Proceeding From Observed Correlation to Causal Inference

The Use of Natural Experiments

Michael Rutter

Institute of Psychiatry and King's College, London, United Kingdom

ABSTRACT—*This article notes five reasons why a correlation between a risk (or protective) factor and some specified outcome might not reflect environmental causation. In keeping with numerous other writers, it is noted that a causal effect is usually composed of a constellation of components acting in concert. The study of causation, therefore, will necessarily be informative on only one or more subsets of such components. There is no such thing as a single basic necessary and sufficient cause. Attention is drawn to the need (albeit unobservable) to consider the counterfactual (i.e., what would have happened if the individual had not had the supposed risk experience). Fifteen possible types of natural experiments that may be used to test causal inferences with respect to naturally occurring prior causes (rather than planned interventions) are described. These comprise five types of genetically sensitive designs intended to control for possible genetic mediation (as well as dealing with other issues), six uses of twin or adoptee strategies to deal with other issues such as selection bias or the contrasts between different environmental risks, two designs to deal with selection bias, regression discontinuity designs to take into account unmeasured confounders, and the study of contextual effects. It is concluded that, taken in conjunction, natural experiments can be very helpful in both strengthening and weakening causal inferences.*

Address correspondence to Michael Rutter, P.O. Box 80, MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London, United Kingdom SE5 8AF; e-mail: j.wickham@iop.kcl.ac.uk.

From an early point in their training, all behavioral scientists are taught that statistically significant correlations do not necessarily mean any kind of causative effect. Nevertheless, the literature is full of studies with findings that are exclusively based on correlational evidence. Researchers tend to fall into one of two camps with respect to how they react to the problem. First, there are those who are careful to use language that avoids any direct claim for causation, and yet, in the discussion section of their papers, they imply that the findings do indeed mean causation. Second, there are those that completely accept the inability to make a causal inference on the basis of simple correlation or association and, instead, take refuge in the claim that they are studying only associations and not causation. This second, "pure" approach sounds safer, but it is disingenuous because it is difficult to see why anyone would be interested in statistical associations or correlations if the findings were not in some way relevant to an understanding of causative mechanisms.

Some researchers argue that only laboratory experiments or randomized controlled trials (RCTs) allow any firm causal inference. Of course, it is true that both provide a much needed control that allows rigorous testing of the causal hypothesis and takes account of unmeasured confounders. Nevertheless, there are many risk factors for which neither laboratory experiments nor RCTs are feasible or ethical. That would apply, for example, to experiences such as maltreatment, life stresses, or child neglect. How should these be studied in order to consider causal effects? That question constitutes the prime focus of this article. In short, the focus is not on planned interventions that could be the subject of RCTs, but rather on the many naturally occurring risk and protective experiences that could not be tested in that way for either practical or ethical reasons. The focus is also strictly on environmental factors that might have a causal effect (as these have a major public health importance) rather than on

the broader range of causal inferences that are not open to manipulation.[1]

WHAT IS MEANT BY *CAUSE*

Before proceeding with a discussion of how natural experiments (using design elements that provide some kind of approximation to experimental conditions) may help in testing causal inferences, it is necessary to start with a consideration of what is meant by *a cause*. When considering either psychopathology or psychological traits, it is essential to appreciate that multifactorial causation is the rule. There are very few, if any, simple direct determinative causal effects on any outcome. MacMahon, Pugh, and Ipsen (1960) proposed the metaphor of a web of causation. Rothman and Greenland (2002) similarly argued that any causal effect is composed of a constellation of components acting in concert. With very few exceptions, there is no such thing as a single necessary and sufficient cause. It is not just that multifactorial traits or disorders have multiple causal inferences, but also that several different causal pathways may all lead to the same endpoint (Rutter, 1997). Furthermore, almost all causal pathways involve several different phases. For example, the pathway to the psychological or psychopathological end point does not begin with a causal risk factor, it must be preceded by the pathway leading to exposure to the risk factor (Rutter, Champion, Quinton, Maughan, & Pickles, 1995). Thus, if the hypothesized cause is, say, maternal binge drinking during pregnancy or paternal physical abuse in early childhood, what led to the parent behaving in this way? The answer may lie in either societal influences or individual propensities or often a complex mixture of both. If that were not complicated enough, it is also the case that influences are often reciprocal, with the causal arrow running in both directions (see e.g., Engfer, Walper, & Rutter, 1994).

For all these (and other) reasons, there is no point in seeking to identify the single cause of any outcome, because there is no such thing. This appreciation led Mackie (1965, 1974) to refer to causes that are insufficient but necessary components of unnecessary but sufficient causes—what most epidemiologists now refer to as *INUS* (Schwartz & Susser, 2006). What this complicated sounding but actually very simple and straightforward concept means is that the overall causal nexus in its entirety is a sufficient cause of the condition being considered—that is, it is enough to cause the outcome without the operation of any other influences. On the other hand, it is unnecessary because it represents only one of several possible causal pathways. Conversely, INUS are necessary because, if all other components are held constant, the outcome will not occur or will not occur at that time. Mackie argued that it was not strictly necessary for a mediating mechanism to be identified in order to determine causation, but nevertheless, such a mechanism was implied and the causal inference is likely to be stronger if there is evidence of such a mechanism. As Robins and Greenland (1996) pointed out, there is the implication that changing a causal factor will actually reduce the population's burden of disorder (if the outcome being investigated concerns a disorder).

HOW MAY A CAUSE BE IDENTIFIED?

All causal reasoning requires an implicit comparison of what actually happened when an individual experienced the supposed causal influence with what would have happened if simultaneously they had not had that experience. Even in a controlled experiment, that observation can never be made. Accordingly, researchers have had recourse to counterfactual reasoning in thinking about causes (i.e., making the thought experiment of what the counterfactual situation might be). Previous researchers have repeatedly pointed out that such reasoning is necessarily subjective even when robust techniques can be used to quantify key features (see Hernán, Herndandez-Diaz, & Robins, 2004; Mackie, 1974; Maldonado & Greenland, 2002; Rubin, 1986).

The first systematic analysis of a causal relationship was provided by the philosopher John Stuart Mill (1843), who argued that three fundamental conditions had to be met: (a) the cause had to precede the effect, (b) the cause had to be statistically associated with the effect, and (c) there had to be no plausible alternative explanation for the effect other than the cause. Hill (1965), in a now classic article, outlined a set of guidelines (not rules) that may be used to consider the possible validity of a causal inference. As Cochran and Chambers (1965) pointed out, this need to consider alternative explanatory hypotheses is the most critical in many ways, but it is often missing from environmental research. Simply providing supporting evidence that might bolster the causal inference should never be acceptable (Shavelson & Towne, 2002).

ALTERNATIVE EXPLANATIONS

In addition, it is necessary to consider some of the key possibilities that must be examined. For present purposes, attention is drawn to the five possibilities that have played the greatest role in the development of natural experiments. First, there is the possibility of genetic mediation of the risk stemming from an environmental feature. This possibility arises because of the

[1] Although this article focuses on the use of natural experiments to test causal inferences with respect to environmental factors that might influence psychopathological outcomes, the strategies considered here apply much more broadly to psychological research as a whole when it involves (as usually it does) the need to examine and test causal inferences. Thus, not only can the strategies be employed (as they sometimes have been) in relation to positive outcomes such as resilience or well-being, but they can also apply to personal attributes (such as temperament or ethnicity; see Rutter & Tienda, 2005) or supposedly "fixed" traits such as biological sex (see Rutter, Caspi, & Moffitt, 2003). The focus here has been largely on epidemiological approaches, but, with brain-imaging findings, there are similar needs to differentiate between neural effects on the workings of the mind and experiential effects on the structure and functioning of the brain (see Frith & Frith, in press; Peterson, 2003).

existence of both passive and active gene–environment correlations (Plomin, DeFries, & Loehlin, 1977). "Passive" in this instance refers to the fact that parents who pass on their risky genes also have a tendency to create risk environments. This is evident, for example, in the rearing provided by parents with a mental disorder (Rutter et al., 1997). "Active" in this case refers to the tendency for children (and adults) to shape and select their environments. As a consequence, there is evidence of genetic influences on individual variations within a range of different types of risk exposure (Kendler & Baker, 2007). To deal with this possibility, researchers have devised a range of genetically sensitive designs to test environmental-risk mediation hypotheses.[2]

Second, there is the possibility of social selection or allocation bias (Caspi, 2003), which suggests that the association between the risk experience and the outcome reflects the origin of the putative risk factor rather than its effects. For example, is the increased risk of a variety of psychopathological outcomes for children born to a teenage mother due to the sort of person likely to have a child during adolescence, or is it due to the rearing or socialization provided by such a mother (Maynard, 1997; Moffitt & The E-Risk Study Team, 2002)? Similarly, does the association between low socioeconomic status (SES) and some forms of psychopathology reflect a causal impact of the former on the latter, or does it reflect the qualities of individuals who land up with low SES (Miech, Caspi, Moffitt, Entner Wright, & Silva, 1999)? These concerns have led to a variety of designs using twins or adoptees to determine the possibility of social selection (or one of the other alternative possibilities). In addition, researchers have used designs capitalizing on circumstances in which total populations have been exposed (without the opportunity of choice or selection) to environments that provide or remove risk.

Third, there is the possibility of reverse causation. In a now classic article, Bell (1968) argued persuasively that some of the supposed effects of socialization experiences might reflect the effects of children on their social environments (meaning parents, teachers, etc.) rather than the effects of the children's rearing environment on them. Thomas, Chess, and Birch (1968) argued much the same from their study of children's temperamental qualities. The reality of such reverse effects has been shown subsequently in both experimental (Anderson, Lytton & Romney, 1986) and naturalistic (Bell & Harper, 1977) studies and in studies of the effects of children on their adoptive parents (Ge et al., 1998; O'Connor, Deater-Deckard, Fulker, Rutter, & Plomin, 1998). The notion of reverse causation has also been extended to associations between one form of psychopathology on another type of disorder (McGue & Iacono, 2005).

Fourth, there is a possibility that the risk feature has been misidentified. For example, in the past it was supposed that "broken homes" were a major cause of antisocial behavior (see Wootton, 1959, for a critique). However, homes may be broken for many different reasons, and research has shown that the main risk stems from either family discord or poor parenting rather than from family change or break-up, as such (Brown, Harris, & Bifulco, 1986; Fergusson, Horwood, & Lynskey, 1992). Designs using twins or adoptees have been used to separate prenatal from postnatal effects and to differentiate among different possible sources of environmental risk.

Fifth, there is the ever-present concern that the association reflects not causation, but some confounding variable (meaning any variable that both differentiates between the groups to be compared and is statistically associated with the outcome being investigated). The general statistical strategies for dealing with possible confounding variables are discussed in Rosenbaum (2002); Rothman and Greenland (1998); Shadish, Cook, and Campbell (2002); and Susser, Schwartz, Morabia, and Bromet (2006) and are not dealt with here. Equally, this article does not deal with the inappropriateness of controlling for factors that are intermediate on the pathway between exposure to the risk factor and the outcome, or those that are caused in part by exposure to the risk experience (see Weinberg, 1993, for a discussion of this issue).

