Genetic Epidemiology 11:463–472 (1994)

Direction of Causation: Reply to Commentaries

Michael C. Neale, David L. Duffy, and Nicholas G. Martin

Department of Psychiatry, Medical College of Virginia, Richmond (M.C.N.); Queensland Institute for Medical Research, The Bancroft Centre, Brisbane, Australia (D.L.D., N.G.M.)

We reply to the commentaries on the lead papers by Neale et al. [1994a] and Duffy and Martin [1994]. Topics covered include power calculations, cross-sectional measurement vs. lifetime reports, the appropriateness of the direction of causation (DOC) model, extensions to study causation between the latent variables, sampling of subjects, and heterogeneity. We consider the potential of combining genetically informative research designs with multivariate longitudinal and experimental methods. © 1994 Wiley-Liss, Inc.

Key words: genes, environment, twins, parents, causation, longitudinal, instrumental variable, design, methodology, structural equation modeling, Mx, LISREL, causal inference

INTRODUCTION

Science, like time, does not generally stand still, and it has not done so since the lead articles were submitted. Two recent and significant contributions to direction of causation (DOC) modeling are chapter 13 in Neale and Cardon [1992] and the article by Heath et al. [1993], which address several issues that were not discussed in our lead articles. In particular, these sources explore the potential for imbalance of error of measurement to bias the estimates of reciprocal causal influence. Heath et al. [1993] provide some power calculations. Further consideration of the use of information from relatives to resolve models that are normally underidentified may be found in Neale et al. [1994b], where the fit of multiple regression models is shown to change according to which variable is selected as dependent. In the present article we discuss the issues raised by the commentators that are not considered by these other papers.

Address reprint requests to Michael C. Neale, Department of Psychiatry, Medical College of Virginia, Box 710, Richmond, VA 23298.

CAREY [1994]

We are grateful that this author points out that inferences about direction of causation can be made using data from relatives, even without a genetically informative study design. This point has also been made explicit in the article by Neale et al. [1993]. We used twins for our examples partly to ease understanding and partly because of the availability of appropriate multivariate data.

With respect to power calculations, Neale et al. [1989c] presented three-dimensional (3-D) graphical displays of the power to reject false causal models. However, it is not easy to summarize these results, because the number of parameters involved in even a bivariate model is large enough to generate a very large number of combinations of true and false models. This may be part of the reason that they have never been published! Nevertheless, Heath et al. [1993] tabulate power for a variety of true models, including some where variables are measured imperfectly.

GOLDBERG AND RAMAKRISHNAN [1994]

The comments these authors make are generally apposite. Especially useful is the discussion of temporary exposure to factors that cause later disease. If the biological causative pathway of interest were lagged and the exposure varied with time, it would obviously be foolish to use contemporaneous measures alone. Most of the variables chosen for the examples in Duffy and Martin [1994] are fairly stable over time. Moreover, in much of social and medical science we are forced to rely on *lifetime* assessments—using questions of the form "Have you ever...?" Ideally, these queries would be answered accurately and honestly, so that we could obtain good measures of earlier temporary exposure. In practise, simple lapses of memory or recall bias may cloud the issue. If such processes were unrelated to the outcome (and did not correlate with relatives' reporting styles), then they would accumulate as measurement error. However, if they were correlated with either the exposure or outcome phenotypes, then more serious consequences would follow. Consider, e.g., the question "Have you ever had amnesia?"—where the answer "No" may be quite uninformative.

On occasion, "problems" such as recall bias or memory deficits turn out to be opportunities for the elucidation of substantively meaningful processes. For example, the lead article by Neale et al. [1994a] specifically discusses the use of DOC models for detection of recall bias in life-history frameworks (which would include the case-control study). In their example, adult depression is correlated with recalled parental coldness, so the question becomes "Is the direction of causation actually from present state of mind to past?" or "Does an episode of depression cause biased responses about one's parents?"

Goldberg and Ramakrishnan's use of the term "confounding" from the single-cause to single-outcome thinking of a lot of epidemiology understates a strength of the direction-of-causation twin method, whether applied to cross-sectional, historical, or concurrent longitudinal designs. Let us further discuss the same or similar (but more plausible for this purpose) examples.

