Psychological Bulletin
2004, Vol. 130, No. 2, 289–303

0033-2909/04/$12.00 DOI: 10.1037/0033-2909.130.2.289

The Intervention Selection Bias: An Underrecognized Confound in Intervention Research

Robert E. Larzelere and Brett R. Kuhn
University of Nebraska Medical Center

Byron Johnson
University of Pennsylvania

Selection bias can be the most important threat to internal validity in intervention research, but is often insufficiently recognized and controlled. The bias is illustrated in research on parental interventions (punishment, homework assistance); medical interventions (hospitalization); and psychological interventions for suicide risk, sex offending, and juvenile delinquency. The intervention selection bias is most adequately controlled in randomized studies or strong quasi-experimental designs, although recent statistical innovations can enhance weaker designs. The most important points are to increase awareness of the intervention selection bias and to systematically evaluate plausible alternative explanations of data before making causal conclusions.

The difficulty of making causal conclusions from nonrandomized studies is widely recognized by methodologists (e.g., Copas & Li, 1997; Shadish, Cook, & Campbell, 2002; Wilkinson & the Task Force on Statistical Inference, 1999). Nonetheless, causal conclusions are regularly made about many interventions from inadequate evidence. If these conclusions are wrong, they can adversely affect clinical practices, public policies, and research directions. We argue in this article that premature causal conclusions are being made about several parental and psychological interventions, hindering scientific progress in these areas.

Selection bias is often the most important threat to making valid causal inferences in intervention research. It is the artifactual part of postintervention outcome differences that is due to differences in preexisting characteristics of the groups being compared, when these characteristics lead to distinct prognoses. For example, those with more difficult presenting problems may be more likely to get selected for an intervention. If so, those initial presenting problems may be related to group differences in postintervention outcomes, which can then masquerade as apparent intervention effects (or lack thereof) if the selection bias is not accounted for. Correct conclusions about intervention effects require convincing evidence that selection differences have been minimized or accounted for. Well-conducted randomized designs accomplish this goal but are rarely used to study some interventions for ethical or pragmatic reasons.

Emphasizing the intervention selection bias does not mean that other issues are unimportant in intervention research. Other threats to internal validity may also play an important role in any given study. Most applications also require external validity, that is, the generalization of intervention effects across persons, settings, and other dimensions. Internal validity and external validity are both essential to know which intervention is optimal in a given situation (Shadish et al., 2002). Without adequate internal validity, however, we can only apply interventions of uncertain effectiveness.

Examples

We begin this article by illustrating selection biases in research on interventions by parents for behavioral or homework problems; by medical personnel for hospitalization; and by psychologists for suicide risk, sex offending, and delinquency. The final sections provide a brief overview of strategies to minimize a selection bias, emphasizing their roles in ruling out plausible alternative explanations.

Parental Interventions

First, we consider two parental interventions: disciplinary punishment and homework assistance. Are these interventions counterproductive, even when intended to help children? Many social scientists think so (Keith, 1986; Straus, 2001), even though their conclusions are based primarily on correlational evidence from passive observational designs. These findings may be artifacts of an intervention selection bias.

Parental Punishment

Many developmental psychologists have concluded that power assertive disciplinary tactics, including all forms of punishment, have detrimental effects on children's prosocial behavior and moral internalization (Bee, 1998; Berger & Thompson, 1995;

Robert E. Larzelere and Brett R. Kuhn, Department of Psychology, Munroe-Meyer Institute, University of Nebraska Medical Center; Byron Johnson, Center for Research on Religion and Urban Civil Society, University of Pennsylvania.

Byron Johnson is now at the Witherspoon Institute, Princeton, New Jersey.

An earlier version of this article was presented at the 106th Annual Convention of the American Psychological Association, San Francisco, August 1998.

Correspondence concerning this article should be addressed to Robert E. Larzelere, Department of Psychology, Munroe-Meyer Institute, 985450 University of Nebraska Medical Center, Omaha, NE 68198-5450. E-mail: rlarzelere@unmc.edu

Bornstein & Lamb, 1988; Etaugh & Rathus, 1995; Grolnick, Deci, & Ryan, 1997; Kochanska, Padavich, & Koenig, 1996). In contrast, behavioral clinicians have emphasized the effectiveness of nonphysical punishment such as time-out to reduce behavior problems in children with externalizing disorders (Aronfreed, 1968; Axelrod & Apsche, 1983; National Institutes of Health, 1991; Patterson, 1982; Walters & Grusec, 1977).

Of the common forms of parental punishment, we focus on physical punishment because it is considered the most detrimental power assertive tactic (Garbarino, 1996; Straus, 1999, 2001). If the presumed detrimental effects of physical punishment can be explained by the intervention selection bias, then the same artifact could easily explain conflicting conclusions about nonphysical punishment.

Evidence. Three lines of evidence indicate that the detrimental correlates of nonabusive physical punishment are due to a selection bias. First, the pattern in three literature reviews indicates that detrimental effects of physical punishment tend to disappear in research designs with stronger internal validity, even to the point of showing beneficial effects in randomized clinical trials. Second, alternative disciplinary tactics are just as strongly associated with detrimental outcomes as is nonabusive physical punishment. Third, the detrimental outcomes of ordinary physical punishment tend to disappear with more adequate statistical controls for initial child misbehavior.

In two overlapping qualitative literature reviews, 47% of 36 prospective or retrospective studies found that physical punishment predicted detrimental child outcomes compared with only 6% that found beneficial child outcomes (Larzelere, 1996, 2000). None of the 36 studies, however, controlled for initial severity of child misbehavior. In contrast, 59% of 17 studies that did control for initial misbehavior found beneficial child outcomes for nonabusive spanking, including all 4 randomized clinical trials.

More recently, Gershoff's (2002) meta-analysis found a similar pattern for child correlates of physical punishment. The unweighted mean of 113 effect sizes from cross-sectional, retrospective, and longitudinal studies indicated that physical punishment was associated with detrimental child characteristics ($d = 0.45$). None of these 113 effect sizes controlled for the initial severity of child misbehavior. In contrast, the mean effect size (d) of four randomized studies was -1.15, indicating that physical punishment was associated with beneficial child outcomes. This is one of the largest differences ever found between randomized and nonrandomized studies in a meta-analysis (Lipsey & Wilson, 1993). Most of the randomized clinical trials were conducted by Roberts and his colleagues. They first confirmed that the traditional spank backup was essential to the effectiveness of behavioral parent training with noncompliant 2- to 6-year-olds (Bean & Roberts, 1981). They later documented that a brief room isolation was just as effective as a spank backup for time-out and showed that each of these two backups was more effective than a restraint backup (Day & Roberts, 1983; Roberts, 1988; Roberts & Powers, 1990).

Whereas the beneficial effects in randomized clinic trials contradicted the detrimental associations from correlational studies, the results of longitudinal studies that controlled for initial misbehavior have produced mixed results. According to Larzelere's (2000) review, six longitudinal studies with those controls found significantly detrimental child outcomes of physical punishment (Adams, 1995; Baumrind, 2001; Larzelere & Smith, 2000; McLeod, Kruttschnitt, & Dornfeld, 1994; Simons, Lin, & Gordon, 1998; Straus, Sugarman, & Giles-Sims, 1997); two found both beneficial and detrimental child outcomes depending on age, ethnicity, and religiosity (Ellison, Musick, & Holden, 1998; Gunnoe & Mariner, 1997); and one found neither beneficial nor detrimental longitudinal outcomes (Larzelere, Sather, Schneider, Larson, & Pike, 1998). Across these nine studies, beneficial outcomes occurred more often than detrimental outcomes for children under the age of 7, for African Americans, and for Conservative Protestants (Larzelere, 2000).

Overall, the apparently detrimental effects of nonabusive physical punishment disappeared and even reversed in research designs with stronger internal validity, especially when spanking was used to back up milder disciplinary tactics in defiant 2- to 6-year-olds or in subcultural groups that view spanking as more normative. If the apparently detrimental correlational evidence is due to a selection bias, then similar correlations should occur for alternative disciplinary tactics when investigated in similar ways.

The implicit assumption of the correlational studies has been that effective disciplinary responses would be associated with beneficial child outcomes (e.g., reduced aggression). This assumption was contradicted by a study that investigated all disciplinary responses reported by mothers of toddlers. It found that the frequency of each disciplinary response was correlated with higher rates of disruptive behavior (aggressive and oppositional behavior), whether disruptive behavior was measured concurrently or 20 months later (Larzelere et al., 1998; Larzelere, Schneider, Larson, & Pike, 1996; see Table 1).

The first data column in Table 1 (Column 2 overall) shows the correlations of the frequency of seven disciplinary responses with disruptive behavior 20 months later. The frequency of physical punishment correlated at about .15 with subsequent disruptive behavior, indicating an apparently detrimental child outcome. This is close to the mean effect size r of .18 (equivalent to $d = 0.37$) for the 17 prospective studies in Gershoff's (2002) meta-analysis that predicted aggression-related outcomes (Baumrind, Larzelere, & Cowan, 2002).

However, the frequency of every alternative disciplinary response in Table 1 also correlated positively with subsequent disruptive behavior. The largest correlations occurred for nonphysical punishment alone (rs = .29 and .33), reasoning alone (.40 and .53), and disciplinary responses that included neither punishment nor reasoning ("other": .34 and .32). Thus, recommended alternative tactics not only failed to correlate negatively with subsequent disruptive behavior, but they correlated more positively than did physical punishment.