There are five points that warrant mention, however. First, Robins (2001) has argued for the crucial importance of both study design and knowledge of background subject matter to provide leads on how to approach the issue of confounding. Rather than rely on a general controlling for confounders, he suggests that substantially greater leverage may be obtained by undertaking hypothesis-driven analyses focusing on alternative pathways. He has also argued that causal graphs spelling out the implications of the background knowledge can lead to statistical modeling that could go a substantial way to increase or decrease the likelihood of the causal inference being correct (Gill & Robins, 2001; Robins, 2001; Robins, Hernán, & Brumback, 2000). Second, Rosenbaum and Rubin, (1983a, 1983b) have advocated the use of propensity scores to equate groups on the basis of likelihood of exposure to risk rather than controlling for differences in risk. If such scores are to do their job adequately, it is crucial to include all relevant covariates that are found to predict the risk exposure under investigation and, in particular, to include covariates that could lead to allocation bias. The very limited findings on how propensity scores compare

[2]Twin and adoptee studies have been dismissed by a few critics because of ideological concerns, criticism of the concept of heritability, and methodological weaknesses of some of the earlier (and, to a lesser extent, more recent) research (see Joseph, 2003; Kamin & Goldberger, 2002). The issues have been discussed in Rutter (2006). It should be accepted (for example) that early researchers have not always questioned the assumptions of the twin design, that there has often been a failure to appreciate the consequences of the restriction in environmental range in adoptive families, and that too little attention has been paid to possible sampling biases. The critics, however, have been equally guilty of selective attention and reporting. Any dispassionate critic would have to conclude that the best researchers in modern behavioral genetics have dealt with the concerns in a rigorous fashion, resulting in considerable confidence in the validity of twin and adoptee designs when employed with all necessary regard to the many methodological issues involved (see Kendler, 2005; Kendler & Prescott, 2006; Rutter, 2006). The researchers using the varied twin and adoptee designs in the natural experiments discussed in this article have paid due regard to these essential issues.

with randomization are reassuring (Luellen, Shadish, & Clark, 2005; Shadish, Luellen, & Clark, 2006), but their main value lies in their indication of nonoverlapping areas that might best be omitted from analyses (Rubin, 1979). Third, sensitivity analyses to quantify how strong a confounder would have to be to overturn a causal inference (see Cornfield et al., 1959; Rosenbaum & Rubin, 1983b; Sampson, Laub, & Wimer, 2006) may be very helpful. Fourth, diverse strategies and diverse samples with different patterns of confounders are very helpful. The solution does not rely solely on replication. If replications include the same biases, the result will be simply be confirmation of an incorrect inference (Rosenbaum, 2002). This is where natural experiments are particularly helpful, as they separate variables that ordinarily go together. Fifth, none of these approaches deal with unmeasured confounders. There is one quasi-experimental design (regressive discontinuity) that takes unmeasured confounders into account, but it has a very limited applicability to naturally occurring causes (rather than planned interventions).

NATURAL EXPERIMENTS

Cook and Campbell (1979) and Shadish et al. (2002) pioneered the use of quasi-experiments or natural experiments. Their work provides a detailed account of a wide range of strategies, including some designed to deal with planned interventions (which are outside the scope of this article). They emphasized that in the interplay between design and statistical analysis, design rules (Shadish & Cook, 1999). Following their lead, this article deals with design issues. Each of these design issues has crucial statistical implications, but these are deliberately left outside the scope of this article. Readers interested in using any of the designs should consult the cited articles for details of the statistical handling of the concepts. Three strategic considerations shaped the structure of this article. First, as noted, the natural experiments described here exclude those dealing with planned treatments. Second, the many strategies well described by Shadish et al. (2002) are not repeated here unless there is a new point to be introduced. Rather, the emphasis is on the use of twin and adoptee designs for a range of purposes, including several that are not directly concerned with the exclusion of genetic mediation, new examples of the use of total population risks (or removal of risks), and somewhat new approaches in longitudinal studies. Third, rather than provide brief mention of the many examples of the use of natural experiments, one or two examples are described in some detail to provide a clearer picture of the strategy. In brief, the key design requirement in all cases is for samples and/or circumstances that serve to separate variables that ordinarily go together and thereby provide manipulation, and not just analysis, of crucial variables. Each of the designs does so in a slightly different way, and their discussion seeks to highlight both their strengths and their limitations (see also Rutter, Pickles, Murray, & Eaves, 2001).

GENETICALLY SENSITIVE STRATEGIES

Multivariate Twin Designs

Twin designs use the difference between monozygotic (MZ) and dizygotic (DZ) twins (the former share 100% of their segregating genes and the latter share just 50%) to partition the population variance into additive genetic and shared and nonshared environmental effects ("shared" meaning that the effect makes the twins similar, and "unshared" meaning that the effect makes them dissimilar). By treating some postulated environmental risk factor as a trait or phenotype, it is possible to estimate the heritability of the risk factor. By means of cross-twin, cross-trait analyses (the cross being between the risk factor and the outcome being investigated), it is further possible to determine the extent to which the risk effect is genetically or environmentally mediated. If the sample is studied longitudinally, this provides additional research leverage.

Thus, Kendler and Prescott (2006) used multivariate twin analyses to test whether, through an environmental effect, early drinking of alcohol predisposed a person to later alcoholism. The competing hypothesis was that both were the result of the same shared predisposing genetic liability. Multivariate twin analyses clearly showed that there was a substantial shared genetic liability between the age at which an individual started drinking and the later development of alcoholism. By contrast, there was no evidence that the age at which a person first started drinking alcohol had any environmentally mediated causal effect on the likelihood of developing alcoholism (Kendler & Prescott, 2006). Multivariate analyses also showed that the genetic liability for one substance-use disorder tended to be shared with other substance-use disorders, suggesting that much of the genetic liability concerned a general tendency to use substances rather than a specific pharmacological response to any one substance. The genetic liability also involved antisocial personality disorder.

Multivariate twin analyses are limited to some extent by the confounding of shared environmental and nonadditive genetic effects (leading to an underestimate of shared environmental influences on risk) and by the confounding of genetic and Genetic × Shared Environment (G × E) interaction effects (so that the supposed "pure" genetic effect includes the G × E interaction). It is possible to model this explicitly, but the statistical power to do so will be low (Heath, Lynskey, & Waldron, in press).

Jaffee et al.'s (2004) longitudinal twin study of the effects of corporal punishment and physical abuse on children's propensity to engage in antisocial behavior provides another example of the strength of multivariate twin analyses. The starting point was a significant association between both parental punishment and abuse and the psychopathological outcome. The twin designs, using the MZ–DZ contrast, showed that there was a substantial (73%) heritability of antisocial behavior, a modest heritability (25%) of corporal punishment, and very little heritability of physical maltreatment. Cross-twin, cross-trait ana-

lyses indicated that, in the case of corporal punishment, the mediation was mainly genetic (85%), whereas it was mainly environmental for physical abuse (94%). Taking the findings as a whole, the implication was that, to a substantial extent, the corporal punishment was a response to, rather than a cause of, the children's behavior. By contrast, this did not appear to be the case with physical abuse. On the other hand, the longitudinal data suggested that the regular use of corporal punishment tended to escalate to physical abuse and that this escalation effect was environmentally mediated. The study constitutes a neat example of how multivariate genetic analyses can do much to test hypotheses about environmentally mediated risks, with the finding that the mechanisms involved in two apparently similar behaviors (corporal punishment and physical abuse) may be quite different.

The Children of Twins (CoT) Design

An ingenious extension of the twin design is provided by the CoT strategy (D'Onofrio et al., 2003; Silberg & Eaves, 2004). The rationale is that the offspring of adult MZ twins will be social cousins, but biologically, they are genetic half-siblings (one parent is part of an MZ pair, and the other parent is not). By contrast, the offspring of DZ pairs are cousins both genetically and socially. Expressed very simplistically, if the parental feature involves a genetically mediated risk, the risk should apply as much to the offspring of the MZ cotwin as to the children reared by the MZ twin with the parental feature being examined. By contrast, this will not apply to the same extent to the offspring of DZ twins (because they share only half their genes, whereas MZ twins share 100%). If, however, the risk is environmentally mediated, there will not be the same MZ–DZ differences. Hierarchical linear modeling is needed to separate genetic and environmental effects while at the same time controlling for possible confounds, but the concept of the design lies in the comparison between half-siblings and cousins. The strategy requires a very large sample because of the small genetic difference between cousins and half-siblings and because of the need to take account of a range of possible confounding variables. It is also constrained by the fact that measures are likely to be available only for the "twin" parent and not for the "nontwin" parent who is the marital partner. Nevertheless, it provides a potentially very informative natural experiment when there is discordance between the experiences provided by the biological mother (or father) and those provided by her (or his) cotwin. Thus, it has been found that it is likely that parental divorce has a causal role with respect to an earlier initiation of sexual intercourse and a greater likelihood of emotional difficulties (D'Onofrio et al., 2006), but that the statistical effect on earlier initiation of drug use was explained by social selection. Similarly, harsh forms of physical punishment appear to have an environmentally mediated influence on both disruptive behavior and drug or alcohol use (Lynch, Turkheimer, D'Onofrio, Mendle, & Emery, 2006). Caution is needed, however, when the putative environmental risk factor involves both parents (as with divorce) and not just the effect of the behavior of one parent (see Eaves, Maes, & Silberg, 2005).

Use of Discordant Twin and Sibling Pairs

Discordant same-sex twin pairs reared together provide an alternative way of tackling similar causal questions. The rationale is that by examining pairs that are discordant both for the hypothesized environmental risk factor and for the outcome being investigated, a substantial degree of control for both genetic and environmental liabilities that are shared by the twins is provided (Kendler & Prescott, 2006). For example, Prescott and Kendler (1999) used the epidemiological Virginia Adult Twin Study to determine whether the widely observed association between early use of alcohol and later alcoholism (Grant & Dawson, 1997) reflected a causal effect of the former on the latter. In the sample as a whole, confirming earlier findings, there was an odds ratio of 2 in males and 4 in females for drinking before age 15 and later alcohol dependence. By sharp contrast, there was an odds ratio of just 1 in discordant MZ twin pairs. The findings in DZ twin pairs were intermediate. Statistical modeling indicated that the association between early use of alcohol and later alcohol dependence was largely a function of a shared genetic liability (but with some effect from shared environmental factors). The strong implication is that the substantial odds ratios in general population samples did not reflect an environmentally mediated causal effect. The discordant twin study is limited by reliance on self-report data, but methodological checks suggested that it is unlikely that reporting bias accounted for the findings.

Lynskey et al. (2003) and Lynskey, Vink, and Boomsma, (2006) used the same design, first using the Australian Twin Register and then the Netherlands Twin Register to examine the so-called "gateway hypothesis" that early use of cannabis led, by some form of causal mechanism, to an elevated risk of other illicit drug use. Retrospective reporting at age 30 was required in order to obtain assessments of lifetime use. In the Australian sample, the twin with early cannabis use had a two- to fourfold increase in other drug use, the increase being affected only slightly when controlling for possible confounders. The findings from the Dutch study were very similar. The odds ratios for MZ and DZ pairs were similar, but the numbers were too small for a full control of genetic influences. Contrary to the early alcohol study, the findings of both these studies confirmed the likely environmentally mediated causal gateway effect. The mechanisms involved, however, remain quite uncertain. They could reflect attitudes induced by early pleasurable cannabis effects, or they could reflect a social effect operating by means of cannabis use both providing an entry to groups using other drugs and bringing about contact with drug dealers. In this case, the discordant twins design confirmed the earlier epidemiological/

longitudinal study findings (Fergusson & Horwood, 2000). The causal inference would have been stronger if it had been possible to restrict the sample to MZ pairs, and it would have been better to have had prospective data, but the studies clearly point to the value of this design in both confirming and disconfirming causal hypotheses.