Study 1

Does smoking cause asthma? Here it seems highly probable that individuals prone to developing respiratory disease later in life are more likely to find cigarette smoke irritating earlier in life and so never take up the habit. A longitudinal study might infer that smoking at the start of the study protects, or is uncorrelated with, later development of asthma. Errors of measurement notwithstanding, a DOC model might "correctly" show a pathway from later asthma to earlier non-smoking.

Study 2

Do men who decide to undergo vasectomy differ in terms of unmeasured risk factors for prostate cancer from those who do not? These factors might be social class or diet related, but would lead to genetic or shared environmental correlations between twins for prostate cancer risk.

Study 3

Does the relationship and direction of the causation between vitamin E and Parkinson's disease vary with time? Is vitamin E a risk factor for Parkinson's disease, but does dietary vitamin E intake decline with the onset of disease? Undoubtedly, the former hypothesis is far more interesting and should be addressed appropriately through, e.g., assessment of vitamin E intake at or before diagnosis.

We hope that these comments confirm that only randomized experiments give unambiguous information about the direction of causation. We can only reiterate that the method is not a panacea for the problem of inferring direction of causation in non-experimental settings.

McARDLE [1994]

McArdle's comments are, as usual, incisive and original, and raise several new issues. We discuss each in turn.

Biometric Group Differences

First is the remark that "biometric group differences can help resolve other important aspects of structural equation systems." This is patently obvious for the estimation of, e.g., additive genetic and shared environmental variance, but how does it help in more conventional contexts? A classic example is the resolution of assortative mating and martial interaction with cross-sectional twin data [Heath, 1987]. Here the general idea is that twin pair resemblance is expected to differ as a function of whether the twins are both married, both single, or one of each. Another example is the resolution of mate selection based on the environment or the phenotype [Heath et al., 1985]. Because monozygotic (MZ) twins correlate more highly, their spouses should resemble each other more strongly if mate selection is based on the phenotypes of the twins. Yet another example is the explicit modeling of mediating variables in multiple regression [Neale et al., 1994b]. Others include the relationship of disease liability to age at onset [Neale et al., 1989a]; rater bias [Heath et al., 1985; Neale and Stevenson, 1989]; dimensionality of scales [Heath et al., 1991]; non-invariance to orthogonal rotation in factor analysis [Neale et al., 1993]; and tests for non-random sampling [Neale and Eaves, 1993]. In general, while modeling of a

dataset collected from relatives is more complex than modeling one from unrelated subjects, it is because it is more informative and will permit tests of hypotheses that are otherwise mere assumptions.

Appropriateness of the DOC Model

The second question—"Is DOC an appropriate causal model?"—raises the measurement error problem with a simple example, as shown in Figure 1. The problem with this case is that there is no difference in the genetic architecture between the two phenotypes, *X* and *Y*, except for measurement error. It would be surprising to find that, e.g., we could learn about direction of causation by simply adding a random number to one of our measures.

We note that where measurement error is zero on one variable, that this is only a submodel (Model 1 in the Duffy and Martin paper [1994]), since in most of the examples chosen, both variables have structured unique factors (in biometrical terms, specific genetic, or shared environmental determinants). Duffy and Martin commented that the situation where one factor has no specific determinants limits the range of testable hypotheses, a point emphasized by McArdle's example.

In the second paragraph of the second section of McArdle's commentary, he suggests that because measurement error can cause erroneous conclusions about di-