This pattern of longitudinal correlations is easily explained by the intervention selection bias. The frequency of any disciplinary tactic tends to reflect the frequency of misbehavior to which the parent is responding. Therefore, frequency measures of any disciplinary tactic tend to be associated with higher subsequent levels of disruptive behavior, merely due to the continuation of children's disruptive behavior over time. In part, the size of the positive correlations corresponds to how commonly the disciplinary tactics are used. More common disciplinary tactics reflect misbehavior frequency more closely and thus predict subsequent misbehavior more strongly. For similar reasons, Straus and Mouradian (1998) found that the frequency of three recommended disciplinary tactics correlated more strongly with antisocial behavior than did physical

Table 1
Apparent Effects of Toddler Disciplinary Responses on Disruptive Behavior[a] 20 Months Later

	Measure of discipline response		
	Frequency	Proportional usage	Proportional usage
Discipline response	*r*		Partial *r*[b]
Discipline response to fighting incidents (N = 27 mother–child pairs)			
Reasoning without punishment (230)	.40*	.36†	.39*
All physical punishment (83)	.16	−.16	−.17
Physical punishment alone (53)	.18	−.04	−.28
Physical punishment and reasoning (30)	.05	−.20	.08
Nonphysical punishment alone (46)	.29	−.15	−.10
Nonphysical punishment and reasoning (24)	.11	−.13	−.03
Other (418)	.34†	−.18	−.22
Discipline response to disobedience incidents (N = 31 mother–child pairs)			
Reasoning without punishment (777)	.53**	.29	.37*
All physical punishment (382)	.14	−.31†	−.04
Physical punishment alone (283)	.11	−.24	−.13
Physical punishment and reasoning (99)	.12	−.24	.24
Nonphysical punishment alone (177)	.33†	.08	−.10
Nonphysical punishment and reasoning (71)	.15	−.34†	−.07
Other (1,557)	.32†	.09	−.20

Note. The number of mother–toddler pairs was the sample size for testing whether these correlations differed from .00. The numbers in parentheses following each disciplinary response are the total number of disciplinary incidents in which a mother used that particular disciplinary response, summed across all the mother–toddler pairs. This table is an expanded version of Table 3 from "Punishment Enhances Reasoning's Effectiveness as a Disciplinary Response to Toddlers," by R. E. Larzelere, P. R. Sather, W. N. Schneider, D. B. Larson, and P. L. Pike, 1998, *Journal of Marriage and the Family, 60,* p. 400. Copyright 1998 by the National Council on Family Relations. Adapted with permission.

[a] Disruptive behavior was based on two subscales of the Toddler Behavior Checklist, completed after the month of structured diary data on responses to fighting and disobedience and again 20 months later. [b] Controlling for disruptive behavior at Time 1 (i.e., for 1 month between 26 and 39 months of age).

† $p < .10$. * $p < .05$. ** $p < .01$.

punishment in a cross-sectional study. A frequency measure of disciplinary tactics exacerbates selection bias effects, because it directly reflects the frequency of misbehavior.

Consistent with this interpretation, Larzelere's (1996, 2000) two literature reviews found six disciplinary tactics that were associated with more detrimental child outcomes than was physical punishment, whereas only grounding of teenagers was associated with more beneficial outcomes than was physical punishment. Similarly, studies in Gershoff's (2002) meta-analysis that investigated alternative disciplinary tactics were as likely to have effect sizes favoring physical punishment as favoring the alternative tactics, especially before the teenage years (Baumrind et al., 2002).

Column 3 of Table 1 shows that this selection bias can be minimized with a proportional measure: the proportion of all discipline incidents in which that disciplinary response was used. Whereas the frequency of each disciplinary response is positively correlated with subsequent disruptive behavior problems, proportional measures have more varied correlations with them. Column 4 adds a statistical control, using the initial level of disruptive behavior as a covariate, with results similar to Column 3.

One crucial assumption for valid causal inferences from covariate-adjusted statistics is that the covariate must be measured without error (Campbell & Kenny, 1999; Freedman, 1987; Huitema, 1980; Rothman & Greenland, 1998b). If detrimental child correlates of physical punishment are due to the intervention selection bias, they should disappear as covariates for the initial severity of child misbehavior are improved to reduce measurement error. Straus et al. (1997) used a covariate adjustment and obtained the strongest evidence of detrimental causal effects of physical punishment to date. Their measure of initial antisocial behavior, however, consisted of only three levels: zero, low, and high antisocial behavior. Larzelere and Smith (2000) replicated the Straus et al. results using the same longitudinal data set, but showed that their findings became nonsignificant after controlling for a more comprehensive 16-item measure of externalizing behavior problems. In addition, with the original trichotomous covariate for antisocial behavior, Larzelere and Smith obtained similar apparently detrimental effects for grounding, sending children to their room, and removing an allowance. These results suggest that the strongest causal evidence for the detrimental effects of customary physical punishment (Straus et al., 1997) is hampered by residual confounding (Rothman & Greenland, 1998b) due to an inadequate measure of initial antisocial behavior.

Implications. The pattern of evidence is consistent with the view that the intervention selection bias accounts for most detrimental child outcomes associated with nonabusive physical punishment. Presumably, the bias may account for correlational evidence for the apparently detrimental effects of nonphysical

punishment as well (e.g., Table 1). Premature conclusions about detrimental effects of physical and nonphysical punishment could potentially affect research, clinical practice, and policy in adverse ways.

Premature conclusions may hinder research to synthesize the divergent perspectives about parental discipline represented by behavioral parent training and cognitive developmental psychology. Those perspectives generally complement each other well (Larzelere, Schneider, et al., 1996), but they often make contradictory recommendations about whether reasoning or nonphysical punishment is preferred to address inappropriate child behavior. One behavioral parent trainer, for example, concluded, "If I were allowed to select only one concept to use in training parents of antisocial children, I would teach them how to punish more effectively," referring to time-out (Patterson, 1982, p. 111). In contrast, major developmental experts have said, "In general, parental use of power-assertive or forceful techniques to effect children's compliance has been considered detrimental to the development of internalization" (Kochanska & Thompson, 1997, p. 67). The two perspectives not only recommend different disciplinary tactics, but they sometimes denigrate the tactic recommended by the other perspective (Blum, Williams, Friman, & Christophersen, 1995; Christophersen, 1988, 1990; Holden, 1997; Kuczynski & Hildebrandt, 1997). Empirically based theories of child rearing therefore remain largely isolated from one another (Holden, 1997).

Premature confidence about the detrimental effects of power assertion, although based largely on correlational evidence, is one factor hindering resolutions of these differences. Developmental psychologists have not explained why nonphysical punishment is a critical component of clinical interventions that teach parents to manage their children's behavior more effectively. At the same time, behavioral clinicians have not explained why the parents of well-behaved children rely on reasoning more than nonphysical punishment. Although several theoretical perspectives could help synthesize these two perspectives, their potential has not been fully exploited (Baumrind, 1973; Bell, 1968; Bell & Harper, 1977; Hoffman, 1977; Larzelere, 2001; Patterson, 1982).

Second, incorrect premature conclusions can adversely affect clinical practices. Behavioral parent training has not only been empirically supported as a treatment package for clinically disruptive children (Brestan & Eyberg, 1998; Kazdin, 1995), but its essential treatment components have been documented in rigorous randomized studies (Roberts & Powers, 1990). Nonetheless, premature conclusions about the detrimental effects of spanking have prompted many behavioral parent trainers to abandon spanking as an option for children who escape time-out, despite its status as one of the two best-supported enforcement options for time-out with 2- to 6-year-olds.

It is hard to imagine a medication being dropped from clinical use on the basis of similar evidence. If there were one leading medication for a given malady, and a new medication demonstrated equivalent effectiveness, this would not result in the first being eliminated from consideration. Rather, clinical practice would be enhanced by having two options available on a case-by-case basis according to the relevant indications and contraindications. Furthermore, if one of the medications appeared ineffective for a particular client, the physician could then prescribe the other. In contrast, the spank backup has not only been largely abandoned, but has often been replaced by physical restraint, which was found to be significantly less effective (Roberts & Powers, 1990). Thus, at a time of increased need for effective treatments to prevent clinically disruptive children from becoming juvenile delinquents, one of the most replicated, effective treatments has been weakened because of premature conclusions about the detrimental effects of nonabusive physical punishment, derived mostly from correlational studies that controlled inadequately for the intervention selection bias. Alternatives to spanking may ultimately prove to be better at enforcing time-out in clinically defiant 2- to 6-year-olds. The point is that currently used alternatives have rarely been established by research as strong as the original research evidence for the spank backup.

Finally, premature conclusions also affect public policymaking. Eleven countries have adopted policies to ban parental spanking (EPOCH-Worldwide, 2002). Four social scientists were expert witnesses supporting a legal challenge against Canadian parents' right to use "reasonable force" to correct their children. They submitted an earlier version of Gershoff's (2002) literature review as evidence (Holden, 1998). In the final published version, Gershoff emphasized that her meta-analysis should not be used to support causal conclusions, because 113 out of 117 effect sizes were based on correlational data. Nonetheless, a previous version of her review has already been used to support a parental spanking ban, a policy implication that requires causal, not correlational evidence.

Homework Assistance

A similar pattern of results can be found in research evaluating parental assistance with children's homework. When children are doing poorly in school, parents tend to respond by supervising homework more closely (Levin et al., 1997; Pomerantz & Eaton, 2001; Singh et al., 1995). This may explain why most correlational studies have found parental assistance with homework to be associated with lower achievement (Balli, Wedman, & Demo, 1997; C. Chen & Stevensen, 1989; Desimone, 1999; McDermott, Goldman, & Varenne, 1984; Miller & Kelley, 1991; Singh et al., 1995). Consequently many have concluded that homework assistance may inadvertently hinder children's development by fostering dependency (C. Chen & Stevensen, 1989; Desimone, 1999; Keith, 1986; Levin et al., 1997; Pomerantz & Eaton, 2001).