A longitudinal study by Caspi et al. (2004) illustrates how a combination of other design features can greatly increase the power of the discordant MZ twin design to test for causal effects. The risk feature being investigated was maternal negative emotion focused on one twin. This was assessed when the twins were aged 5 years using standardized ratings based on audiotaped interviews with mothers. Outcome was assessed 2 years later using teacher reports. The analysis took into account the children's behavior at age 5 and examined the prospective effect on the behavior at age 7. That is, within-individual change in behavior was the focus of the analysis, and the use of different informants for the independent and dependent variables obviated the problem of halo effects in measurement and requiring that the effect be evident across time periods. The contrast between twins within MZ pairs meant that it was highly unlikely that any family-wide social confound was operative. The prospective strategy similarly ruled out any mediating effect of the child's behavior on the mother's emotional expression. The main limitation, as with any other quasi-experimental design, is the uncertainty as to whether the psychopathological risk stemmed from the negative expressed emotion per se or from some associated environmental feature impacting one twin in the MZ pair. Also, the findings leave open the question of what set of features led the mother to respond so differently to the two twins.

Discordant twin pairs cannot be used to examine prenatal risk effects, but discordant sibling pairs can serve the same purpose. Because meiosis results in a random allocation of parental genes across siblings, there is a good control for genetic liability, except in a situation in which there is a shared genetic liability between the discordance in the risk factor and individual differences in the outcome being considered. Thus, D'Onofrio et al. (in press) used the U.S. National Longitudinal Study of Youth to examine the possible effects of mothers' smoking during the pregnancy on externalizing behaviors in the offspring by contrasting pregnancies when the mother smoked with those in which she did not. The measure of mothers' smoking in the pregnancy was obtained soon after the child's birth, and the children's behavior was assessed by repeated maternal reports in middle childhood. Some caution is needed because of the reliance on maternal report for both the independent and dependent variables. Uncertainty is also presented by the limitations of evidence on differences between the mothers who smoked in both pregnancies and the mothers who smoked only in one. Nevertheless, it is striking that the discordant sibling design suggested no environmentally mediated causal effect of smoking in pregnancy on externalizing behavior, whereas structural equation modeling seeking to control for confounders had suggested a causal effect. The implication would seem to be that relying on measured covariates to account for confounding may well give rise to false positive conclusions. The findings showed that there was a prenatal effect on birth weight (confirming numerous other correlational studies) but that (disconfirming other studies) although there was a statistical association with the offspring's externalizing behavior, this was not due to a prenatal influence.

A Finnish study used the same design for the same purpose, but with a focus on hyperkinetic disorders (Obel et al., 2007). The sample size of pairs discordant for maternal smoking in the pregnancy was over 18,000 at the time of the children's birth. The clinic record register data were used for the diagnosis of attention deficit/hyperactivity disorder (ADHD), and the maternal report at the time of the child's birth was used for the assessment of smoking in pregnancy. In the sample as a whole, there was a very highly significant odds ratio of 2.2 for the effect of smoking in pregnancy on the risk for ADHD in males, after control for confounders. Moreover, there was a dose-response gradient according to the number of cigarettes per day smoked, as well as a lower risk in those who quit smoking before the pregnancy than in those who continued smoking during the pregnancy. The causal inference might seem to be justified. Nevertheless, the discordant sibling comparison showed a much reduced, statistically nonsignificant odds ratio of 1.2 in males, casting serious doubt on such an inference. The findings in females, based on a sample size far smaller than that for males, were more ambiguous. Note that, unlike the U.S. study, the Finnish study had a measure of ADHD that did not depend on maternal report. Also, both the Finnish and the U.S. study showed that the discordant siblings analysis confirmed the effect of maternal smoking on birth weight and disconfirmed a strong prenatal effect on risk for ADHD. Both studies showed that, in the absence of a natural experiment, the usual control for confounders was inadequate to protect against a misleading inference on causation.

Migration Strategies

The use of migration strategies to control for genetic mediation is based on the observation that particular adverse outcomes differ among ethnic groups. At one time, it was assumed that these differences reflected differences in genetic background of the ethnic groups. That was one possibility, but the ethnic variations were also associated with a wide range of lifestyle differences. The natural experiment, therefore, arose from determining what happened when high-risk (or low-risk) ethnic groups moved to a different country and adopted entirely different lifestyles. Two medical examples illustrate the point very well. The Pima Indians have long been observed to have unusually high rates of obesity and associated obesity-connected conditions. It seems likely that a genetic vulnerability plays a

role in this. Nevertheless, the experimental contrast was provided by comparison of the medical outcomes between Pima Indians living in affluent circumstances and those living in areas with quite different (less adequate) nutritional opportunities. The findings were striking in showing that when the lifestyle and dietary conditions varied so did the rate of medical conditions in Pima Indians (Ravussin, Valencia, Esparza, Bennett, & Schulz, 1994; Valencia et al., 1999).

A somewhat parallel, although less dramatic, example is provided by an early study of coronary artery disease in people of Japanese origin living in Japan, Hawaii, and California (Marmot & Syme, 1976). The findings showed that when the Japanese people adopted a Californian lifestyle, their rate of coronary artery disease rose to much the same levels of those of Caucasian individuals living in California. In other words, although there may well have been genetic factors involved in the vulnerability, the prime risk factor lay in lifestyle features, including diet. It is interesting to note that, in this instance, smoking was not implicated in the migration difference because smoking levels were very high in Japan, as well as in California. It should be noted, too, that the findings with respect to hypertension were rather different.

The best-documented psychopathological example concerns the raised rate of schizophrenia spectrum disorders in people of Afro-Caribbean origin living in either the UK or the Netherlands—the two countries in which the phenomenon has been most extensively examined (Jones & Fung, 2005). The migration design relies on two key comparisons: the rate of the disorder in migrants compared with the rate of the disorder in their country of origin, and the rate in migrants compared with the rate in nonimmigrants in the host country to which they migrated. In the case of schizophrenia spectrum disorders, it was found that the rate was substantially higher in people of Afro-Caribbean descent living in the U.K. or the Netherlands than in Caucasians living in the same two countries and in people of presumably the same ethnic origins living in one of the islands in the West Indies. A very detailed, carefully controlled study showed that the difference in the incidence of schizophrenia spectrum disorders was not an artifact of recognition or diagnosis but, rather, that it appeared to reflect some aspect of adversities associated with migration to the U.K. or the Netherlands (Jones & Fung, 2005). The precise mechanisms have yet to be identified, but the evidence strongly points to some form of environmental influence. The possibility of differential migration of individuals with a genetic liability to schizophrenia was shown to be unlikely because the ethnic difference applied much more strongly to the siblings of Afro-Caribbean people with schizophrenia than it did to the parents. Nevertheless, it could be that gene–environment interactions were involved.

Recently, a parallel study in the U.S. showed a somewhat similar ethnic difference, with African-Americans about three times more likely to be diagnosed as having schizophrenia than are Whites (Bresnahan et al., 2007). The causal inference with respect to some kind of environmental influence is rather strong, but it would be strengthened if there was evidence of environmental factors differentially operating within ethnic groups. In summary, as applied to schizophrenia, the migration findings make a compelling case for some kind of environmental influence, but they do not identify the key components of such an influence.

Adoption

Studies of risk experiences within adoptee samples also provide a good way of examining possible environmental risk mediation because it excludes the possibility of genetic mediation. Thus, Case, Lubotsky, and Paxson (2002) found that the association between low SES and children's health outcomes was closely similar in biological and adoptive families (unfortunately the findings are reported rather briefly without adequate detail). The inference (as in the Duyme, Dumaret, & Tomkiewicz, 1999, study dealing with IQ on the dependent variable—see below) is that some form of causal environmental effect was operative. The adoptee design, however, requires specific attention to the possibility of selective placement. Also, it is limited by the restricted range of risk environments in most adoptee samples, the difficulty of getting access to large representative samples of early adopted infants, and the paucity of good information on the qualities of the adoptive parents. As a result of these and other considerations, the design has been of less use in practice than anticipated by theory.

OTHER USES OF TWIN AND ADOPTEE STRATEGIES

Mendelian Randomization (MR)

MR is one of the most recently developed techniques, with a rationale first outlined by Katan (1986) and then by Gray and Wheatley (1991) but more fully developed by Davey-Smith and Ebrahim (2003, 2005). At first sight, the title sounds misleading because genes are of course not randomly distributed in the sense of everyone having the same chance of receiving particular genetic variants. Nevertheless, the allelic variations in the parental genes are randomly distributed among offspring at the stage of meiosis, as noted in Mendel's second law. What this means is that the inheritance of any one trait is independent of the inheritance of other traits. The ingenious point with respect to testing a hypothesis regarding environmental causation, however, is that the allelic variations are likely to be randomly distributed with respect to the environmental risk hypothesis being tested. In cases where the pathway from genotype to phenotype is relatively well understood and when the gene in question has functionally varying alleles, allelic status can be used as a randomly distributed proxy for the behavioral differences that emerge from it (Irons, McGue, Iacono, & Oetting, 2007). In other words, the groups defined by the allelic variation known to affect exposure to some intermediate phenotype (such as drinking behavior) can be used to study the effect of variation

in that phenotype on some disease or disorder hypothesized to be caused by the intermediate phenotype. Note that the design is employed primarily to rule out reverse causation; it does that because the genotype is unaffected by the disease, disorder, or trait being studied. The design is also not invalidated by confounding, provided that there is no pleiotropic effect of the regulatory allelic variation (Keavney, 2004).

The design was originally developed for use in internal medicine to determine if the disease being studied had caused the environmental risk factor, rather than the other way round. It has had some success both in providing support for a causal hypothesis (Casas, Bautista, Smeeth, Sharma, & Hingorani, 2005) and in casting doubt on its validity (Keavney et al., 2006). Nevertheless, it usually requires huge samples, and it is weak if the risk effect is small, if there are difficulties in measuring the intermediate phenotype, or if the risk operates only above some threshold (Meade, Humphries, & De Stavola, 2006).

One key example of using the MR (Nitsch et al., 2006; Tobin, Minelli, & Burton, 2004) design in the psychopathological arena (Irons et al., 2007) exhibits fewer of these problems than do other examples: testing the supposed gateway effect of heavy early alcohol use in predisposing one to both the later use of "hard" drugs (such as cocaine or heroin) and to a broad range of antisocial behavior. That there is a substantial overall statistical association in the general population is not in doubt, but the query is whether it derives from a general liability to disinhibited behavior (influenced by either genetic or environmental factors) or whether the intermediate phenotype of heavy drinking causes the other problems. The opportunity to employ the MR design is provided by an allelic variation of a gene that leads to greatly reduced aldehyde dehydrogenase (ALDH) enzyme activity. Individuals with ALDH2 enzyme deficiency have an unpleasant flushing response to any ingestion of alcohol. The relevant gene is found almost exclusively in East Asians and, in this ethnic group, it is associated with a substantially decreased rate of alcoholism as compared with those having normal enzyme activity. Irons et al. (2007) were creative in realizing that the design could be used in their study of adoptees born in Korea and placed in adoptive homes in the U.S. in infancy (none of the adopting parents being of East Asian descent). The findings showed (as expected) a substantial genetic effect on alcohol abuse but no significant difference between the ALDH2 deficient and nondeficient groups in either drug abuse or antisocial behavior.