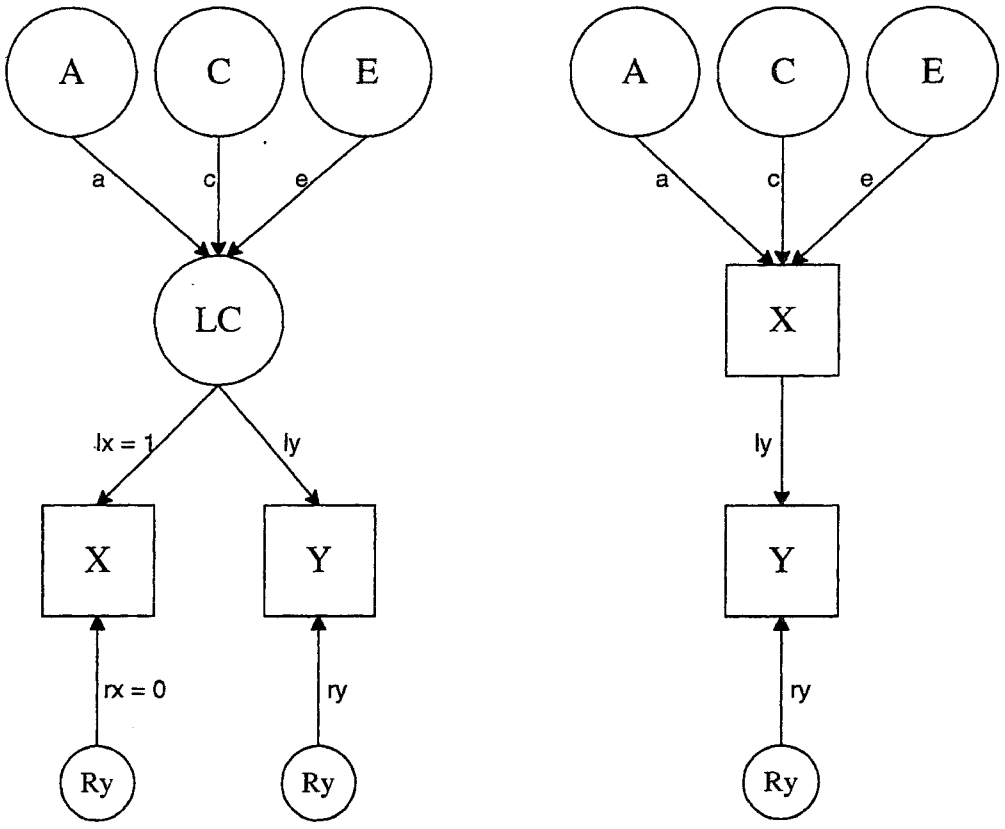

Fig. 1. Path diagrams illustrating the equivalence of the common factor and causal models when one variable has no residual error.

rection of causation, we should ideally test these hypotheses with factor scores which are, in theory, error free. Furthermore, at least three measured variables are needed to reject the (single common) factor model. McArdle seems to refer to the use of the data from unrelated subjects to reject the model, but this is not the only way it could fail. It might prove inadequate when fitted as a "psychometric factor" model to pairwise data, despite being adequate for unrelated individuals. When we move to the use of the pair or family as the unit of analysis, we are using the same data summary as would be used to fit the direction-of-causation models—and the arguments are in danger of becoming circular. Nevertheless, it seems wise to take this more complex route *initially* to give the factor model its strongest possible test.

The choices of research design and which models to test are not independent of each other nor of the particular variables under study. For example, if we had three trained physicians measure each subject's subscapular skinfold, a single factor model would a priori seem more plausible than if we used three different measures (e.g., self-report questionnaire, personal interview, and some physiological test) to assess depression. Thus for study design we might wish to take short-interval test-retest measures to deal with the error problem. We should, however, remain aware of the possible effects of practise or memory on physical and psychological tests.

Biometric Causal Models

The third issue raised in this commentary is perhaps the most exciting. The question is "How far can we assess direction of causation between the latent genetic and environmental factors?" While at first this might seem to be virgin territory, we do in fact know quite a lot about models that specify a chain of causal connections. Such "simplex" or "Markov" structures form one of the most widely used paradigms in longitudinal research, both genetic [Eaves et al., 1986; Boomsma et al., 1989; Cardon et al., 1992; Neale and Cardon, 1992] and non-genetic [Guttman, 1954; Wohlwill, 1973]. Within genetic research, the paper by Hewitt et al. [1988] is particularly relevant here, as it discusses the number of occasions required to reject a set of models that include both simplex causal processes and common factor effects. For many purposes, this number seems to be four, which may therefore be considered the minimum number of variables to be linked in a unidirectional causal chain for hypothesis testing in multivariate data. However, we must emphasize that the longitudinal study has a considerable advantage over the analogous multivariate DOC one, namely that causal paths going back in time may be eliminated. A second advantage is the possibility that the paths from occasion to occasion are all equal in the longitudinal study, an argument that is difficult to retain in the multivariate causal model. These differences imply that the multivariate DOC study would require more than four variables to allow the resolution of biometric causal processes in addition to phenotypic ones. Further research in this area would be of considerable value.