Studies with stronger internal validity have found beneficial or mixed effects of homework assistance, suggesting that the correlational evidence may be an artifact, possibly due to selection bias. First, clinical outcome studies have consistently found improvements in homework quantity and quality after training parents to supervise homework and use behavioral management techniques (Anesko & O'Leary, 1982; Dougherty & Dougherty, 1977; Forgatch & Ramsey, 1994; Goldberg, Merbaum, Even, Getz, & Safir, 1981; Kahle & Kelley, 1994; Koven & LeBow, 1973; Loitz & Kratochwill, 1995; Maertens & Johnston, 1972; Miller & Kelley, 1991; Rhoades & Kratochwill, 1998). These studies investigated specifically trained supervision skills, not untrained parental assistance with homework. When longitudinal studies have controlled statistically for prior academic achievement, they have found that untrained homework assistance improves children's academic performance as often as not (Levin et al., 1997; Singh et al., 1995). For example, Pomerantz and Eaton (2001) found that "intrusive"

homework support predicted improved academic achievement once prior achievement was statistically controlled for.

The correlational evidence for the detrimental effects of parental homework assistance has been accepted sufficiently to weaken conclusions based on stronger evidence to the contrary. From the above findings, for example, Pomerantz and Eaton (2001) concluded, "The achievement of children whose mothers frequently used intrusive support improved over time but did not exceed that of children whose mothers infrequently used intrusive support" (p. 174). Despite its technical accuracy, a similar conclusion would be highly unusual in an equivalent study of a professional intervention (e.g., "Head Start improved children's academic achievement, but not to a level exceeding children who did not receive [or need] Head Start").

This pattern is remarkably similar to the pattern of results and conclusions for disciplinary punishment. The intervention effects are consistently found to be effective for clinically supervised interventions, mixed for longitudinal studies that control statistically for selection bias, and apparently counterproductive in correlational studies. Furthermore, recommendations for research and practice appear to be more strongly influenced by correlational evidence than by studies with greater internal validity. Although longitudinal studies have stronger ecological validity than randomized studies, the latter rule out the intervention selection bias more adequately. In this case, clinical outcome studies have shown that it is possible for trained parents to help their children perform better on homework, providing strong evidence against correlationally based conclusions that parental assistance is counterproductive.

Medical Interventions

The strong association between hospitalization and mortality provides an informative example of the intervention selection bias. For diseases tracked by the Health Care Financing Administration in Medicare patients, the mortality rate is around 6.5% for hospital inpatients (Green, Passman, & Wintfield, 1991). It is not surprising that mortality rates in intensive care units are higher, ranging from 15% to 25% (Goldhill & Withington, 1998; Zimmerman et al., 1998). By comparison, actuarial tables indicate that mortality rates for a 30-day period in 1980 were 0.1% for 65-year-olds, 0.2% for 70-year-olds, and about 2.5% for 100-year-olds in the United States (Faber, 1982). Thus the relative risk of death in a hospital appears to be about 30 times what it would be outside the hospital and even higher in an intensive care unit.

Despite these statistics, differences in mortality rates have not been used to make conclusions about the overall effectiveness of hospitalization, which would be directly analogous to the logic used to make causal inferences about parental interventions. Mortality rates have been used, however, to compare hospitals' quality of care with each other since 1986 (Kahn et al., 1988). Critics immediately raised the issue of selection bias. They pointed out that the hospital with the highest mortality rate turned out to be a hospice caring for terminally ill patients (Randolph, Guyatt, & Carlet, 1998). Hospitals with high mortality rates tended to have a higher percentage of patients 85 or older with high-risk diagnoses who required nursing home care (Green et al., 1991). One study concluded that patient characteristics were 315 times more important than hospital characteristics in predicting mortality after simple surgery (Silber & Rosenbaum, 1997).

Such concerns led to extensive research to improve risk-adjusted mortality rates with better measures of illness severity. Leading severity indices used 14 to 17 variables and accounted for 11.8% to 17.8% of variance in mortality rates (Daley et al., 1988; Escarce & Kelly, 1990; LeGall, Lemeshow, & Saulnier, 1993; Lemeshow et al., 1993).

Despite extensive research to improve severity indices, statistical controls for the severity of medical illness are generally considered inadequate to transform risk-adjusted mortality rates into valid indicators of the quality of care (L. M. Chen, Martin, Keenan, & Sibbald, 1998; Daley et al., 1988; Schuster & Kollef, 1994; Silber & Rosenbaum, 1997; Thomas & Hofer, 1998). Silber and Rosenbaum (1997) concluded that "it was the concern about incomplete severity adjustment that led to the failure of the Health Care Financing Administration mortality models" (p. OS88). Nonetheless, risk-adjusted mortality rates continue to be used to evaluate the quality of hospital care. For example, they constitute one part of the basis for determining the best American hospitals in an annual listing by *U.S. News & World Report.*

In two respects, risk-adjusted mortality rates constitute stronger causal evidence of the effects of hospitalization than the evidence for detrimental effects of the two parental interventions. First, hospitalization clearly precedes its outcome (death), whereas the temporal sequence is more ambiguous in most correlational studies of parental interventions. Second, risk adjustment for initial severity has been studied thoroughly in the medical literature, whereas it has been generally ignored in parental intervention research. Nonetheless, causal inferences are more readily made about the two parental interventions than about hospitals' quality of care.

Psychotherapy Interventions

The intervention selection bias can also play an important role in research on psychotherapy interventions. Outcome studies of interventions for suicide risk, sex offending, and juvenile delinquency often rely on weaker quasi-experimental designs, which are especially vulnerable to the intervention selection bias.

Suicide Risk

The association of psychotherapy with subsequent suicide attempts provides the first illustration. Larzelere, Smith, Batenhorst, and Kelly (1996) reviewed prospective longitudinal studies of suicide attempts in children and adolescents. Psychotherapy was one of the most replicated predictors of subsequent suicides and attempts. The only predictor that was replicated more often was a previous suicide attempt. Not only did the longitudinal associations appear to suggest detrimental effects of psychotherapy, but the magnitude of the association was substantial. Table 2 shows the relative suicide rates of children and adolescents who had participated in psychological treatment compared with those who had not. The median relative risk from the nine studies indicates that youth who received psychological treatment were 14.3 times more likely to commit suicide than their comparison group.

Psychological treatment also predicted higher subsequent rates of suicide attempts (see Table 3). With the exception of family therapy, youth who received psychotherapy were substantially more likely to make a future attempt than youth who did not receive treatment (median odds ratio = 6.2).

Table 2
Suicide Rates Following Psychological or Medical Treatment of Children and Adolescents

Study	Sex	*N*	Age (years)	Treatment	Years followed	Suicide rate (%)		RR ratio
						Treatment	Comparison	
Garfinkel et al. (1982)	M, F	505	6–21	Emergency room treatment for attempters	1–9	1.00	0.03[a]	32.6
Goldacre & Hawton (1985)	M	641	12–20	Hospitalized for self-poisoning	1–5	0.62	0.04[a]	15.9
	F	1,851	12–20	Hospitalized for self-poisoning	1–5	0.11	0.01[a]	11.0
Kuperman et al. (1988)	M	881	2–18	Psychiatric inpatients	4–15	1.02	0.12[b]	8.3
	F	450	2–18	Psychiatric inpatients	4–15	0.44	0.03[b]	14.3
Motto (1984)	M	122	10–19	Psychiatric inpatients	4–10	9.00	0.09[a]	99.1
Otto (1972)	M	321	10–20	Psychiatric or medical treatment for attempt	10–15	10.00	0.31[c]	32.0
	F	1,226	10–20	Psychiatric or medical treatment for attempt	10–15	2.90	0.73[c]	3.9
Shafii et al. (1985)	M, F	20	12–19	Prior contact with mental health professionals	n/a[d]	69.22	45.80	2.7[e]

Note. RR = relative risk; M = male; F = female; n/a = not applicable.
[a] Estimated from Figure 2 for 15- to 19-year-olds in Shaffer et al. (1988). [b] Estimated for all Iowans of same age and gender. [c] Estimated from excess deaths in comparison sample compared to nonsuicide deaths in attempters. [d] Case-control retrospective postmortem study. [e] Odds ratio (in other cases, the odds ratio is slightly larger than the RR ratio).

Three points can be made from this example. First, if causal interpretations were made from these longitudinal correlations in the same way as for parental interventions, one would conclude that psychotherapy is detrimental for children and adolescents. Second, the magnitude of the intervention selection bias can be substantial, assuming that psychotherapy does not actually increase the risk of suicide. Third, the intervention selection bias can produce findings that conceal an effective intervention or that make an intervention appear worse than it actually is. For example, the effects of family therapy in two small studies were much more positive than any other intervention (see Table 3). This pattern of results could occur if family therapy was the only intervention sufficiently effective to overcome the intervention selection bias. To meet conventional standards for effectiveness, however, family therapy had to not only overcome the intervention selection bias but do so to a statistically significant degree in the opposite direction. This unrealistically high standard was not achieved in either small study, suppressing the potential effectiveness of family therapy. Family therapy has been shown to be particularly effective for certain other problems (e.g., drug abuse), despite retaining more difficult cases in therapy (Stanton & Shadish, 1997). Still, it could be that family therapy only appears more effective for suicide risk because of the selection of cases with better prognoses than other interventions.

It is important to note that none of the studies listed in Tables 2 and 3 concluded that therapy caused an increase in suicides or suicide attempts. The purpose of the studies was to determine what factors predicted subsequent suicides and suicide attempts, yet the authors were ambivalent about including psychotherapy as a risk factor for suicide. All 4 studies that discussed therapy implications suggested that the standard treatment was ineffective and should be changed in some way (Barter, Swaback, & Todd, 1968; Brent et al., 1993; Cohen-Sandler, Berman, & King, 1982; Garfinkel, Sroese, & Hood, 1982). None of the 10 studies made any attempt to control for selection bias, and few even acknowledged it.