The implication is that the association was derived from a shared general liability for problem behavior and not from the causal effect of early alcohol drinking. As with all natural experiments, it is desirable whenever possible to include more than one design element to separate variables. Adoption provided that further opportunity here. The results showed no effect of adoptive parental alcoholism (confirming the twin studies' conclusions of a genetic influence that will not apply with nonbiological parentage) but did show a significant effect of drinking by siblings, suggesting that this may well represent an environmental effect. In that connection, Irons et al. (2007) pointed out that although the ALDH2 allelic variation has a strong effect on alcoholism, it is not determinative. Not only do some individuals with ALDH2 deficiency become alcoholic, but there is a gender difference in effect and a secular trend. As with all natural experiments, there are limitations, and it is crucial that users of the design take steps to deal with these (as was the case in this example). In conclusion, MD provides a good means of eliminating reverse causation and controlling for social selection or allocation bias, but it works best when the relevant gene has a strong, highly focused effect, when pleiotropic effects are slight, and when there is a well defined intermediate phenotype (Katan, 2004; Keavney, 2004). Caution is needed, even when the genetic effect is strong (as it is with ALDH2), if environmental influences also affect the genetic pathway to the phenotype. The design is certainly a useful addition to the natural experiments armentarium, but, equally, it is likely to be practicable in only a few circumstances.

Use of an Instrumental Variable External to the Liability to the Psychopathological Outcome

Instrumental variables refer to circumstances that do not affect the outcome being studied but do influence the putative risk factor being considered (Foster & McLanahan, 1996). Researchers often embed instrumental variables within a Rubin causal model as part of a statistical approach designed to deal with omitted confounding variables, two-way influences, and imprecision in measurement (see Angrist, Imbens, & Rubin, 1996a, 1996b; Robins & Greenland, 1996; Rosenbaum, 1996). This broader statistical usage is outside the scope of this article. Our focus is only on how instrumental variables are used in a natural experiment.

The question of whether unusually early use of alcohol constitutes a risk factor for later alcoholism or alcohol dependence in adult life can also be investigated through the use of an instrumental variable that is external to the psychopathological liability and that provides an increased propensity to early use of alcohol. An unusually early onset of puberty in girls constitutes just such an independent factor. The rationale is similar to that for MR. That is, if early use of alcohol truly predisposes one to alcoholism, the same early use effect should be found even if the early use was a result of early puberty rather than an overall alcoholism liability. The findings across three large-scale general population epidemiological studies in Sweden, New Zealand, and Finland are consistent (Caspi & Moffitt, 1991; Pulkkinen, Kaprio, & Rose, 2006; Stattin & Magnusson, 1990). All show that early puberty is indeed associated with an increased rate of drinking alcohol and with drunkenness during the teenage years. The test, however, lies in what is found in early adult life. All three studies are consistent in showing that although there was an effect on alcohol use in adolescence, this

was no longer apparent by the mid-twenties. In other words, the effects were quite different from those found in ordinary circumstances. The implication is that the early use was not a causal factor despite its strong correlation. Rather, it reflected the same underlying liability. Longitudinal studies tell the same story. Thus, McGue and Iacono (2005) showed that early alcohol use was associated not only with later alcoholism but also with smoking, the use of drugs, involvement with the police, early sexual activities, conduct problems, and educational underachievement. The inference is that, in the case of the timing of first alcohol use, it is likely that there is not a causal effect; rather, the early use reflects a shared liability to a broader range of problem behavior.

The use of an instrumental variable as a means of testing causal inferences involves two key design features. First, the instrumental variable must be outside the control of the individual (to avoid allocation bias), and second, it must affect the outcome by some means that is independent of the usual liability to that outcome (to avoid both possible genetic mediation and possible social confounding). There are few examples of its use in this way as a form of natural experiment, possibly because there are few clear cut circumstances of variables operating in this way. Nevertheless, the strategy is worth considering when the possibility arises.

Adoption/Fostering as a Means of Separating Prenatal Drug/Alcohol Effects From Effects of the Postnatal Environment

With some supposed prenatal effects, there is the problem of differentiating prenatal and postnatal effects. Thus, most, but not all, longitudinal studies have shown that fetal cocaine exposure is associated with adverse effects on neurodevelopment and cognitive performance (see e.g., Jacobson, Jacobson, Sokol, Martier, & Chiodo, 1996; Singer, Arendt, Minnes, Farkas, & Salvator, 2000). Not only could this association be mediated by genetic influences on the mother's liability to take cocaine, but it could also be derived from environmentally mediated postnatal effects stemming from the adversities in rearing that are generally more common in drug-dependent mothers.

A possible way forward is provided by comparison of drug-exposed infants who are reared by their biological mothers with those reared by foster or adoptive parents. Thus, Singer et al. (2004), as part of a longitudinal study from birth, compared the outcome at 4 years of 48 prenatally cocaine-exposed infants reared by their biological families and 42 reared by adoptive or foster parents. The fostered/adopted children had a slightly higher mean IQ (83 vs. 79), but there was also a much higher proportion of them with an IQ below 70 (10% vs. 37%). The implication is that the cognitive deficit was due to the prenatal cocaine exposure. An alternative explanation is that the foster/adopted children had less prenatal cocaine exposure. This possibility can be ruled out, however, because prenatal cocaine exposure was actually twice as high (which probably led to the children being removed from the biological mother). Genetic risk cannot be excluded but seems unlikely as a sufficient explanation. What is more problematic is that the prenatal cocaine exposure was also accompanied by a greater use of other drugs (including alcohol and tobacco), so that the risk could involve other substances rather than cocaine by itself.

Moe (2002) used a design that was similar in its focus on prenatal drug/alcohol exposure but used a matched control group who had not experienced prenatal exposure. She found that the cognitive functioning at 4.5 years in substance-exposed children who were adopted or fostered in infancy was substantially and significantly lower than in the control group. The findings supported the inference that the adverse effects on cognition and development were primarily a function of prenatal substance exposure rather than exposure to an adverse rearing environment.

The adoption/fostering strategy provides a useful means of separating prenatal and postnatal effects but, on its own, it is much less informative regarding either the precise risk feature (because the prenatal exposure will often involve several substances) or the particular risk mechanisms.

Adoption as a Radical Change in Environment

There are two examples of the radical, sudden change of environment caused by adoption providing a natural experiment. First, there is the study of children from extremely depriving institutions in Romania being adopted into generally well-functioning adoptive families in the U.K. (Beckett et al., 2006; Rutter et al., 2007; Rutter & The English and Romanian Adoptees Study Team, 1998). The causal inference with respect to the effects of prior institutional deprivation could be tested in two rather different ways. First, there was the test of developmental recovery postadoption. Because this recovery was very substantial and was not a function of selective placement, the inference that the profound developmental impairment was caused by the institutional deprivation was strong. Second, there was the test of whether the remaining deficits were a function of preadoption institutional experiences and not variations in the quality of postadoption rearing. Follow-up findings at age 4, 6, and 11 years were consistent in showing that deficits were a function of the preadoption environment and were seemingly not influenced by variations in the postadoption environment. Two further findings much strengthened the causal inference. First, the effects of institutional deprivation were about as strong at age 11 as they had been at ages 4 and 6 years. Second, the effects were largely restricted to outcomes (such as quasi-autistic patterns and disinhibited attachment) that were rare in groups not experiencing institutional deprivation (Kreppner et al., 2007; Rutter et al., 1999).

Two main threats to validity had to be dealt with. First, a major difficulty in most previous studies of institutional care has been

the likelihood that, for some children, admission to the institutions had been a consequence of some type of preexisting handicap. That was implausible in this case, because almost all the children were admitted in the first few weeks of life. Second, there was often the likelihood that children were chosen for adoption on the basis of their positive functioning. That was less likely here, because no children had been adopted before the fall of the Ceauşescu regime in 1989 (so that the children's age at leaving the institution depended on how old they were in 1989 rather than any aspect of their developmental progress). Parental choice will have played some part in who was selected for adoption, but detailed analyses gave no evidence that this accounted for findings (Beckett et al., 2006). In summary, despite the unavoidable absence of measures of the children's functioning while in the Romanian institutions, several design features meant that the causal inference was strong.

Duyme et al.'s (1999) study of late-adopted children in France provides a rather different example. The sample comprised children who had been neglected or abused during infancy and who had been compulsorily removed from their biological families as a result. In addition, they had multiple foster family or institutional placements and an IQ test preadoption showing an IQ between 60 and 86. They were age 4 to 6 at the time of adoptive placement and age 11 to 18 at follow-up. The focus of the study was entirely on IQ. Two main results were evident. First, the children showed a substantially higher IQ postadoption than they did preadoption (91 vs. 78), and second, the degree of change postadoption was systematically associated with the socioeconomic level of the adoptive family (being 19 points in the highest group and 8 points in the lowest). It is the latter finding that provides the main basis of the causal inference with respect to the influence of the adoptive family environment. Note, however, that even in the higher SES groups, the mean IQ (98) was just below the general population average of 100. The conclusion that the qualities of the adoptive home environment mattered should not be seen as contradicting the Romanian study findings because the preadoption circumstances were so different (the profound deprivation in Romanian institutions being far greater).

Subgroups Within Adoptees

Another research strategy is provided by the comparison of subgroups within adoptees in order to delineate the specifics of the environmental risk. The Romanian adoptee study provides the example (Sonuga-Barke et al., 2007). One key question concerns the extent to which the psychological sequelae of the institutional deprivation derived from subnutrition or psychological deprivation. All the subnourished children suffered from both, but a subgroup of the psychologically deprived were not subnourished (at least as indexed by body weight at the time of U.K. entry). Three main findings were evident. First, after at least 6 months of institutional deprivation, the group without subnutrition showed a marked impairment in head growth. The implication is that the deprivation impaired brain growth even when the overall level of nutrition was adequate. Second, there were marked adverse psychological sequelae of deprivation even when it was not associated with subnutrition, pointing to the likely causative role of psychological deprivation as such. Third, subnutrition had far less effect on outcome than did length of institutional deprivation. The implication was that the main risk effect derived from the psychological deprivation rather than the subnutrition.

The main threat to the validity of this conclusion stems from the lack of evidence on why some children were subnourished and others were not. The obvious supposition would seem to be that the more nourished children were better treated and hence should have better outcomes. The empirical evidence that, to the contrary, they fared badly makes that confound implausible. Nevertheless, the uncertainty over the origin of the differences in experiences necessarily makes for caution.

Twin-Singleton Comparisons

Twin-singleton comparisons provide a useful design to compare the effects of obstetric/perinatal complications with patterns of parent–child interactions on language development. The starting point is the well-demonstrated finding that, as a group, twins lag behind singletons in their language development by about 3 months at 3 years of age (Rutter, Thorpe, Greenwood, Northstone, & Golding, 2003; Thorpe, Rutter, & Greenwood, 2003). The natural experiment arises from two key features. First, although genetic influences on language will of course operate within both groups of twins and groups of singletons, there is no reason to suppose that they will differ in either strength or type between the two groups. Second, overall social disadvantage is unlikely to be responsible for the language impairment because it is not particularly associated with twinning. In other words, the comparison virtually rules out the two major confounding factors operative in the general population. The leading contenders were obstetric complications (known to be much more common in twins) and patterns of parent–child interaction and communication that had been altered by having to deal with two children of the same age at the same time. The design provided a test of causal effects by focusing on language performance at age 3 after taking into account the level of functioning at 20 months of age, using within-individual change as the criterion. The findings showed no effect of obstetric complications (in a group that excluded those born after less than 34 weeks gestation). By contrast, there was a significant effect of parent–child interaction/communication, which eliminated the twin-singleton difference when put into an overall model. It was this demonstrated mediation effect (see Baron & Kenny, 1986) in combination with the within-individual change between 20 and 36 months and the elimination of the main competing hypothesis that makes the causal inference plausible.

DESIGNS FOR AVOIDING OR OTHERWISE DEALING WITH SELECTION BIAS

Universal Experiences to Avoid Selection Bias

The one straightforward way to test environmental mediation effects by avoiding social selection or allocation biases is to study the effects of experiences that apply to total populations in which the experience being studied is not open to the influence of personal choice. Three rather different examples illustrate this strategy.