While much attention has been devoted to multivariate and longitudinal designs and models individually, there has been little work on their *joint* use. One application was described by Boomsma et al. [1989]; other uses include the relationship between personality and major depressive disorder [Katz and McGuffin, 1987; Kendler et al., 1993]. None of these studies examines the potential of the longitudinal multivariate design for resolving causal relations between the observed variables. Here is another

area ripe for investigation. While McArdle states that the "DOC paradigm probably cannot help us," it would indeed be nice to know for sure!

We seem to be similarly ignorant about the potential of a mixture of genetically informative and experimental designs. From an epidemiological perspective, experimental designs are often unethical, impractical, or somewhat unrealistic compared to the natural history of diseases. Nevertheless, we agree that they offer the most compelling evidence for causation available and should be used wherever possible.

Sampling of Subjects

McArdle succinctly delivers a slew of questions under this heading. First is whether or not all subjects have the same direction of causation. Structural models of covariances provide an averaged view of the sample and may well mask heterogeneity. Presumably, if half the subjects suffered depression because of life events and half caused themselves life events because they were depressed, a reciprocal causal model would provide the best fit to the data. However, it would be a good idea to test this with some simulations, because it seems possible that population heterogeneity could give rise to patterns of covariance that could not be accounted for by a stable reciprocal causal model (i.e., one with a matrix of reciprocal paths whose eigenvalues lie within the unit circle [Jöreskog and Sörbom, 1989]). Perhaps better insights could be obtained through analysis of the raw data and examination of the fit of individual (and family) likelihoods [Hopper and Culross, 1983; Neale, 1991]. A model of unidirectional causation should fit one part of the population well and the other part poorly, and vice versa when the direction is changed. However, the estimates in both cases would be biased by the different causal processes in the two subpopulations. Bootstrapping procedures might help to detect heterogeneity while minimizing bias. A better approach would be to fit a joint model, estimating both the parameters for each subgroup together with the probabilities of group membership.

The question of whether all subjects are at the same point in the dynamic process is a form of heterogeneity that might be easier to detect through the use of ancillary variables such as age. At a crude level we could split the sample into "old" and "young" groups and test for heterogeneity of parameter estimates. More sophisticated approaches that use continuous indices of heterogeneity present no special problems in theory, but would seem to require software development [Neale and Cardon, 1992; ch. 17].

Whether we espouse a particular causal theory or simply respond to the convenience of having available data relates to fundamental aspects of scientific method. As illustrated by Eaves et al. [1989] and Neale and Cardon [1992: Fig. 1.6], models lie at the interface between theory and data. It is thus natural for either theory or data analysis to generate new models. But perhaps McArdle's comment is most pertinent in the question of how we assess goodness-of-fit. To fit a set of models according to theoretical guidelines is, statistically speaking, quite a different procedure from allowing the empirical observation of non-significant parameters to guide our path to "the most parsimonious model."

McArdle's question about whether the direction of causation is the same for MZ and dizygotic (DZ) pairs is easier to answer. In the structural model we assume that it is. If in reality it were to differ, there would probably be some empirical conse-

quences for the phenotypic variances of the two groups, just as there are when twins cooperate or compete with one another [Eaves, 1976; Carey, 1986]. The model predicts the same expected variance for both MZ and DZ groups, so significant group heterogeneity would cause it to fit badly.