These studies nevertheless may represent the strongest evidence regarding the effectiveness of suicidal interventions for children and adolescents. Linehan (1997) found only 20 comparison-group studies that evaluated suicidal interventions, all of which targeted adults. She concluded that little is known about effective interventions for suicide risk for adults, and even less for minors. Unlike the previous examples, premature conclusions do not appear to

Table 3
Rates of Suicide Attempts Following Psychological or Medical Treatment of Children and Adolescents

Study	Sex	*N*	Age (years)	Treatment	Years followed	Subsequent suicide attempts (%)		Odds ratio
						Treatment	No-treatment	
Barter et al. (1968)	M, F	45	13–20	Social welfare or mental health contact	0.3–3.7	62.5	19.0	7.08
Brent et al. (1993)	M, F	134	13–18	Rehospitalized	0.3–1.2	36.4	4.5	12.23
	M, F	134	13–18	Received psychotropic medication	0.3–1.2	15.3	5.3	3.20
	M, F	134	13–18	Family therapy	0.3–1.2	2.8	12.2	0.20
Cohen-Sandler et al. (1982)	M, F	20	6–16	Individual or group therapy	0.4–3.0	30.8	0.0	5.33[a]
	M, F	20	6–16	Family therapy	0.4–3.0	11.1	30.0	0.29
Garfinkel et al. (1982)	M, F	505	6–21	Psychosocial services prior to hospitalization	n/a[b]	80.4	34.1	7.91
Pfeffer et al. (1991)	M, F	69	4–14	Psychiatric inpatients[c]	6–8	23.2	6.3	4.53

Note. M = male; F = female; n/a = not applicable.
[a] Based on 0.5 attempts in no-treatment group. [b] Retrospective matched case-control study (the percentages are inflated, but the odds ratio is unbiased). [c] Compared with community controls.

have produced adverse effects on recommendations or clinical practice. Rather, the empirical evidence about suicidal interventions has simply been ignored.

Sex Offending

Sex offending is purported to be one of the fastest growing violent crimes in the United States (Shaw & Work Group on Quality Issues, 1999). Sex offenders constitute about one third of prison populations in some locations (Polizzi, MacKenzie, & Hickman, 1999), and 60% to 80% of sex offenders are rearrested for a sexual or violent crime within 2 decades of their initial arrest. A child molester averages more than 75 victims, whereas an adult rapist averages 7.5 victims (Polizzi et al., 1999). Recidivism concerns have led to recent laws that permit state authorities to extend the confinement of some sex offenders (Janus, 2000; Wood, Grossman, & Fichtner, 2000). Although one motivation is to protect the public, the constitutionality of these laws could depend on the effectiveness of rehabilitation (Janus, 2000; Wood et al., 2000).

Most treatment outcome studies for sex offenders have used weaker quasi-experimental designs, such as posttreatment comparisons of nonequivalent groups. There have been few randomized clinical trials because no-treatment control groups are usually considered unethical. There is some evidence that sex offender treatment is effective, but that evidence is confounded with the intervention selection bias. Although Furby, Weinrott, and Blackshaw (1989) concluded that there was "no evidence that clinical treatment reduces rates of sex offenses" (p. 27), more recent reviews have reached modestly positive conclusions (Hall, 1995; Polizzi et al., 1999; Shaw & Work Group on Quality Issues, 1999; but see Harris, Rice, & Quincey, 1998, for a more pessimistic review).

Hall's (1995) meta-analysis of 12 comparison-group studies may be the most cited review in this area. He concluded that the average effect size was moderate but promising ($d = 0.24$). However, Harris et al. (1998) concluded that Hall's result was inflated by several studies that compared treatment groups with comparison groups dominated by treatment refusers or dropouts. In contrast, the five studies whose comparison groups were unbiased according to Harris et al. had an average effect size (d) of -0.01, indicating no effect. Treatment dropouts and treatment refusers are known to have a higher risk of recidivism (Harris et al., 1998; Marques, 1999). Attrition from sex offender treatment programs is common, introducing a large selection bias into the subgroup that completes treatment. It is typical, for example, to have 70% of sex offenders refuse or drop out of antiandrogen drug treatments (Grossman, Martis, & Fichtner, 1999; Harris et al., 1998). Selection bias is exacerbated when refusers or dropouts are included in the no-treatment control group.

Thus selection bias due to attrition is a plausible explanation of the evidence used to support the effectiveness of sex offender treatment. This led Harris et al. (1998) to conclude,

> Agreeing to and persisting with treatment over the long term serves as a filter for detecting those offenders who are relatively less likely to reoffend, but the nature of the treatment has little or no detectable specific effect on outcome. (p. 103)

Selection bias not only undermines causal inferences about the average effectiveness of treatments for sex offenders, but it also makes comparisons among alternative treatments more ambiguous. Hall's (1995) comparisons of different types of treatment were compromised because the degree of selection bias varied by treatment (Harris et al., 1998). Treatments that appeared more effective may merely have controlled less adequately for selection bias, which may explain why the treatment considered most effective by Hall (cognitive–behavioral therapy) later proved ineffective in a major randomized clinical trial (Marques, 1999). In fact, participants assigned to treatment had a nonsignificantly higher recidivism rate than controls.

Once again, premature conclusions about treatment effectiveness often hinder future progress in research and practice guidelines. A recent "research overview" (Grossman et al., 1999) concluded that a 30% reduction in recidivism due to treatment for sex offenders was a "robust finding," mostly on the basis of Hall's (1995) meta-analysis. Other research summaries have followed Hall's meta-analysis in recommending cognitive–behavioral treatment and antiandrogen treatment, even though the latter has a 70% refusal/dropout rate (Grossman et al., 1999; Heilbrum, Nezu, Keeney, Chung, & Wasserman, 1998).

As for clinical practice, recent guidelines from the American Academy of Child and Adolescent Psychiatry stated, "Realistic assurance should be provided that [sexual perpetrator] problems are treatable" and "In spite of the lack of empirical rigor, there is considerable evidence that treatment interventions are effective in interrupting the course of sexually abusive behavior" (as cited in Shaw & Work Group on Quality Issues, 1999, p. 68S). Hall's (1995) meta-analysis was the primary reference provided to support these statements.

Premature conclusions about effectiveness can also undermine the development of new or improved treatments. One example is multisystemic family therapy, which significantly reduced recidivism in sex offenders in a small randomized clinical trial (Borduin, Henggeler, Blaske, & Stein, 1990). Such promising treatments and other innovative approaches may not be encouraged or funded if existing treatments are considered effective. Although some have questioned whether randomized clinical trials are ethical for sex offenders (Maletzky, 1997; Shaw & Work Group on Quality Issues, 1999), this assumes the availability of effective treatments. As Harris et al. (1998) put it, "Recommended treatment continues to evolve, but the evolution is not based on an empirical foundation of effective treatment and has no chance to be" (p. 103).

Juvenile Delinquency

Outcome research on juvenile delinquency yields a similar pattern of results despite more widespread use of stronger research designs. Like sex offender research, delinquency studies rely primarily on recidivism as the outcome variable, which rarely has an equivalent pretest measure. This, in turn, hinders the ability of quasi-experiments to rule out selection biases related to recidivism differences.

In the most extensive meta-analysis of delinquency interventions, 45% of 200 studies of serious delinquents used random assignment (Lipsey & Wilson, 1998). The other 55% used a variety of quasi-experimental designs, such as matched or nonequivalent group designs. The effect size estimates were not adjusted for pretreatment differences, probably because few of the

original studies provided the necessary information (Lipsey & Wilson, 1998).

The meta-analysis concluded that delinquency interventions are effective, despite a small average effect size ($d = 0.12$; Lipsey, 1999; Lipsey & Wilson, 1998). The authors, Lipsey and Wilson (1998), noted that randomized studies had a significantly smaller effect size than nonrandomized studies, without specifying the precise statistic. Any effect size significantly less than 0.12 would approach zero, indicating no effect. Similar results occurred in a meta-analysis of European studies of adult and juvenile offenders, which found an overall mean effect size (*d*) of 0.24 but of only 0.04 in the three randomized studies (Redondo, Sanchez-Meca, & Garrido, 1999).

Despite the disappointing results from randomized designs, confident conclusions have been made that interventions for delinquency work (Andrews et al., 1990; D. C. Gibbons, 1999; Hollin, 1999; Lipsey & Wilson, 1998; Redondo et al., 1999). Although these conclusions stem from commendable motivations (e.g., to support rehabilitation as an alternative to punitive sentences for delinquents; D. C. Gibbons, 1999), they are based on very small effect sizes that nearly vanish when the analyses are limited to randomized outcome studies. Despite such marginal evidence, the field proceeds as though most interventions for delinquency are effective.

Premature conclusions about the effectiveness of interventions may impede the development of better interventions. Fortunately, in this case the situation is mitigated by two positive developments. First, several meta-analyses have identified promising components of interventions for delinquency (Andrews et al., 1990; Dowden & Andrews, 2000; Izzo & Ross, 1990; Lipsey, 1999; Lipsey & Wilson, 1998). The hope is that future interventions will be enhanced by incorporating these components. Second, there have been systematic efforts to identify "blueprint" delinquency interventions that have proven effective in replicated evaluations that rule out the intervention selection bias (Mihalic, Irwin, Elliott, Fagan, & Hansen, 2001). Federal funding is promoting rigorous evaluations of new replications of these blueprint delinquency treatments. This illustrates the kind of research and development that can yield cumulative progress based on systematic clinical research that minimizes the intervention selection bias.

Other Psychotherapy Examples

Clinical research often produces a similar pattern of results when it depends primarily on posttreatment measures from nonrandomized outcome studies. Research on family preservation (Littell & Schuerman, 1995; Westat, Inc., Chapin Hall Center for Children, & James Bell Associates, 2001) and substance abuse treatment (Pearson & Lipton, 1999; Simon, 1998) also addresses resistant presenting problems (Knight, Hiller, & Simpson, 1999), has substantial relapse and attrition rates (Simon, 1998), and does little to control for pretreatment differences (Pearson & Lipton, 1999). Frequent statements are made that random assignment would be unethical and about the lack of cumulative progress (Simon, 1998), yet leading reviews insist that something works despite small effect sizes that diminish in studies with stronger research designs (Pearson & Lipton, 1999).