First, there are two studies of prenatal famine imposed on a total population. The first concerns the Dutch famine in World War II, in which researchers examined the risk of a range of psychopathological outcomes in offspring (Hoek, Brown, & Susser, 1998; Neugebauer, Hoek, & Susser, 1999). It constituted a natural experiment because the famine was both severe and time limited and because it was externally imposed on the total population and not just a socially disadvantaged subgroup, thus avoiding the usual social selection confound. The initial finding was that exposure to the famine in early gestation was associated with a higher frequency of central nervous system congenital anomalies (Stein, Susser, Saenger, & Marolla, 1975). This suggested the operation of some kind of biological pathway. Much later, researchers used the same sample to examine possible risks for schizophrenia (Susser et al., 1996) and found a two-fold increase in risk associated with exposure to famine conditions in early gestation. Three criteria were used to define the exposed birth cohorts: low food rations in the first trimester of gestation, conception at the height of the famine as indicated by adverse health effects in the general population, and a detectable excess of congenital neural defects. Schizophrenia was ascertained in individuals through a national psychiatric registry. A variety of steps was taken to examine the possibility of both selection bias and selective survival, and neither seemed likely to account for the effect. The main limitation of the study concerns the reliance on groups rather than exposure at the individual level. The causal inference was strengthened, however, by the gestational period specificity, by the diagnostic specificity, and by the parallel association with congenital central nervous system deficits.

A parallel example was provided by the Chinese famine in 1959–1961 (St. Clair et al., 2005). The findings showed a very similar two-fold increase in the risk of schizophrenia. The study had the advantage of larger numbers than those seen in the Dutch study but the weakness of the uncertainties on the exact timing of the prenatal exposure by month of gestation. It is notable, however, how similar the findings were despite the samples being so ethnically and culturally different. The reality of the biological famine effect was shown by the decrease in birth rate during the famine and by the increase in mortality rate. Because this is such an unusual experience, it cannot be concluded that prenatal famine ordinarily constitutes a causal influence on schizophrenia. On the other hand, it may be that the effect of folate deficiency in causing de novo genomic mutations might constitute a possible mechanism (McClellan, Susser, & King, 2006).

The same basic strategy of a universal intervention may be applied to the removal of risks. Thus, Honda, Shimizu, and Rutter (2005) used the natural experiment provided by the fact that Japan stopped using the measles, mumps, and rubella (MMR) vaccine at a time when its use was continuing in other parts of the world to test the hypothesis that MMR was responsible for the marked rise in the rate of diagnosed autism spectrum disorders (ASD). The strength of this natural experiment lay in the availability of systematic standardized diagnostic data for a defined geographical area, in the exact timing of the total withdrawal of use of the MMR vaccine, in follow-up to age 7, and in the fact that ASD rates elsewhere in the world were still rising. The findings showed that the withdrawal of MMR had no effect on the rising trajectory in the rate of ASD. In addition to a variety of other checks, the findings made a universal causal effect of MMR on ASD implausible, especially in the light of the fact that other research strategies all gave rise to the same negative conclusion (Rutter, 2005). It should be appreciated, however, that although the finding virtually rules out the possibility of MMR as a cause of an overall rise in the rate of ASD, it cannot rule out the possibility of an idiosyncratic effect in a small minority of individuals. The same applies to the comparable strategy, with similarly negative findings, on the withdrawal of thimerosal (a mercury preservative) from vaccines in Scandinavia in the early 1990s. The removal of risk was unassociated with any change in the rate of ASD (Atladóttir et al., 2007; Madsen et al., 2003).

The natural experiment of the opening of a casino on an American Indian reservation was used by Costello, Compton, Keeler, and Angold (2003) to determine if the alleviation of poverty brought mental health benefits. The law required that a particular proportion of the profits from the casino had to be distributed to all those living on the reservation without any actions by individuals (i.e., as with MMR, eliminating the possibility of allocation bias). By good fortune, the casino was set up in the middle of a prospective longitudinal study of child mental health. The results showed that the casino profits had indeed resulted in a substantial reduction in poverty among the Indians living on the reservation and that this was followed by a reduction in the rate of some (but not all) kinds of child psychopathology. More detailed analyses indicated that the benefits were likely to have been mediated by changes in the family. Yet another example is provided by the effects of forced school closure on the scholastic attainments of the children who experienced a loss of educational input for the period of the closure (de Groot, 1951; Jencks et al., 1972). Once more, it is unlikely that the adverse effects were either chance fluctuations or due to some unconnected happening. The design in each study provided a good test of the causal hypothesis on general time trends in the outcomes being examined, but it could not effectively examine effects at the individual level.

The last example, namely the effect of major catastrophes (such as shipwreck or earthquake), similarly suffers from the inevitable lack of preexposure time trends in psychological functioning. On the other hand, by the nature of the experience, it cannot have been brought about by the individuals themselves. Nevertheless, that does not necessarily rule out social selection effects (which are brought on by influences on individuals in a situation that exposed them to the catastrophe). Three features are valuable in testing for environmental mediation. First, the comparison group needs to be chosen to deal with possible social selection effects. Thus, in Yule, Udwin, and Murdoch's (1990) study of the sinking of a cruise ship, children who had applied to go on the cruise but who had not been able to get a place were used as controls. Second, attempts must be made to test for dose–response relationships within the exposed group. For example, in their study of the California school sniper tragedy, Pynoos et al. (1987) showed such a relationship with respect to the degree of impact of the incident. Third, the specificity of the psychopathological effects must be taken into account. The particular symptoms associated with posttraumatic stress disorder provide that (Yule, 2002), although it should be noted that these are not the only sequelae.

Within-Individual Change

Numerous cross-sectional studies have shown that friends tend to be similar in their behavior and interests (homophily). The question is whether this likeness reflects selection (i.e., individuals choose friends like themselves; a selection process) or socialization (i.e., individuals are influenced by the behavior and attitudes of their friends). Kandel (1978) recognized that, in order to test these alternatives, it was necessary to have longitudinal data to examine within-individual changes over time in relation to the formation and dissolution of friendships. Her sample comprised New York high-school students and the dependent variables included marijuana use. Measures were available on both friends-to-be and former friends, and stable and unstable friendships could be contrasted. In brief, it was found that homophily among stable pairs increased over time, that it was greater among newly formed pairs after the pairs had been formed than it was before the friendship, and that pairs that remained stable over time were more similar than the subsequently unstable pairs. Quantitative analyses showed that both selection and socialization were operative.

A rather similar issue arises with respect to the possible effect of juvenile gangs in facilitating delinquent behavior. Numerous studies had shown that gang members are more likely to commit serious and violent offences at high frequency (Spergel, 1990). The query arising from these between-group comparisons is whether this is a consequence of individuals with a greater antisocial liability being more likely to join a gang or, rather, a deviant socialization effect of gang membership on delinquent activities. Thornberry, Krohn, Lizotte, and Chard-Wiershem (1993) tested the latter possibility by examining within-individual change over time using data from the multiwave Rochester Youth Development Study. Between-group comparisons were undertaken with respect to non-gang members, transient gang members, and stable gang members, and comparisons over time were made with respect to within-individual change, according to self-reported delinquent acts before, during, and after gang membership. For transient gang members, there was no evidence of a selection effect (i.e., they did not differ from non-gang members before joining a gang), but there was substantial evidence for social facilitation (i.e., delinquent activities were higher during the period of gang membership). For stable gang members, by contrast, the findings suggested both selection and social facilitation effects.

A comparable issue arises from the numerous between-group comparisons showing that married men are less likely than unmarried men to engage in crime (see Sampson & Laub, 1993). The question is whether this difference reflects selection into marriage or a causal effect of marriage. Using their long-term follow-up (to age 70) of the Gluecks' serious adolescent delinquency sample, Sampson et al. (2006) addressed this issue by determining whether individuals were less likely, over time, to engage in crime during their periods of being married than during their periods of nonmarriage. They used 10 individual-specific and 10 family and parental background features and a range of adult time-varying covariates to assess selection into marriage. They used an inverse probability of treatment weighting, devised by Robins et al. (2000), to weight each person and period by the inverse of the predicted probability of receiving the "treatment" (i.e., marriage) that they actually received in that period. In short, the model in effect creates a pseudopopulation of weighted replicates allowing a comparison of status (married vs. unmarried) without making distributional assumptions about counterfactuals. The model allows age effects to be taken into account, and the approach also allows a quantitative estimate of marriage effects on crime. Researchers found an average crime reduction of about 35% in periods of marriage (using a conservative approach rating ruling out reciprocal effects and imposing a strict causal order). The robustness of the estimate was shown by the demonstration that it was broadly similar across both the shorter time span of 17 to 32 years and the longer time span of 17 to 70 years. As with any other nonexperimental study, it is impossible entirely to rule out the possibility of the operation of some unconceptualized and unmeasured confounder. Nevertheless, as the authors argue, given the unusually rich range of measures available, it is difficult to imagine what time-stable or time-invariant covariate could overcome the magnitude and robustness of the marriage effect.

Rutter, Maughan, Mortimore, Ouston, and Smith's (1979) study of school effects constitutes another, somewhat different, example of using changes over time to assess a possible environmentally mediated causal effect. The hypothesis being tested was that the qualities of schools influenced the attainments and

behavior of the children attending those schools. The competing alternative hypothesis stated that any associations found reflected either variations in the intake of pupils to each school or the influence of the pupils on school functioning. The natural experiment was provided by the fact that the transfer from elementary to high schools in inner London at that time involved a marked lack of continuity between the two (so that the children at any one elementary school moved to a large number of different secondary schools). Data were available for pupil functioning pretransfer (at age 11–12) and at the end of compulsory schooling 5 years later (also as part of a later follow-up). It was argued that if the within-schools variations observed for pupil outcomes were truly a function of school influences, then the differences should not be accounted for by intake differences—they should instead be a function of measured features in school quality, and the effects should increase over the course of schooling. The findings supported this argument. The causal inference was further strengthened by comparable findings in other longitudinal studies and by the beneficial effects on pupil outcomes of the appointment of new school principals in two out of three failing schools on the point of closure (see Rutter & Maughan, 2002). The causal inference in this example is supported by the overall pattern of findings but is much weakened by the need to rely on group tendencies rather than within-person change at the individual level for specifics of school experience.

DEALING WITH UNMEASURED CONFOUNDERS: REGRESSION DISCONTINUITY (RD) DESIGNS

The only natural experiment that can adequately deal with unmeasured confounders is the RD design. RD designs were first introduced by Thistlethwaite and Campbell (1960) nearly half a century ago as an alternative to RCTs. The key defining feature is that allocation for some planned intervention is by means of an assignment variable that uses a strict predetermined cutoff rather than randomization. In other words, the design capitalizes on a major selection bias, provided that it is under strict control. The basic point is that the assignment cannot be caused by the intervention—it does not matter whether or not it is related to the outcome. It does, however, require that all participants belong to the same population prior to assignment. The statistical analysis also requires accurate specification of the intervention effect (e.g., whether it is linear or curvilinear) and inclusion of an interaction term when this is relevant. Note that, unlike an RCT, effects are measured in terms of a discontinuity in regression lines (hence its name) instead of a difference in means. Although at first sight it is not obvious that RD allows an unbiased estimate of a causal effect, it has been shown mathematically that it does (Rubin, 1977; Shadish et al., 2002), and this constitutes its major advantage (Laird & Mosteller, 1990).