Non-random sampling of subjects can lead to biased results in almost any study; the DOC method is no exception. Just how the parameter estimates change is of course a function of the type of non-random sampling. Treatments for hard and soft selection [e.g., Neale et al., 1989; Martin and Wilson, 1982; van Eerdewegh, 1982] show a non-linear reduction of the correlation when the sampling of a subject depends on their position on the scale being measured. For a substantial range of initial correlations, truncation can lead to underestimation of the impact of shared environmental effects. Should this type of non-randomness exist for only one of the variables in a DOC analysis, incorrect inferences about causality would be quite possible. Fortunately, many twin studies offer a test for non-random sampling by comparing pairs concordant for participation with those from which only one twin was available [Neale and Cardon, 1992; Neale and Eaves, 1993]. In principle, these data also allow inference about the relationship between the genetic and environmental factors and subject selection. It may prove more difficult to detect non-random sampling as a function of pair resemblance [e.g., Lykken et al., 1987].

Generally, we should not ask whether or not we need other groups of relatives; the question is *when*. Initial studies typically focus on highly informative sets of relatives such as MZ and DZ twins or adopted and natural siblings and build on these designs if the results and funding agencies so warrant. The same will presumably apply to studies aimed primarily at assessing causal relations between variables. To answer the specific questions on whether we could estimate parameters reflecting $A \leftrightarrow E$ correlation and $A \times E$ interaction, it seems that some form of measurement of the environment is required for the former. This assessment of environmental indices could be either direct or indirect, as with models of cultural transmission where parents' phenotypes may be modeled as either a cause of their children's environment or as an index of their own shared environment which is transmitted to their children [Eaves et al., 1978; Fulker, 1982; Neale and Cardon, 1992; ch. 17]. For $A \times E$ interaction, we could, in principle, use either measured indices or higher moments or other statistics such as sum-difference regressions. While the former approach has been described and applied in a number of contexts [e.g., see Neale and Cardon, 1992; ch. 11; Kendler et al., 1991], the latter has not [Molenaar et al., 1990]. Melding these methods with DOC models will not be simple, even with improvements in structural modeling software.

Grammar

Because of the limitations of direction-of-causation hypothesis testing, McArdle is right to emphasize the confirmatory approach through falsification of a priori hypotheses, rather than exploratory model fitting. We strongly agree that terminology such as "genes cause X" should be avoided. Interestingly, "X causes genes" is rarely considered a problem, although it is implicit in models of $A \times E$ interaction. From a different perspective, natural selection may be said to "cause" the genotype, particularly the presence of genetical non-additivity [Fisher, 1958]. These processes are generally beyond the scope of many studies, whose primary foci are usually current social and medical problems.

CONCLUSION

If, among other advantages, studies of relatives are capable of inferring causation from correlation, then one begins to wonder why anyone would *not* collect data from this type of sample. Twin, family, and adoption studies can address every question answered by studies of unrelated subjects, and many more besides. In the unlikely event that none of the additional scope of behavior genetic designs were of interest, then there would be a small advantage to collecting information from unrelated individuals: their statistical independence yields slightly greater power. Thus sample sizes may be smaller when familial resemblance is not an issue. The choice between a small, efficient study of unrelated persons or one with the potential to resolve numerous confounded effects should be made on a case-by-case basis. However, as more and more capabilities of the genetically informative study are discovered, so the balance tilts in its favor.

One area that neither the lead articles nor the commentaries directly addressed is the action of mediating variables. It seems that the action of genes through enzymes is often catalytic; i.e., A will cause B as long as catalyst C is present in sufficient quantities. The same may be true for environmental or phenotypic processes; e.g., social support may serve to decrease depression by reducing the adverse impact of life events. Likewise, the medication may occur across modalities, so that, e.g., genetically controlled substrate oxidation might mediate the environmental effects of diet composition on phenotypic measures of body fat. Merely placing the mediating variable in the middle of a causal chain in a structural equation model does not adequately model this type of process. Rather, we wish to allow the size of the path coefficient from cause to effect to be a function of the mediating variable. One such model has been described for continuous $A \times E$ interaction [Neale and Cardon, 1992: ch. 17] but its potential applications are legion.

The development of DOC models has implications for future research designs. Clearly, any single approach has its limitations, as does comparing results across studies using different methods. The major difficulties with *joint* design types seem to be their cost and logistics. If these problems can be overcome, we should be aiming at a global design for genetically informative multivariate longitudinal experimental studies.