Failure to recognize the pervasiveness of the intervention selection bias hinders discriminations between clinical treatments that have well-documented evidence and those that do not. Without rigorous evaluation there will be continued debate over whether interventions work for many difficult clinical problems (Kluger, Alexander, & Curtis, 2000; Kutash & Rivera, 1996). For example, a recent book about "what works" in social work gave the impression that something works for every problem, regardless of the adequacy of supporting causal evidence (e.g., Chamberlain, 2000; Nelson, 2000; see also Westat, Inc. et al., 2001). When strong evidence and weak evidence lead to equivalent recommendations, premature conclusions will continue to be confidently promoted beyond that warranted by their empirical support, hindering the development of more effective intervention practices and policies.

Premature conclusions may become more widespread in clinical practice when accrediting agencies require benchmark comparisons. Because of limited resources, outcomes and decisions are likely to be based on posttreatment measures such as recidivism, consumer satisfaction, and level of functioning. With no correction for selection biases, posttreatment measures will discriminate against treatment programs that serve more difficult clients (Lyons et al., 1997; Thakur, Hoff, Druss, & Catalanotto, 1998). For example, the Joint Commission on Accreditation of Healthcare Organizations is planning to combine juvenile detention centers, residential treatment centers, group homes, and emergency shelters in the same group for benchmarking purposes. These agencies differ widely in the prognoses of the youth they serve.

Although space limitations prevent the discussion of educational interventions, a selection bias often makes compensatory education appear less effective than it actually is. Reanalyses of an initially negative evaluation of Head Start (Westinghouse Learning Corporation & Ohio University, 1969) have accounted for most of the apparently detrimental effects but did not turn them into widespread beneficial effects (Campbell & Boruch, 1975; Campbell & Erlebacher, 1970; Magidson, 1977, 2000; Wu & Campbell, 1996). The current national evaluation of school performance may have similar biases against those serving disadvantaged students.

Minimizing the Intervention Selection Bias

Perhaps the most important step toward minimizing a selection bias is to recognize its pervasiveness and potential magnitude in intervention research. This article tries to increase that awareness in several ways. First, the above examples suggest that premature causal conclusions are being made about several interventions despite the plausibility that a selection bias accounts for the relevant data. Second, we intentionally coin the phrase *intervention selection bias* to refer specifically to selection bias in intervention research. The additional term (*intervention*) is intended to sensitize readers and researchers to the importance of minimizing selection bias artifacts before making causal inferences from intervention studies.

Third, we offer Figure 1 to clarify when the intervention selection bias generally operates against or in favor of an intervention. A *corrective* intervention is one in which a selection bias makes the intervention look less effective than it actually is (Campbell & Boruch, 1975). This typically occurs in nonequivalent group designs or passive longitudinal designs, when the intervention is chosen because of the severity of a presenting problem, whether that problem is behavioral, educational, medical, or psychological.

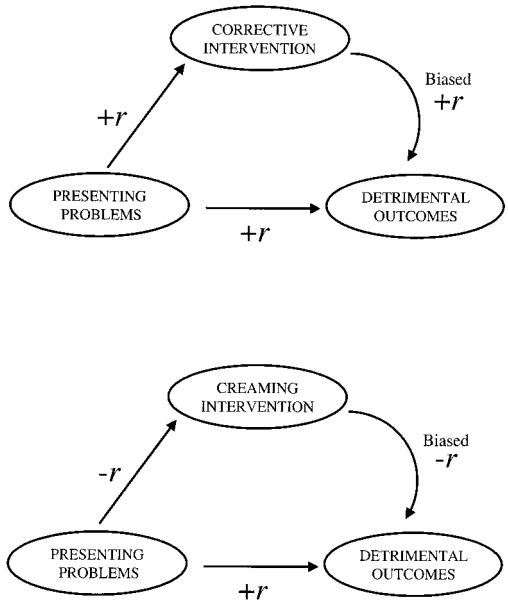

Figure 1. The intervention selection bias for corrective and creaming interventions.

A *creaming* intervention represents the opposite case, in which selection bias makes the intervention look more effective than it actually is. This typically occurs when interventions are intended for advantaged clients (e.g., gifted programs) or when the intervention clients generally have a better prognosis than the comparison group. The treatment outcome studies of sex offenders showed that this could occur by attrition, which has its own set of corrective procedures (e.g., Stanton & Shadish, 1997). In any case, *creaming* refers to the analogy of skimming the cream off the top of the milk, implying that the intervention group includes cases with better preexisting prognoses than the comparison group.

The crucial comparison requires estimating what the outcomes would have been without the intervention (i.e., the *counterfactual*, or missing, results; see Rubin, 1990; Wilkinson & the Task Force on Statistical Inference, 1999). Even if the purpose of an intervention were to correct presenting problems, an evaluation could be biased in its favor if the comparison group had an even poorer prognosis. The direction of this bias would then be depicted as a creaming intervention in Figure 1, despite the corrective purpose of the intervention.

The direction of the typical bias in Figure 1 applies mostly to weaker designs that compare postintervention outcomes only, as in our examples. If intervention and comparison groups were compared on gain scores, then the direction of the bias could reverse because of regression toward the mean (Campbell & Kenny, 1999; Lambert & Bickman, 2001). For example, a disadvantaged group selected for an intervention (e.g., homework assistance) could already have been more likely to improve than an advantaged group that did not require that intervention. Selection bias is a greater threat to internal validity in weaker designs lacking a pretest.

Figure 1 is an oversimplification of the multivariate world in at least two ways. First, other threats to generalized causal inferences may be more relevant in many situations. Second, selection biases and other confounds may apply in complex ways beyond the simple three-variable case, complicating causal inferences further. Although the two paths from the presenting problems are drawn as causal paths, they are labeled as correlations, recognizing that they could be misspecified as casual effects because of other omitted relevant variables. These additional confounds could change both the magnitude and the sign of the apparent causal effect of the intervention. Although other systematic biases may be more crucial in a given application, the emphasis of this article is on the intervention selection bias, which is often the most important bias to consider first.

Confirmation Bias

Failure to recognize plausible threats to internal validity, such as the intervention selection bias, is often due to the *confirmation bias*—that is, the pervasive tendency to emphasize confirmations of favored explanations (Myers, 2004, p. 388; Wason & Johnson-Laird, 1972). The confirmation bias facilitates the logical error of affirming the consequent (Damer, 1980). If the presumed causal pattern is true (the antecedent in a conditional argument), then a given correlational pattern is implied (the consequent). The logical error of affirming the consequent occurs when one observes the implied correlational pattern and concludes that the presumed causal pattern is therefore confirmed. This is a logical error because many other causal patterns could also generate the same correlational pattern. Plausible alternative explanations for the correlational pattern must thus be ruled out to increase confidence in a particular causal explanation.

For over a century, many statistical innovations have burst on the scene with a new flurry of affirming the consequent when proponents claimed too much for the causal conclusiveness of the new method (Copas & Li, 1997; Freedman, 1997; Schuessler, 1978; Turner, 1997). Such exuberance then subsided after experience showed that many alternative interpretations of the underlying data remained plausible even after applying the new method to substantive issues of the day.

Plausible Alternative Interpretations

The key to making valid causal inferences is to systematically rule out plausible alternative interpretations of the relevant data (Rindskopf, 2000; Shadish et al., 2002). Investigators should generate as many plausible explanations of existing data as possible and subject the leading alternatives to competitive empirical tests. Systematically testing plausible alternatives has long been considered crucial for cumulative scientific progress (Chamberlin, 1890/1965; Larzelere & Skeen, 1984; Platt, 1964; Popper, 1935/1959; Rindskopf, 2000). This approach resists the confirmation bias, highlights multiple rational explanations, promotes thoroughness, and suggests plausible explanations that might otherwise be overlooked (Chamberlin, 1890/1965). The intervention selection bias is one plausible alternative explanation that deserves more consideration.

Standard approaches for making valid causal inferences can be considered ways to rule out classes of plausible alterative explanations. The well-known threats to internal validity highlight the most common types of plausible alternatives, including selection biases (Shadish et al., 2002). Epidemiologists emphasize Hill's

(1965) criteria for causality, each of which rules out a set of plausible alternatives. New statistical innovations to enhance causal validity are designed to rule out classes of plausible alternatives that were not as easily ruled out before.

Internal Validity

The internal validity of research designs corresponds directly with the range of alternative explanations that are ruled out. Properly implemented randomized designs rule out most alternative explanations. Creative ways to implement such designs should therefore be a priority whenever possible (Campbell, 1969; Shadish et al., 2002). Regression discontinuity and interrupted time series designs are the strongest quasi-experimental designs. The first specifies the selection process explicitly, and the second controls for both preexisting differences and trends prior to the intervention. Nonequivalent comparison-group designs come next in supporting valid causal inferences (Heinsman & Shadish, 1996; Shadish et al., 2002; Shadish, Matt, Navarro, & Phillips, 2000). Passive nonexperimental designs are generally weaker in internal validity, but longitudinal designs rule out more alternative explanations than cross-sectional designs.

The strongest designs eliminate self-selection by assigning participants to intervention conditions, whereas slightly weaker designs take into account preexisting differences and trends in the outcomes. Nonequivalent group designs account for initial group differences (see Shadish et al., 2000, for details), but rarely for group trends. The weakest designs, which dominated most of this article's examples, adjust for neither preexisting differences nor trends. The recent Shadish et al. (2002) book expands on these designs by emphasizing how design elements and convergent and divergent data patterns can enhance internal validity across a range of designs.