Because RD was an alternative to RCT for planned experimental interventions, it has rather limited use as a natural experiment to test prior causes of a nonexperimental variety. Nevertheless, there are two examples showing that it can occasionally be used in this way. Cahan and Cohen (1989) showed its applicability in their study determining whether the amount of formal education, as opposed to increasing chronological age, has an effect on cognitive performance. The design was possible because all children in Jerusalem's Hebrew Language state-controlled elementary schools (other than those providing only special education) in 1987 had a single admission date in December. Thus, within each school year, there was a 12-month age span from the oldest to the youngest child (all of whom would have received the same amount of schooling), and between school classes there was a 12-month difference in duration of schooling (but a similar within-group variation in age). There was a problem with the key assumption that there could be no exceptions to the age cut-off. Although grade retention and grade skipping were rare, admission was sometimes delayed and sometimes accelerated; moreover, this variation was nonrandom (with variation greatest near the cut-off point and with delayed-admission students tending to be less able, whereas accelerated-admission students were more able). To keep this bias to a very low level, the study excluded all overage and underage children in each of the three age-specified year classes studied. It also excluded all children born in November and December. To allow direct comparisons between tests, researchers standardized effect sizes by using the pooled within-age standardization in Grade 4 (the youngest age group). The findings were striking in showing that the school effect exceeded the age effect for 10 out of the 12 cognitive measures used. The two exceptions were both figural tests; the greatest school effects were found for verbal tests and the one numerical test used. There was no reason to expect anything other than a linear effect (it might have been different in a less standardized school system). It should be noted that, although the data are not individually longitudinal, they do represent changes over time. Thus, although technically a cross-sectional comparison, the RD design (by ensuring strict comparability of groups) allows the strong inference of within-individual change.

The study is described in some detail in order to emphasize the crucial importance of ensuring that the RD design assumptions are truly met and to illustrate the practical steps that may be taken to bring this about. The one key assumption that was not explicitly tested was the absence of an Age × Schooling interaction, but the figures provided for the three grades studied gave no indication that there was an interaction. It should be noted that the findings are in keeping with the somewhat different form of natural experiment provided by forced school closure (see Rutter & Madge, 1976). None of these schooling experiments tested the possibility of an interaction between schooling effects and the quality of family influences. The findings from studies of gains and losses in the long U.S. summer school vacation indications that this is highly likely (Entwisle, Alexander, & Olson, 2004; Heyns, 1978).

A second example of the use of the RD design is provided by a retrospective assessment of the effects of Head Start on children's health and school progress using a discontinuity in program funding across counties as the research lever (Ludwig & Miller, in press). Specifically, during the spring of 1965, the U.S. Office of Economic Opportunity provided technical assistance to the 300 poorest counties to develop Head Start funding proposals. Funding rates were found to be 50% to 100% higher in countries with poverty levels just above the Office of Economic Opportunity's cut-off. The treatment group comprised those just above the cut-off, and the control group comprised those just below. Because the cut-off was on a predetermined variable, the design was operative and allowed little opportunity for strategic behavior by individuals that could invalidate the required assumptions. The findings derive from somewhat rough county level funding data and from the National Educational Longitudinal Study, which tracked a national sample of eighth graders in 1988. The funding difference persisted through the late 1970s. In brief, the findings showed a substantial effect of Head Start on children's health, a lesser but still significant effect on school graduation, and no significant effect on reading or math scores or on noncognitive outcomes. The validity of the findings was tested by determining whether similar effects were found in age groups that could not have benefited from Head Start (they were not) and whether the health benefits applied to outcomes that could not plausibly be influenced by Head Start (they did not). The limitations stem from reliance on countywide measures, from reliance on residence at follow-up rather than at program initiation, and from lack of satisfactory data on in- and out-migration. The study well illustrates the value of incorporating multiple tests of validity but also demonstrates the uncertainties that usually operate when seeking to apply the RD design to naturally occurring circumstances outside of experimental control. The RD design provides an important alternative to RCT, but it is quite limited in its applicability to naturally occurring prior risk or protective factors.

CONTEXTUAL EFFECTS

The last type of natural experiment to consider is quite different in that it focuses not on testing an environmental mediation hypothesis as such, but rather on the possibility that effects apply only in certain contexts. Dredging data in search of subgroup differences is likely to lead to artifactual false positive findings. By contrast, a hypothesis-driven examination of contextual effects may be quite informative.

The context may be either genetic or psychosocial. The former is exemplified by the existence of interactions between specific identified genetic variants and specific measured environments (G × E; Caspi & Moffitt, 2006; Moffitt, Caspi, & Rutter, 2006; Rutter, 2006, 2007; Rutter, Moffitt, & Caspi, 2006). Thus, Caspi et al. (2002) found that physical maltreatment predisposed individuals to antisocial behavior only in the presence of an allelic variant of a gene on the X chromosome that influenced monoamine-oxidase-A activity. The finding has been replicated by other groups and confirmed in a meta-analysis (Kim-Cohen et al., 2006). Similarly, Caspi et al. (2003) showed that a variant of the serotonin transporter promoter moderated the effects of both child abuse and life stresses in relation to the outcome of depression. This finding has also been broadly replicated 14 times with only three failures to replicate. A third example is provided by Caspi et al.'s (2005) finding that the Val/Val variant of the COMT gene is associated with the risk factor linking early heavy use of cannabis with the later onset of a schizophrenia spectrum disorder. Cannabis use in the absence of this variant is unassociated with a schizophrenia spectrum outcome. It is relevant that the first two of these G × E epidemiological findings have been confirmed by both animal studies and human brain imaging studies examining neural effects.

A psychosocial contextual effect may be illustrated by Geoffroy et al.'s (2007) finding that nonmaternal care in the first year of life was associated with higher language scores at age 4 in children from lower SES families (as compared with language scores of children in maternal care) but not among those from moderate or higher SES backgrounds. The finding was evident in a longitudinal study linking early nonmaternal care with later language performance, thus establishing the temporal relationship. The most important threat to validity was provided by the possibility of systematic bias in the selection of nonmaternal care, but a rigorous examination of possible social selection effects provided no support for this alternative explanation. Nevertheless, some caution is needed with respect to the causal inference because of the lack of direct measurement of the quality of rearing by either the biological mothers or the nonmaternal care. Nevertheless, the implausibility of alternative explanations for the interaction effect found suggests that there was likely to be a causal effect and, hence, that there was need for further research to test this possibility. The implication is that when family circumstances are good, nonmaternal care brings no particular benefits, but when family circumstances are poor, it may contribute to an improved outcome.

The crucial point in both the genetic and psychosocial examples is that it may be informative to test for specific hypothesized contextual effects rather than assume a universally applicable environmental effect.

OVERALL CONCLUSIONS

In this article, I have described fifteen rather different types of natural experiments that may be used as a means of testing whether a naturally occurring risk experience had a statistical association with some relevant outcome and whether a causal inference might be justified. Each design has its own particular strengths and limitations, and none is free of the latter. Nevertheless, taken in conjunction (when multiple designs are possible), they can do much to strengthen or weaken the causal

inference. Thus, for example, a multivariate twin design, a discordant twin pair design, MR, and the use of early puberty as an instrumental variable all suggested that the correlation found between early alcohol consumption and later alcohol dependence and antisocial behavior reflected a shared genetic liability and not causation. Conversely, both a regression discontinuity design and school closure studies suggested that there was a causal effect of overall duration of education. Similarly, there are several examples of natural experiments that can be useful in differentiating between different types of environmental risk effect.

As with most other research designs, natural experiments tend to be more effective for identifying a probable causal effect than they are for determining precisely which aspect of the experience carries the main risk and what form of mediating mechanism is involved. In future research, natural experiments will almost always need to be combined with true human or animal experiments. On the other hand, they do provide a substantial advantage over cross-sectional correlational studies and, to a lesser extent, over multiple timepoint longitudinal studies. They constitute sets of strategies that warrant much greater use.

Acknowledgments—Discussions with members of the Academy of Medical Sciences Working Party on "The Identification of Environmental Causes of Disease: How Should We Decide What to Believe and When to Act?" have most helpfully contributed to my thinking on testing causal inferences and, hence, to the contents of this article. I am also most grateful for the comments and suggestions of anonymous reviewers and of the Editor.

REFERENCES

Anderson, K.E., Lytton, H., & Romney, D.M. (1986). Mothers' interactions with normal and conduct-disordered boys: Who affects whom? *Developmental Psychology*, *22*, 604–609.

Angrist, J.D., Imbens, G.W., & Rubin, D.B. (1996a). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

Angrist, J.D., Imbens, G.W., & Rubin, D.B. (1996b). Identification of causal effects using instrumental variables: Rejoinder. *Journal of the American Statistical Association*, *91*, 468–472.

Atladóttir, H.O., Parner, E.T., Schendel, D., Dalsgaard, S., Thomsen, P.H., & Thorsen, P. (2007). Time trends in reported diagnoses of childhood neuropsychiatric disorders. *Archives of Pediatric and Adolescent Medicine*, *161*, 193–198.

Baron, R.M., & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Beckett, C., Maughan, B., Castle, J., Colvert, E., Groothues, C., Kreppner, J., et al. (2006). Do the effects of early severe deprivation on cognition persist into early adolescence? Findings from the English and Romanian Adoptees Study. *Child Development*, *77*, 696–711.

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.

Bresnahan, M., Begg, M.D., Brown, A., Schaefer, C., Sohler, N., Insel, B., et al. (2007). Race and risk of schizophrenia in a US birth cohort: Another example of health disparity [Electronic version]. *International Journal of Epidemiology*, *36*, 751–758.

Brown, G.W., Harris, T.O., & Bifulco, A. (1986). Long-term effects of early loss of parent. In M. Rutter, C. Izard, & P. Reed (Eds.), *Depression in childhood: Developmental perspectives* (pp. 251–296). New York: Guilford Press.

Cahan, S., & Cohen, N. (1989). Age versus schooling effects on intelligence development. *Child Development*, *60*, 1239–1249.

Casas, J.P., Bautista, L.E., Smeeth, L., Sharma, P., & Hingorani, A.D. (2005). Homocystine and stroke: Evidence on a causal link from Mendelian randomization. *Lancet*, *365*, 222–232.

Case, A., Lubotsky, D., & Paxson, C. (2002). Economic status and health in childhood: The origins of the gradient. *American Economic Review*, *92*, 1308–1334.

Caspi, A. (2003). Life-course development: The interplay of social-selection and social-causation within and across generations. In P.L. Chase-Lansdale, K. Kiernan, & R.J. Friedman (Eds.), *Human development across lives and generations: The potential for change* (pp. 10–43). New York: Cambridge University Press.

Caspi, A., McClay, J., Moffitt, T.E., Mill, J., Marin, J., Craig, I.W., et al. (2002). Role of genotype in the cycle of violence in maltreated children. *Science*, *297*, 851–854.

Caspi, A., & Moffitt, T.E. (1991). Individual differences are accentuated during periods of social change: The sample case of girls at puberty. *Journal of Personality and Social Psychology*, *61*, 157–168.

Caspi, A., & Moffitt, T.E. (2006). Gene-environment interactions in psychiatry: joining forces with neuroscience. *Nature Reviews. Neuroscience*, *7*, 583–590.

Caspi, A., Moffitt, T.E., Cannon, M., McClay, J., Murray, R., Harrington, H., et al. (2005). Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the COMT gene: Longitudinal evidence of a gene × environment interaction. *Biological Psychiatry*, *57*, 1117–1127.

Caspi, A., Moffitt, T.E., Morgan, J., Rutter, M., Taylor, A., Arseneault, L., et al. (2004). Maternal expressed emotion predicts children's antisocial behavior problems: Using monozygotic-twin differences to identify environmental effects on behavioral development. *Developmental Psychology*, *40*, 149–161.

Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H.L., et al. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science*, *301*, 386–389.

Cochran, W.G., & Chambers, S.P. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A (General)*, *128*, 234–266.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., & Wynder, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, *22*, 173–203.

Costello, E.J., Compton, F.N., Keeler, G., & Angold, A. (2003). Relationships between poverty and psychopathology: A natural experiment. *Journal of the American Medical Association*, *290*, 2023–2029.

Davey Smith, G., & Ebrahim, S. (2003). "Mendelian randomization": Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, *32*, 1–22.

Davey Smith, G., & Ebrahim, S. (2005). What can Mendelian randomization tell us about modifiable behavioural and environmental exposures. *British Medical Journal*, *330*, 1076–1079.

de Groot, A.D. (1951). War and the intelligence of youth. *Journal of Abnormal Social Psychology*, *46*, 596–597.

D'Onofrio, B.M., Turkheimer, E., Eaves, L.J., Corey, L.A., Berg, K., Solaas, M.H., & Emery, R.E. (2003). The role of the Children of Twins design in elucidating causal relations between parent characteristics and child outcomes. *Journal of Child Psychology and Psychiatry*, *44*, 1130–1144.

D'Onofrio, B.M., Turkheimer, E., Emery, R.E., Slutske, W.S., Heath, A.C., Madden, P.A., & Martin, N.G. (2006). A genetically informed study of the processes underlying the association between parental marital instability and offspring adjustment. *Developmental Psychology*, *42*, 486–499.

D'Onofrio, B.M., Van Hulle, C.A., Waldman, I.D., Rodgers, J.L., Harden, K.P., Rathouz, P.J., & Lahey, B.B. (in press). Smoking during pregnancy and offspring externalizing problems: An exploration of genetic and environmental confounds. *Development and Psychopathology*.

Duyme, M., Dumaret, A.-C., & Tomkiewicz, S. (1999). How can we boost IQs of "dull children"? A late adoption study. *Proceedings of the National Academy of Sciences, USA*, *96*, 8790–8794.

Eaves, L.J., Maes, H.M., & Silberg, J.L. (2005). Revisiting the children of twins: Can they be used to resolve the environmental effects of dyadic parental treatment on child behavior? *Twin Research & Human Genetics*, *8*, 283–290.

Engfer, A., Walper, S., & Rutter, M. (1994). Individual characteristics as a force in development. In M. Rutter & D.F. Hay (Eds.), *Development through life: A handbook for clinicians* (pp. 79–111). Oxford, United Kingdom: Blackwell Scientific.

Entwisle, D.R., Alexander, K.L., & Olson, L.S. (2004). Young children's achievement in school and socioeconomic background. In D. Conley & K. Albright (Eds.), *After the bell: Family background, public policy and educational success* (pp. 86–108). London: Routledge.

Fergusson, D.M., & Horwood, L.J. (2000). Does cannabis use encourage other forms of illicit drug use? *Addiction*, *95*, 505–520.

Fergusson, D.M., Horwood, L.J., & Lynskey, M.T. (1992). Family change, parental discord and early offending. *Journal of Child Psychology and Psychiatry*, *33*, 1059–1075.

Foster, E.M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods*, *1*, 249–260.

Frith, C., & Frith, U. (in press). What can we learn from structural and functional brain imaging? In M. Rutter, D. Bishop, D. Pine, S. Scott, J. Stevenson, E. Taylor, & A. Thapar (Eds.), *Rutter's child and adolescent psychiatry* (5th ed.). Oxford, United Kingdom: Blackwell.

Ge, X., Conger, R.D., Cadoret, R.J., Neiderhiser, J.M., Yates, W., Troughton, E., & Stewart, M.A. (1996). The developmental interface between nature and nurture: A mutual influence model of child antisocial behavior and parenting. *Developmental Psychology*, *32*, 574–589.

Geoffroy, M.-C., Côté, S.M., Borge, A.I.H., Larouche, F., Séguin, J.R., & Rutter, M. (2007). Association between nonmaternal care in the first year of life and children's receptive language skills prior to school entry: The moderating role of socioeconomic status. *Journal of Child Psychology and Psychiatry*, *48*, 490–497.

Gill, R.D., & Robins, J.M. (2001). Causal inference for complex longitudinal data: The continuous case. *Annals of Statistics*, *29*, 1785–1811.

Grant, B.F., & Dawson, D.A. (1997). Age at onset of alcohol use and its association with DSM-IV alcohol abuse and dependence: Results from the National Longitudinal Alcohol Epidemiologic Survey. *Journal of Substance Abuse*, *9*, 103–110.

Gray, R., & Wheatley, K. (1991). How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant*, *7*(Suppl. 3), 9–12.

Heath, A., Lynskey, M.T., & Waldron, M. (in press). Child and adolescent substance use and substance use disorder. In M. Rutter, D. Bishop, D. Pine, S. Scott, J. Stevenson, E. Taylor, & A. Thapar (Eds.), *Rutter's child and adolescent psychiatry* (5th ed.). Oxford, United Kingdom: Blackwell.

Hernán, M.A., Herndandez-Diaz, S., & Robins, J.M. (2004). A structural approach to selection bias. *Epidemiology*, *15*, 615–625.

Heyns, B. (1978). Summer learning and the effects of schooling. New York: Academic.

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

Hoek, H.W., Brown, A.S., & Susser, E. (1998). The Dutch famine and schizophrenia spectrum disorders. *Social Psychiatry and Psychiatric Epidemiology*, *33*, 373–379.

Honda, H., Shimizu, Y., & Rutter, M. (2005). No effect of MMR withdrawal on the incidence of autism: A total population study. *Journal of Child Psychology and Psychiatry*, *46*, 572–579.

Irons, D.E., McGue, M., Iacono, W.G., & Oetting, W.S. (2007). Mendelian randomization: A novel test of the gateway hypothesis and models of gene-environment interplay. *Development and Psychopathology*, *19*, 1181–1195.

Jacobson, S.W., Jacobson, J.L., Sokol, R.J., Martier, S.S., & Chiodo, L.M. (1996). New evidence of neurobehavioral effects of in utero cocaine exposure. *Journal of Pediatrics*, *129*, 581–588.

Jaffee, S.R., Caspi, A., Moffitt, T.E., Polo-Tomas, M., Price, T.S., & Taylor, A. (2004). The limits of child effects: Evidence for genetically mediated child effects on corporal punishment but not on physical maltreatment. *Developmental Psychology*, *40*, 1047–1058.

Jencks, C., Smith, M., Acland, M., Bane, M.J., Cohen, D., Gintis, H., et al. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York: Basic Books.

Jones, P.B., & Fung, W.L.A. (2005). Ethnicity and mental health: The example of schizophrenia in the African-Caribbean population in Europe. In M. Rutter & M. Tienda (Eds.), *Ethnicity and causal mechanisms* (pp. 227–261). New York: Cambridge University Press.

Joseph, J. (2003). *The gene illusion: Genetic research in psychiatry and psychology under the microscope*. Ross on Wye, United Kingdom: PCCS Books.

Kamin, L.J., & Goldberger, A.S. (2002). Twin studies in behavioral research: A skeptical view. *Theoretical Population Biology*, *61*, 83–95.

Kandel, D.B. (1978). Homophily, selection and socialization in adolescent friendships. *American Journal of Sociology*, *84*, 427–436.

Katan, M.B. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, *1*, 507–508.

Katan, M.B. (2004). Commentary: Mendelian randomization, 18 years on. *International Journal of Epidemiology*, *33*, 10–11.

Keavney, B. (2004). Commentary: Katan's remarkable foresight: Genes and causality 18 years on. *International Journal of Epidemiology*, *33*, 11–14.

Keavney, B., Danesh, J., Paris, S., Palmer, A., Clark, S., Youngman, L., et al. (2006). Fibrinogen and coronary heart disease: Test of causality by "Mendelian randomization." *International Journal of Epidemiology*, *35*, 935–943.

Kendler, K.S. (2005). Psychiatric genetics: A methodological critique. *American Journal of Psychiatry*, *162*, 3–11.

Kendler, K.S., & Baker, J.H. (2007). Genetic influences on measures of the environment: a systematic review. *Psychological Medicine*, *37*, 615–626.

Kendler, K.S., & Prescott, C.A. (2006). *Genes, environment, and psychopathology: Understanding the causes of psychiatric and substance use disorders*. New York: Guilford Press.

Kim-Cohen, J., Caspi, A., Taylor, A., Williams, B., Newcombe, R., Craig, I.W., & Moffitt, T.E. (2006). MAOA, maltreatment, and gene-environment interaction predicting children's mental health: New evidence and a meta-analysis. *Molecular Psychiatry*, *11*, 903–913.

Kreppner, J., Rutter, M., Beckett, C., Castle, J., Colvert, E., Groothues, C., et al. (2007). Normality and impairment following profound early institutional deprivation: A longitudinal follow-up into early adolescence. *Developmental Psychology*, *43*, 931–946.

Laird, N.M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care*, *6*, 5–30.

Ludwig, J., & Miller, D.L. (in press). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*.

Luellen, J.K., Shadish, W.R., & Clark, M.H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review*, *29*, 530–558.

Lynch, S.K., Turkheimer, E., D'Onofrio, B.M., Mendle, J., & Emery, R.E. (2006). A genetically informed study of the association between harsh punishment and offspring behavioural problems. *Journal of Family Psychology*, *20*, 190–198.

Lynskey, M.T., Heath, A.C., Bucholz, K.K., Slutske, W.S., Madden, P.A., Nelson, E.C., et al. (2003). Escalation of drug use in early-onset cannabis users vs. co-twin controls. *Journal of the American Medical Association*, *289*, 427–433.

Lynskey, M.T., Vink, J.M., & Boomsma, D.I. (2006). Early onset cannabis use and progression to other drug use in a sample of Dutch twins. *Behavioural Genetics*, *10*, 1–6.

Mackie, J.L. (1965). Causes and conditions. *American Philosophical Quarterly*, *2*, 245–264.

Mackie, J.L. (1974). *The cement of the universe: A study of causation*. Oxford, United Kingdom: Oxford University Press.

MacMahon, B., Pugh, T., & Ipsen, J. (1960). *Epidemiologic methods*. Boston: Little Brown.

Madsen, K., Lauritsen, M.B., Pedersen, C.B., Thorsen, P., Plesner, A.-M., Andersen, P.H., & Mortensen, P.B. (2003). Thimerosal and the occurrence of autism: Negative ecological evidence from Danish population-based data. *Pediatrics*, *112*, 604–606.

Maldonado, G., & Greenland, S. (2002). Estimating causal effects. *International Journal of Epidemiology*, *31*, 422–429.

Marmot, M.G., & Syme, S.L. (1976). Acculturation and coronary heart disease in Japanese–Americans. *American Journal of Epidemiology*, *104*, 225–247.

Maynard, R.A. (Ed.). (1997). *Kids having kids: Economic costs and social consequences of teen pregnancy*. Washington, DC: Urban Institute Press.

McClellan, J.M., Susser, E., & King, M.C. (2006). Maternal famine, de novo mutations, and schizophrenia. *Journal of the American Medical Association*, *296*, 582–584.

McGue, M., & Iacono, W.G. (2005). The association of early adolescent problem behavior with adult psychopathology. *American Journal of Psychiatry*, *162*, 1118–1124.

Meade, T.W., Humphries, S.E., & De Stavola, B.L. (2006). Commentary: Fibrinogen and coronary heart disease. Test of causality by "Mendelian" randomization by Keavney et al. *International Journal of Epidemiology*, *35*, 944–947.