Epidemiological Criteria

Epidemiological criteria for making valid causal inferences from nonexperimental studies also depend on the extent to which they rule out plausible threats to internal validity. Consider the four criteria most widely accepted by epidemiologists (Morton, Heber, & McCarter, 1996; Rothman & Greenland, 1998a, 1998b; Sackett, Haynes, Guyatt, & Tugwell, 1991). First, the strength of an association between a purported cause and effect rules out all plausible alternatives except those that could produce an equally large confound. The suicide and hospitalization examples show, however, that the intervention selection bias can be huge, at least as large as the association between smoking and lung cancer. Temporal sequence, the second criterion, is the only one considered necessary for causality (Hill, 1965), but its necessity applies to the actual temporal sequence, not to when the relevant variables happen to be measured. Measuring the outcome after the intervention does not prove that it reflects intervention effects more than preintervention influences. A related temporal issue is that the time lag between measurements must fit reasonably well with the time lag for a purported cause to occur (Gollub & Reichardt, 1987). Third, consistency is a useful criterion as long as the consistency is not produced by a pervasive systematic confound, such as a selection bias, or by a set of systematic biases that together produce a consistent bias in the same direction across studies. Coherence, the fourth criterion, concerns evidence for relevant causal mechanisms, which often need to be investigated more thoroughly (Eddy, Dishion, & Stoolmiller, 1998).

Statistical Innovations

Four types of statistical innovations can also enhance the validity of causal inferences to the extent that they successfully rule out confounds such as a selection bias. Some statistical innovations model selection factors directly, such as the propensity score (Rosenbaum & Rubin, 1983) and econometric selection modeling (Heckman, 1979; Winship & Morgan, 1999). Others analyze change over time directly, implicitly controlling for preexisting differences (e.g., latent growth models: MacCallum & Austin, 2000; McArdle & Epstein, 1987; Stoolmiller, 1995; dynamic modeling of latent difference scores: Ferrer & McArdle, 2003; McArdle & Hamagami, 2001; and multilevel models: Affleck, Zautra, Tennen, & Armeli, 1999; R. D. Gibbons et al., 1993; Osgood & Smith, 1995; Raudenbush & Bryk, 2002). A third approach is to remove confounds by means of an instrumental variable, which is a variable associated with the outcome only through the intervention, independent of any confounds (Newhouse & McClellan, 1998). A fourth strategy is to estimate how large an unobserved confound would have to be to make a difference in the study's main substantive conclusions, using sensitivity analyses (Rosenbaum, 1995).

Analysts need to exploit the advantages and minimize the disadvantages of these statistical innovations. The first three innovations feature reasonable ways to enhance causal inferences by minimizing confounds such as selection bias. Yet older innovations have rarely lived up to their initially overstated promises (Copas & Li, 1997; Glazerman, Levy, & Myers, 2002; Heckman, 1979; West, Biesanz, & Pitts, 2000; Winship & Morgan, 1999), which has typically led to new refinements of those methods (e.g., Heckman, Ichimura, Smith, & Todd, 1998; Shadish et al., 2002). These statistical innovations generally depend on untested assumptions, mostly to rule out specification errors (Copas & Li, 1997; Winship & Morgan, 1999). Their complexity distances them further from the data and makes them less straightforward to implement correctly.

Innovations using latent variables or multiple occasions can minimize problems of measurement error, thus satisfying one of the crucial assumptions for making valid causal inferences from nonrandomized designs. But the innovations are only partially adequate for ruling out specification errors (e.g., omitted relevant variables), which is the most crucial assumption for valid causal inferences (Clogg & Haritou, 1997; Freedman, 1991; Kaplan & Berry, 1990; Rosenbaum, 1995; Rubin, 1978; Shadish et al., 2002). Like traditional statistical controls, appropriate implementation of these statistical innovations can generally reduce confounds such as selection bias but can rarely eliminate them completely. Therefore they should be used in conjunction with a strong research design, giving careful consideration to plausible alternative explanations (Rindskopf, 2000; Shadish & Cook, 1999; Shadish et al., 2002).

Conclusion

Selection bias can be a major threat to the validity of casual conclusions about interventions. Preexisting group differences and

their trends must be taken into account before valid conclusions about intervention effects can be made. Although several methodological corrections have been mentioned, the initial and most important step is to be aware of this particular confound. The principle behind most methodological corrections is to rule out plausible alternative interpretations until the only remaining plausible interpretation involves the causal effect of the intervention.

It is impractical, however, for practitioners and policymakers to wait until research has ruled out every conceivable alternative interpretation. It may be helpful to think of the validity of causal inferences as varying along a continuum, according to the number and plausibility of remaining alternative interpretations. Improved discriminations between reasonably valid and more questionable causal inferences can, in turn, provide a stronger basis for recommending certain interventions and increasing priority on the development and refinement of others.

The pervasiveness of the intervention selection bias does not prove that any of the premature conclusions featured in this article are necessarily incorrect. Failure to control for the intervention selection bias, however, prevents research from distinguishing between more or less effective interventions, thus impeding scientific progress on these topics and undermining the confidence with which psychologists can recommend various parental or psychotherapeutic interventions.

Failure to recognize the intervention selection bias can disproportionately affect those clients most in need, as well as those devoted to helping them. Uncontrolled evaluations may be especially biased against interventions to correct more difficult problems. The bias could have dire consequences when third-party payers are attempting to cut costs for expensive cases and improve accountability with simple posttreatment outcome measures (Lyons et al., 1997). Increased recognition and control of the intervention selection bias can only lead to further advancements in psychological research, practice, and policy, especially for corrective interventions designed to address society's most difficult problems.

References

Adams, M. J. (1995). *Youth in crisis: An examination of adverse risk factors affecting children's cognitive and behavioral/emotional development, children ages 10–16.* Unpublished doctoral dissertation, University of Texas at Dallas.

Affleck, G., Zautra, A., Tennen, H., & Armeli, S. (1999). Multilevel daily process designs for consulting and clinical psychology: A preface for the perplexed. *Journal of Counseling and Clinical Psychology, 67,* 746–754.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28,* 369–404.

Anesko, K. M., & O'Leary, S. G. (1982). The effectiveness of brief parent training for the management of children's homework problems. *Child & Family Behavior Therapy, 4*(2–3), 113–126.

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.

Balli, S. J., Wedman, J. F., & Demo, D. H. (1997). Family involvement with middle-grades homework: Effects of differential prompting. *Journal of Experimental Education, 66,* 31–48.

Barter, J. T., Swaback, D. O., & Todd, D. (1968). Adolescent suicide attempts: A follow-up study of hospitalized patients. *Archives of General Psychiatry, 19,* 523–527.

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. (2001, August). *Does causally relevant research support a blanket injunction against disciplinary spanking by parents?* Paper presented at the 109th Annual Convention of the American Psychological Association, San Francisco.

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. (1968). A reinterpretation of the direction of effects in studies of socialization. *Psychological Review, 75,* 81–95.

Bell, R. Q., & Harper, L. V. (1977). *Child effects on adults.* Hillsdale, NJ: Erlbaum.

Berger, K. S., & Thompson, R. A. (1995). *The developing person through childhood and adolescence* (4th ed.). New York: Worth.

Blum, N. J., Williams, G. E., Friman, P. C., & Christophersen, E. R. (1995). Disciplining young children: The role of verbal instructions and reasoning. *Pediatrics, 96,* 336–341.

Borduin, C. M., Henggeler, S. W., Blaske, D. M., & Stein, R. J. (1990). Multisystemic treatment of adolescent sexual offenders. *International Journal of Offender Therapy and Comparative Criminology, 34,* 105–113.

Bornstein, M. H., & Lamb, M. E. (Eds.). (1988). *Developmental psychology: An advanced textbook* (2nd ed.). Hillsdale, NJ: Erlbaum.

Brent, D. A., Kolko, D. J., Wartella, M. E., Boylan, M. B., Moritz, G., Baugher, M., & Zelenak, J. P. (1993). Adolescent psychiatric inpatients' risk of suicide attempt at 6-month follow-up. *Journal of the American Academy of Child & Adolescent Psychiatry, 32,* 95–105.

Brestan, E. V., & Eyberg, S. M. (1998). Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids. *Journal of Clinical Child Psychology, 27,* 180–189.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24,* 409–429.

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.

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

Chamberlain, P. (2000). What works in treatment foster care. In M. P. Kluger, G. Alexander, & P. A. Curtis (Eds.), *What works in child welfare* (pp. 157–162). Washington, DC: CWLA Press.

Chamberlin, T. C. (1965, May 7). The method of multiple working hypotheses. *Science, 148,* 754–759. (Reprinted from *Science,* 1890, February 7, *15,* 92–96)

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

Chen, L. M., Martin, C. M., Keenan, S. P., & Sibbald, W. J. (1998). Patients readmitted to the intensive care unit during the same hospitalization: Clinical features and outcomes. *Critical Care Medicine, 26,* 1834–1841.

Christophersen, E. R. (1988). *Little people: Guidelines for common sense child rearing* (3rd ed.). Kansas City, MO: Westport.

Christophersen, E. R. (1990). *Beyond discipline: Parenting that lasts a lifetime.* Kansas City, MO: Westport.

Clogg, C. C., & Haritou, A. (1997). The regression method of causal inference and a dilemma confronting this method. In V. R. McKim & S. P. Turner (Eds.), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 83–112). Notre Dame, IN: University of Notre Dame Press.

Cohen-Sandler, R., Berman, A. L., & King, R. A. (1982). A follow-up study of hospitalized suicidal children. *Journal of the American Academy of Child Psychiatry, 21,* 398–403.

Copas, J. B., & Li, H. G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society, Series B, 59,* 55–95.

Daley, J., Jencks, S., Draper, D., Lenhart, G., Thomas, N., & Walker, J. (1988). Predicting hospital-associated mortality for Medicare patients. *Journal of the American Medical Association, 260,* 3617–3624.

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.

Desimone, L. (1999). Linking parent involvement with student achievement: Do race and income matter? *Journal of Educational Research, 93,* 11–30.

Dougherty, E. H., & Dougherty, A. (1977). The daily report card: A simplified and flexible package for classroom behavior management. *Psychology in the Schools, 14,* 191–195.

Dowden, C., & Andrews, D. A. (2000). Effective correctional treatment and violent reoffending: A meta-analysis. *Canadian Journal of Criminology, 42,* 449–467.

Eddy, J. M., Dishion, T. J., & Stoolmiller, M. (1998). The analysis of intervention change in children and families: Methodological and conceptual issues embedded in intervention studies. *Journal of Abnormal Child Psychology, 26,* 53–69.

Ellison, C., Musick, M., & Holden, G. (1998). *A longitudinal study of the effects of corporal punishment: The moderating effect of Conservative Protestantism.* Unpublished manuscript, University of Texas at Austin, Department of Sociology.

EPOCH-Worldwide. (2002). *Corporal punishment of children in the family.* Retrieved October 3, 2003, from http://www.stophitting.com/laws/legalReform.php

Escarce, J. J., & Kelly, M. A. (1990). Admission source to the medical intensive care unit predicts hospital death independent of APACHE II score. *Journal of the American Medical Association, 264,* 2389–2394.

Etaugh, C., & Rathus, S. A. (1995). *The world of children.* Orlando, FL: Harcourt Brace.

Faber, J. F. (1982). *Life tables for the United States* (SSA Publication No. 11-11534). Rockville, MD: U.S. Department of Health and Human Services.

Ferrer, E., & McArdle, J. J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Structural Equation Modeling, 10,* 493–524.

Forgatch, M. S., & Ramsey, E. (1994). Boosting homework: A videotape link between families and schools. *School Psychology Review, 23,* 472–484.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics, 12,* 101–128.

Freedman, D. A. (1991). Statistical models and shoe leather. In P. V. Marsden (Ed.), *Sociological methodology* (Vol. 21, pp. 291–313). Oxford, England: Blackwell.

Freedman, D. A. (1997). From association to causation via regression. In V. R. McKim & S. P. Turner (Eds.), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 113–161). Notre Dame, IN: University of Notre Dame Press.

Furby, L., Weinrott, M. R., & Blackshaw, L. (1989). Sex offender recidivism: A review. *Psychological Bulletin, 105,* 3–30.

Garbarino, J. (1996). CAN reflections on 20 years of searching. *Child Abuse & Neglect, 20,* 157–160.

Garfinkel, B. C., Sroese, A., & Hood, J. (1982). Suicide attempts in children and adolescents. *American Journal of Psychiatry, 139,* 1257–1261.

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.

Gibbons, D. C. (1999). Changing lawbreakers: What have we learned since the 1950s? *Crime & Delinquency, 45,* 272–293.

Gibbons, R. D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H. C., Greenhouse, J. B., et al. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Archives of General Psychiatry, 50,* 739–750.

Glazerman, S., Levy, D. M., & Myers, D. (2002). *Nonexperimental replications of social experiments: A systematic review* (MPR Ref. No. 8813-300). Washington, DC: Corporation for the Advancement of Policy Evaluation.

Goldacre, M., & Hawton, K. (1985). Reception of self-poisoning and subsequent death in adolescents who take overdoses. *British Journal of Psychiatry, 146,* 395–398.

Goldberg, J., Merbaum, M., Even, T., Getz, P., & Safir, M. P. (1981). Training mothers in contingency management of school-related behavior. *Journal of General Psychology, 104,* 3–12.

Goldhill, D. R., & Withington, P. S. (1998). Is excess intensive-care mortality in the United Kingdom concealed by ICV mortality prediction models? *Anaesthesia, 53,* 89–90.

Gollub, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development, 58,* 80–92.

Green, J., Passman, L. J., & Wintfield, N. (1991). Analyzing hospital mortality: The consequences of diversity in patient mix. *Journal of the American Medical Association, 265,* 1849–1853.

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.

Grossman, L. S., Martis, B., & Fichtner, C. G. (1999). Are sex offenders treatable? A research overview. *Psychiatric Services, 50,* 349–361.

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, G. C. N. (1995). Sexual offender recidivism revisited: A meta-analysis of recent treatment studies. *Journal of Consulting and Clinical Psychology, 63,* 802–809.

Harris, G. T., Rice, M. E., & Quincey, V. L. (1998). Appraisal and management of risk in sexual aggressors: Implications for criminal justice policy. *Psychology, Public Policy, and Law, 4,* 73–115.

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

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

Heilbrum, K., Nezu, C. M., Keeney, M., Chung, S., & Wasserman, A. L. (1998). Sexual offending: Linking assessment, intervention, and decision making. *Psychology, Public Policy, and Law, 4,* 138–174.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in

experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1,* 154–169.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58,* 295–300.

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.

Holden, G. W. (1998). Affidavit of George W. Holden. In Canadian Foundation for Children Youth and the Law v. Canada (Attorney General), *Ontario Superior Court of Justice,* (2000), 49 O. R. (3d), 662. (Court File no. 98-CV-158948, Applicant's Record, Vol. 5).

Hollin, C. R. (1999). Treatment programs for offenders: Meta-analysis, "what works," and beyond. *International Journal of Law and Psychiatry, 22,* 361–372.

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

Izzo, R. L., & Ross, R. R. (1990). Meta-analysis of rehabilitation programs for juvenile delinquents. *Criminal Justice and Behavior, 17,* 134–142.

Janus, E. S. (2000). Sexual predator commitment laws: Lessons for law and the behavioral sciences. *Behavioral Sciences and the Law, 18,* 5–21.

Kahle, A. L., & Kelley, M. L. (1994). Children's homework problems: A comparison of goal setting and parent training. *Behavior Therapy, 25,* 275–290.

Kahn, K. L., Brook, R. H., Draper, D., Keeler, E. B., Rubenstein, L. V., Rogers, W. H., et al. (1988). Interpreting hospital mortality data: How can we proceed? *Journal of the American Medical Association, 260,* 3625–3628.

Kaplan, R. M., & Berry, C. C. (1990). Adjusting for confounding variables. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (DHHS Publication No. PHS 90-3454, pp. 105–114). Rockville, MD: U.S. Department of Health and Human Services.

Kazdin, A. E. (1995). *Conduct disorders in childhood and adolescence* (2nd ed.). Thousand Oaks, CA: Sage.

Keith, T. Z. (1986). *Homework.* West Lafayette, IN: Kappa Delta Pi.

Kluger, M. P., Alexander, G., & Curtis, P. A. (Eds.). (2000). *What works in child welfare.* Washington, DC: CWLA Press.

Knight, K., Hiller, M. L., & Simpson, D. D. (1999). Evaluating corrections-based treatment for the drug-abusing criminal offender. *Journal of Psychoactive Drugs, 31,* 299–304.

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.

Kochanska, G., & Thompson, R. A. (1997). The emergence and development of conscience in toddlerhood and early childhood. In J. E. Grusec & L. Kuczynski (Eds.), *Parenting and children's internalization of values* (pp. 53–77). New York: Wiley.

Koven, J. T., & LeBow, M. D. (1973). Teaching parents to remediate the academic problems of their children. *Journal of Experimental Education, 41*(4), 64–73.

Kuczynski, L., & Hildebrandt, N. (1997). Models of conformity and resistance in socialization theory. In J. E. Grusec & L. Kuczynski (Eds.), *Parenting and the internalization of values: A handbook of contemporary theory* (pp. 227–256). New York: Wiley.

Kuperman, S., Black, D. W., & Burns, T. L. (1988). Excess suicide among formerly hospitalized child psychiatry patients. *Journal of Clinical Psychiatry, 49,* 88–93.

Kutash, K., & Rivera, V. R. (1996). *What works in children's mental health services?* Baltimore: Brookes.

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, Tampa, FL.

Larzelere, R. E. (1996). A review of the outcomes of parental use of nonabusive or customary physical punishment. *Pediatrics, 98,* 824–828.

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., 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., & Skeen, J. H. (1984). The method of multiple hypotheses: A neglected research strategy in family studies. *Journal of Family Issues, 5,* 474–492.

Larzelere, R. E., & Smith, G. L. (2000, August). *Controlled longitudinal effects of five disciplinary tactics on antisocial behavior.* Paper presented at the 108th Annual Convention of the American Psychological Association, Washington, DC.

Larzelere, R. E., Smith, G. L., Batenhorst, L. M., & Kelly, D. B. (1996). Predictive validity of the Suicide Probability Scale among adolescents in group home treatment. *Journal of the American Academy of Child & Adolescent Psychiatry, 35,* 166–173.

LeGall, J. R., Lemeshow, S., & Saulnier, F. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association, 270,* 2957–2963.

Lemeshow, S., Teres, D., Klar, K., Avrunin, J. S., Gehlbach, S. H., & Rapoport, J. (1993). Mortality Probability Models (MPMII) based on an international cohort of intensive care unit patients. *Journal of the American Medical Association, 270,* 2478–2486.

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.

Linehan, M. M. (1997). Behavioral treatments of suicidal behaviors: Definitional obfuscation and treatment outcomes. In D. M. Stoff & J. J. Mann (Eds.), *Annals of the New York Academy of Sciences: Vol. 836. Neurobiology of suicide: From the bench to the clinic* (pp. 302–328). New York: New York Academy of Sciences.

Lipsey, M. W. (1999). Can intervention rehabilitate serious delinquents? *Annals of the American Academy of Political and Social Science, 564,* 142–166.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1211.

Lipsey, M. W., & Wilson, D. B. (1998). Effective intervention for serious juvenile offenders: A synthesis of research. In R. Loeber & D. P. Farrington (Eds.), *Serious & violent juvenile offenders: Risk factors and successful interventions* (pp. 313–345). Thousand Oaks, CA: Sage.

Littell, J. H., & Schuerman, J. R. (1995). *A synthesis of research on family preservation and family reunification programs* (Report to the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services). Retrieved October 3, 2003, from http://aspe.hhs.gov/hsp/cyp/FPLITREV.HTM

Loitz, P. A., & Kratochwill, T. R. (1995). Evaluation of a self-help manual for children's homework problems. *School Psychology International, 16,* 389–396.