Miech, R.A., Caspi, A., Moffitt, T.E., Entner Wright, B.R., & Silva, P.A. (1999). Low socio-economic status and mental disorders: A longitudinal study of selection and causation during young adulthood. *American Journal of Sociology*, *104*, 1096–1131.

Mill, J.S. (1843). *A system of logic*. London: Parker.

Moe, V. (2002). Foster-placed and adopted children exposed in utero to opiates and other substances: Prediction and outcome at four and a half years. *Journal of Developmental and Behavioural Pediatrics*, *23*, 330–339.

Moffitt, T.E., Caspi, A., & Rutter, M. (2006). Measured gene-environment interactions in psychopathology: Concepts, research strategies, and implications for research, intervention, and public understanding of genetics. *Perspectives on Psychological Science*, *1*, 5–27.

Moffitt, T.E., & The E-Risk Study Team. (2002). Teenaged mothers in contemporary Britain. *Journal of Child Psychology and Psychiatry*, *43*, 727–742.

Neugebauer, R., Hoek, H.W., & Susser, E. (1999). Prenatal exposure to wartime famine and development of antisocial personality disorder in early adulthood. *Journal of the American Medical Association*, *282*, 455–462.

Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B.L., Whittaker, J.C., & Leon, D.A. (2006). Limits to causal inference based on Mendelian randomization: A comparison with randomized controlled trials. *American Journal of Epidemiology*, *163*, 397–403.

Obel, C., Olsen, J., Henriksen, T.B., Moilannen, I., Rodriguez, A., Linnet, K.M., et al. (2007). *Exposure to maternal smoking during pregnancy and diagnosed hyperkinetic disorder in the offspring: A sibling design based on complete follow up of all children born in Finland 1987–2001*. Manuscript submitted for publication.

O'Connor, T.G., Deater-Deckard, K., Fulker, D., Rutter, M., & Plomin, R. (1998). Genotype-environment correlations in late childhood and early adolescence: Antisocial behavioral problems and coercive parenting. *Developmental Psychology*, *34*, 970–981.

Peterson, B.S. (2003). Conceptual, methodological, and statistical challenges in brain imaging studies of developmentally based psychopathologies. *Development and Psychopathology*, *15*, 811–832.

Plomin, R., DeFries, J.C., & Loehlin, J.C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin*, *84*, 309–322.

Prescott, C.A., & Kendler, K.S. (1999). Age at first drink and risk for alcoholism: A noncausal association. *Alcoholism: Clinical and Experimental Research*, *23*, 101–107.

Pulkkinen, L., Kaprio, J., & Rose, R. (2006). *Socioemotional development and health from adolescence to adulthood*. Cambridge, United Kingdom: Cambridge University Press.

Pynoos, R.S., Frederick, C., Nader, K., Arroyo, W., Steinberg, A., Eth, S., Eth, S., et al. (1987). Life threat and post-traumatic stress in school–age children. *Archives of General Psychiatry*, *44*, 1057–1063.

Ravussin, E., Valencia, M.E., Esparza, J., Bennett, P.H., & Schulz, L.O. (1994). Effects of a traditional lifestyle on obesity in Pima Indians. *Diabetes Care*, *17*, 1067–1074.

Robins, J.M. (2001). Data, design and background knowledge in etiologic inference. *Epidemiology*, *11*, 313–320.

Robins, J.M., & Greenland, S. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, *91*, 456–458.

Robins, J.M., Hernán, M.A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550–560.

Rosenbaum, P.R. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, *91*, 465–468.

Rosenbaum, P.R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rosenbaum, P.R., & Rubin, D.B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, *45*, 212–218.

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

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

Rothman, K.J., & Greenland, S. (2002). Causation and causal inference. In R. Detels, J. McEwen, R. Beaglehole, & H. Tanaka (Eds.), *Oxford textbook of public health* (4th ed., pp. 641–653). Oxford, United Kingdom: Oxford University Press.

Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1–26.

Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*, 318–328.

Rubin, D.B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers? *Journal of the American Statistical Association*, *81*, 961–962.

Rutter, M. (1997). Comorbidity: Concepts, claims and choices. *Criminal Behaviour and Mental Health*, *7*, 265–286.

Rutter, M. (2005). Incidence of autism spectrum disorders: Changes over time and their meaning. *Acta Paediatrica*, *94*, 2–15.

Rutter, M. (2006). *Genes and behavior: Nature-nurture interplay explained*. Oxford, United Kingdom: Blackwell.

Rutter, M. (2007). Gene-environment interdependence. *Developmental Science*, *10*, 12–18.

Rutter, M., Andersen-Wood, L., Beckett, C., Bredenkamp, D., Castle, J., Groothues, C., et al. (1999). Quasi-autistic patterns following severe early global privation. *Journal of Child Psychology and Psychiatry*, *40*, 537–549.

Rutter, M., Caspi, A., & Moffitt, T.E. (2003). Using sex differences in psychopathology to study causal mechanisms: unifying issues and research strategies. *Journal of Child Psychology and Psychiatry*, *44*, 1092–1115.

Rutter, M., Champion, L., Quinton, D., Maughan, B., & Pickles, A. (1995). Understanding individual differences in environmental risk exposure. In P. Moen, G.H. Elder Jr., & K. Lüscher (Eds.), *Examining lives in context: Perspectives on the ecology of human development* (pp. 61–93). Washington, DC: American Psychological Association.

Rutter, M., Colvert, E., Kreppner, J., Beckett, C., Castle, J., Groothues, C., et al. (2007). Early adolescent outcomes for institutionally-deprived and non-deprived adoptees: I. Disinhibited attachment. *Journal of Child Psychology and Psychiatry*, *48*, 17–30.

Rutter, M., Dunn, J., Plomin, R., Simonoff, E., Pickles, A., Maughan, B., et al. (1997). Integrating nature and nurture: Implications of person-environment correlations and interactions for developmental psychopathology. *Development & Psychopathology*, *9*, 335–366.

Rutter, M., & The English and Romanian Adoptees Study Team. (1998). Developmental catch-up, and deficit, following adoption after severe global early privation. *Journal of Child Psychology and Psychiatry*, *39*, 465–476.

Rutter, M., & Madge, N. (1976). *Cycles of disadvantage: A review of research*. London: Heinemann Educational.

Rutter, M., & Maughan, B. (2002). School effectiveness findings: 1979–2002. *Journal of School Psychology*, *40*, 451–475.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. London: Open Books.

Rutter, M., Moffitt, T.E., & Caspi, A. (2006). Gene-environment interplay and psychopathology: Multiple varieties but real effects. *Journal of Child Psychology and Psychiatry*, *47*, 226–261.

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

Rutter, M., Thorpe, K., Greenwood, R., Northstone, K., & Golding, J. (2003). Twins as a natural experiment to study the causes of mild language delay: I. Design; twin-singleton differences in language, and obstetric risks. *Journal of Child Psychology and Psychiatry*, *44*, 326–334.

Rutter, M., & Tienda, M. (Eds.). (2005). *Ethnicity and causal mechanisms*. New York: Cambridge University Press.

Sampson, R.J., & Laub, J.H. (1993). *Crime in the making: Pathways and turning points through life*. Cambridge, MA: Harvard University Press.

Sampson, R.J., Laub, J.H., & Wimer, C. (2006). Does marriage reduce crime? A counterfactual approach to within-individual causal effects. *Criminology*, *44*, 465–508.

Schwartz, S., & Susser, E. (2006). What is a cause? In E. Susser, S. Schwartz, A. Morabia, & E.J. Bromet. *Psychiatric epidemiology: Searching for the causes of mental disorders* (pp. 33–42). Oxford, United Kingdom: Oxford University Press.

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., Luellen, J.K., & Clark, M.H. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R.R. Bootzin & P.E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143–157). Washington, DC: American Psychological Association.

Shavelson, R.J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Silberg, J.L., & Eaves, L.J. (2004). Analysing the contributions of genes and parent-child interaction to childhood behavioural and emotional problems: A model for the children of twins. *Psychological Medicine*, *34*, 347–356.

Singer, L.T., Arendt, R., Minnes, S., Farkas, K., & Salvator, A. (2000). Neurobehavioral outcomes of cocaine-exposed infants. *Neurotoxicology and Teratology*, *22*, 653–666.

Singer, L.T., Minnes, S., Short, E., Arendt, R., Farkas, K., Lewis, B., et al. (2004). Cognitive outcomes of preschool children with prenatal cocaine exposure. *Journal of the American Medical Association*, *291*, 2448–2456.

Sonuga-Barke, E.J.S., Beckett, C., Kreppner, J., Castle, J., Colvert, E., Stevens, S. et al. (2007). *Is subnutrition necessary for a poor outcome following severe and pervasive early institutional deprivation? Brain growth, cognition and mental health.* Manuscript submitted for publication.

Spergel, I.A. (1990). Youth gangs: Continuity and change. In M. Tonry & N. Morris (Eds.), *Crime and justice: A review of research* (Vol. 12, pp. 171–275). Chicago: University of Chicago Press.

Stattin, H., & Magnusson, D. (1990). *Paths through life: Vol. 2. Pubertal maturation in female development*. Hillsdale, NJ: Erlbaum.

St. Clair, D., Xu, M., Wang, P., Yu, Y., Fang, Y., Zhang, F., et al. (2005). Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959–1961. *Journal of the American Medical Association*, *294*, 557–562.

Stein, Z.A., Susser, M., Saenger, G., & Marolla, F. (1975). *Famine and human development: The Dutch hunger winter of 1944–1945*. New York: Oxford University Press.

Susser, E., Neugebauer, R., Hoek, H.W., Brown, A.S., Lin, S., Labovitz, D., & Gorman, J.M. (1996). Schizophrenia after prenatal famine: Further evidence. *Archives of General Psychiatry*, *53*, 25–31.

Susser, E., Schwartz, S., Morabia, A., & Bromet, E.J. (2006). *Psychiatric epidemiology: Searching for the causes of mental disorders*. New York: Oxford University Press.

Thistlewaite, D.L., & Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, *51*, 309–317.

Thomas, A., Chess, S., & Birch, H.G. (1968). *Temperament and behaviour disorders in childhood*. New York: New York University Press.

Thornberry, T.P., Krohn, M.D., Lizotte, A.J., & Chard-Wiershem, D. (1993). The role of juvenile gangs in facilitating delinquent behavior. *Journal of Research in Crime and Delinquency*, *30*, 55–87.

Thorpe, K., Rutter, M., & Greenwood, R. (2003). Twins as a natural experiment to study the causes of mild language delay: II. Family interaction risk factors. *Journal of Child Psychology and Psychiatry*, *44*, 342–355.

Tobin, M.D., Minelli, C., & Burton, P.R. (2004). Development of Mendelian randomization: From hypothesis test to "Mendelian disconfounding." *International Journal of Epidemiology*, *33*, 21–25.

Valencia, M.E., Bennett, P.H., Ravussin, E., Esparza, J., Fox, C., & Schulz, L.O. (1999). The Pima Indians in Sonora, Mexico. *Nutrition Reviews*, *57*, S55–S58.

Weinberg, C.R. (1993). Toward a clearer definition of confounding. *American Journal of Epidemiology*, *137*, 1–8.

Wootton, B. (1959). *Social science and social pathology*. London: George Allen & Unwin.

Yule, W. (2002). Post-traumatic stress disorders. In M. Rutter & E. Taylor (Eds.), *Child and adolescent psychiatry* (4th ed., pp. 520–528). Oxford, United Kingdom: Blackwell.

Yule, W., Udwin, O., & Murdoch, K. (1990). The Jupiter sinking: Effects on children's fears, depression and anxiety. *Journal of Child Psychology and Psychiatry*, *31*, 1051–1061.