Lyons, J. S., O'Mahoney, M. T., Miller, S. I., Neme, J., Kabat, J., & Miller, F. (1997). Predicting readmission to the psychiatric hospital in a managed care environment: Implications for quality indicators. *American Journal of Psychiatry, 154,* 337–340.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201–226.

Maertens, N. W., & Johnston, J. (1972). Effects of arithmetic homework upon the attitudes and achievement of fourth, fifth, and sixth grade pupils. *School Science and Mathematics, 72,* 117–126.

Magidson, J. (1977). Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation: A general alternative to ANCOVA. *Evaluation Quarterly, 1,* 399–420.

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.

Maletzky, B. M. (1997). Editor's note. *Sexual Abuse: A Journal of Research and Treatment, 9,* 147.

Marques, J. K. (1999). How to answer the question "Does sex offender treatment work?" *Journal of Interpersonal Violence, 14,* 437–451.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58,* 110–133.

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 137–176). Washington, DC: American Psychological Association.

McDermott, R. P., Goldman, S. V., & Varenne, H. (1984). When school goes home: Some problems in the organization of homework. *Teachers College Record, 85,* 391–409.

McLeod, J. D., Kruttschnitt, C., & Dornfeld, M. (1994). Does parenting explain the effects of structural conditions on children's antisocial behavior? A comparison of Blacks and Whites. *Social Forces, 73,* 575–604.

Mihalic, S., Irwin, K., Elliott, D., Fagan, A., & Hansen, D. (2001, July). Blueprints for violence prevention (Office of Juvenile Justice and Delinquency Prevention Publication No. NCJ 187079). *Juvenile Justice Bulletin.* Retrieved October 3, 2003, from http://www.ncjrs.org/html/ojjdp/jjbul2001_7_3/contents.html

Miller, D. L., & Kelley, M. L. (1991). Interventions for improving homework performance: A critical review. *School Psychology Quarterly, 6,* 174–185.

Morton, R. F., Heber, J. R., & McCarter, R. J. (1996). *A study guide to epidemiology and biostatistics* (4th ed.). Gaithersburg, MD: Aspen.

Motto, J. A. (1984). Suicide in male adolescents. In H. S. Sudak, A. B. Ford, & N. B. Rushforth (Eds.), *Suicide in the young* (pp. 227–244). Boston: Wright.

Myers, D. (2004). *Psychology* (7th ed.). New York: Worth.

National Institutes of Health. (1991). *Treatment of destructive behaviors in persons with developmental disabilities* (NIH Publication No. 91-2410). Rockville, MD: Author.

Nelson, K. (2000). What works in family preservation services. In M. P. Kluger, G. Alexander, & P. A. Curtis (Eds.), *What works in child welfare* (pp. 11–22). Washington, DC: CWLA Press.

Newhouse, J. P., & McClellan, M. (1998). Econometrics in outcomes research: The use of instrumental variables. *Annual Review of Public Health, 19,* 17–34.

Osgood, D. W., & Smith, G. L. (1995). Applying hierarchical linear modeling to extended longitudinal evaluations: The Boys Town Follow-Up Study. *Evaluation Review, 19,* 3–38.

Otto, U. (1972). Suicidal acts by children and adolescents: A follow-up study. *Acta Psychiatrica Scandinavica, 48*(1, Suppl. 233).

Patterson, G. R. (1982). *Coercive family process.* Eugene, OR: Castalia Press.

Pearson, F. S., & Lipton, D. S. (1999). A meta-analytic review of the effectiveness of corrections-based treatments for drug abuse. *Prison Journal, 79,* 384–410.

Pfeffer, C. R., Klerman, G. L., Hurt, S. W., Lesser, M., Peskin, J. R., & Siefker, C. A. (1991). Suicidal children grow up: Demographic and clinical risk factors for adolescent suicide attempts. *Journal of the American Academy of Child & Adolescent Psychiatry, 30,* 609–616.

Platt, J. R. (1964, October 16). Strong inference. *Science, 146,* 347–353.

Polizzi, D. M., MacKenzie, D. L., & Hickman, L. J. (1999). What works in adult sex offender treatment? A review of prison- and non-prison-based treatment programs. *International Journal of Offender Therapy and Comparative Criminology, 43,* 357–374.

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

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books. (Original work published 1935)

Randolph, A. G., Guyatt, G. H., & Carlet, J. (1998). Understanding articles comparing outcomes among intensive care units to rate quality of care. *Critical Care Medicine, 22,* 773–781.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Redondo, S., Sanchez-Meca, J., & Garrido, V. (1999). The influence of treatment programmes on the recidivism of juvenile and adult offenses: An European meta-analytic review. *Psychology, Crime & Law, 5,* 251–278.

Rhoades, M. M., & Kratochwill, T. R. (1998). Parent training and consultation: An analysis of a homework intervention program. *School Psychology Quarterly, 13,* 241–264.

Rindskopf, D. (2000). Plausible rival hypotheses in measurement, design, and scientific theory. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 1–12). Thousand Oaks, CA: Sage.

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.

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

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rothman, K. J., & Greenland, S. (1998a). Hill's criteria for causality. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (Vol. 3, pp. 1920–1924). New York: Wiley.

Rothman, K. J., & Greenland, S. (1998b). *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.

Rubin, D. B. (1990). Formal models of statistical inference for causal effects. *Journal of Statistical Planning and Inference, 25,* 279–292.

Sackett, D. L., Haynes, R. B., Guyatt, R. B., & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Boston: Little, Brown.

Schuessler, K. F. (Ed.). (1978). *Sociological methodology: 1979.* San Francisco: Jossey-Bass.

Schuster, D. P., & Kollef, M. H. (1994). Predicting intensive care unit outcomes [Special issue]. *Critical Care Clinics, 10*(1).

Shadish, W. R., & Cook, T. D. (1999). Design rules: More steps towards a complete theory of quasi-experimentation. *Statistical Science, 14,* 294–300.

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., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin, 126,* 512–529.

Shaffer, D., Garland, A., Gould, M., Fisher, P., & Trautman, P. (1988). Preventing teenage suicide: A critical review. *Journal of the American Academy of Child & Adolescent Psychiatry, 27,* 675–687.

Shafii, M., Carrigan, S., Whittinghill, J. R., & Derrick, A. (1985). Psychological autopsy of completed suicide in children and adolescents. *American Journal of Psychiatry, 142,* 1061–1064.

Shaw, J. A., & Work Group on Quality Issues. (1999). Practice parameters for the assessment and treatment of children and adolescents who are sexually abusive of others. *Journal of the American Academy of Child & Adolescent Psychiatry, 38,* 55S–76S.

Silber, J. H., & Rosenbaum, P. R. (1997). A spurious correlation between hospital mortality and complication rates: The importance of severity adjustment. *Medical Care, 35,* OS77–OS92.

Simon, L. M. J. (1998). Does criminal offender treatment work? *Applied & Preventive Psychology, 7,* 137–159.

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.

Singh, K., Bickley, P. G., Trivette, P., Keith, T. Z., Keith, P. B., & Anderson, E. (1995). The effects of four components of parental involvement on eighth grade student achievement: Structural analysis of NELS-88 data. *School Psychology Review, 24,* 299–317.

Stanton, M. D., & Shadish, W. R. (1997). Outcome, attrition, and family–couples treatment for drug abuse: A meta-analysis and review of the controlled, comparative studies. *Psychological Bulletin, 122,* 170–191.

Stoolmiller, M. (1995). Using latent growth curve models to study developmental processes. In J. M. Gottman (Ed.), *The analysis of change* (pp. 105–138). Hillsdale, NJ: Erlbaum.

Straus, M. A. (1999). Is it time to ban corporal punishment of children? *Canadian Medical Association Journal, 161,* 821–822.

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., 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.

Thakur, N. M., Hoff, R. A., Druss, B., & Catalanotto, J. (1998). Using recidivism rates as a quality indicator for substance abuse treatment programs. *Psychiatric Services, 49,* 1347–1350.

Thomas, J. W., & Hofer, T. P. (1998). Research evidence on the validity of risk-adjusted mortality rate as a measure of hospital quality of care. *Medical Care Research & Review, 55,* 371–404.

Turner, S. (1997). "Net effects": A short history. In V. R. McKim & S. P. Turner (Eds.), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 23–45). Notre Dame, IN: University of Notre Dame Press.

Walters, G. C., & Grusec, J. E. (1977). *Punishment.* San Francisco: Freeman.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* Cambridge, MA: Harvard University Press.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). Cambridge, England: Cambridge University Press.

Westat, Inc., Chapin Hall Center for Children, & James Bell Associates. (2001, January 8). *Evaluation of family preservation and reunification programs: Interim report.* Retrieved October 3, 2003, from the U.S. Department of Health and Human Services Web site: http://aspe.hhs.gov/hsp/fampres94/index.htm

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. 1–2). Athens, OH: Authors.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25,* 659–707.

Wood, R. M., Grossman, L. S., & Fichtner, C. G. (2000). Psychological assessment, treatment, and outcome with sex offenders. *Behavioral Sciences and the Law, 18,* 22–41.

Wu, P., & Campbell, D. T. (1996). Extending latent variable LISREL analyses of the 1969 Westinghouse Head Start evaluation to Blacks and full year Whites. *Evaluation & Program Planning, 19*(3), 183–191.

Zimmerman, J. E., Wagner, D. P., Draper, E. A., Wright, L., Alzola, C., & Knaus, W. A. (1998). Evaluation of Acute Physiology and Chronic Health Evaluation III predictions of hospital mortality in an independent database. *Critical Care Medicine, 26,* 1317–1326.

Received September 15, 2000
Revision received August 6, 2003
Accepted September 28, 2003 ■