ACKNOWLEDGMENTS

M.C.N. was supported by ADAMHA grants MH-40828, AG-04954, and MH-45268.

REFERENCES

Boomsma DI, Martin NG, Molenaar PCM (1989): Factor and simplex models for repeated measures: Application to two psychomotor measures of alcohol sensitivity in twins. Behav Genet 19:79–96.

Cardon LR, Fulker DW, DeFries JC, Plomin R (1992): Continuity and change in general cognitive ability from 1 to 7 years. Dev Psychol 28:64–73.

Carey G (1986): A general multivariate approach to linear modeling in human genetics. Am J Hum Genet 39:775–786.

Carey G (1994): Direction of causality: A comment. Genet Epidemiol 11:473–475.

Duffy DL, Martin NG (1994): Inferring the direction of causation in cross-sectional twin data: Theoretical and empirical considerations. Genet Epidemiol 11:483–502.

Eaves LJ (1976): A model for sibling effects in man. Heredity 36:205–214.
Eaves LJ, Last KA, Young PA, Martin NG (1978): Model-fitting approaches to the analysis of human behavior. Heredity 41:249–320.
Eaves LJ, Long J, Heath AC (1986): A theory of developmental change in quantitative phenotypes applied to cognitive development. Behav Genet 16:143–162.
Eaves LJ, Eysenck HJ, Martin NG (1989): "Genes, Culture and Personality: An Empirical Approach." London: Oxford University Press.
Fisher RA (1958): "The Genetical Theory of Natural Selection." 2nd Ed. New York: Dover.
Fulker DW (1982): Extensions of the classical twin method. In Bonne-Tamir B (ed): "Human Genetics, Part A: The Unfolding Genome." New York: Alan R. Liss, pp 395–406.
Goldberg J, Ramakrishnan V (1994): Commentary: Direction of causation models. Genet Epidemiol 11:457–461.
Guttman L (1954): A new approach to factor analysis: The radex. In Lazarsfeld PF (ed): "Mathematical Thinking in the Social Sciences." Glencoe, IL: Free Press, pp 258–349.
Heath AC (1987): The analysis of marital interaction in cross-sectional twin data. Acta Genet Med Gemellol 36:41–49.
Heath AC, Berg K, Eaves LJ, Solaas MH, Sundet J, Nance WE, Corey LA, Magnus P (1985): No decline in assortative mating for educational level. Behavior Genetics 15:349–370.
Heath A, Meyer JM, Jardine R, Martin NG (1991): The inheritance of alcohol consumption in a general population twin sample. I. Multidimensional scaling of quantity and frequency data. J Stud Alcohol 52:345–352.
Heath AC, Kessler RC, Neale MC, Hewitt JK, Eaves LJ, Kendler KS (1993): Testing hypotheses about direction-of-causation using cross-sectional family data. Behav Genet 23:29–50.
Hewitt JK, Eaves LJ, Neale MC, Meyer JM (1988): Resolving causes of developmental continuity or "tracking." I. Longitudinal twin studies during growth. Behav Genet 18:133–151.
Hopper JL, Culross PR (1983): Covariation between family members as a function of cohabitation history. Behav Genet 13:459–471.
Jöreskog KG, Sörbom D (1989): "LISREL 7: A Guide to the Program and Applications." 2nd Ed. Chicago: SPSS, Inc.
Katz R, McGuffin P (1987): Neuroticism in familial depression. Psychol Med 17:155–162.
Kendler KS, Neale MC, Heath AC, Kessler RC, Eaves LJ (1991): Life events and depressive symptoms: A twin study perspective. In McGuffin P and Murray R (eds): "The New Genetics of Mental Illness." London: Butterworth Heinemann, pp 144–162.
Kendler KS, Neale, MC, Kessler RC, Heath AC, Eaves LJ (1993): A longitudinal twin study of personality and depression in women. Arch Gen Psychiatry 50:853–862.
Lykken DT, McGue M, Tellegen A (1987): Recruitment bias in twin research: The rule of two-thirds reconsidered. Behav Genet 17:343–362.
Martin NG, Wilson SR (1982): Bias in the estimation of heritability from truncated samples of twins. Behav Genet 12:467–472.
McArdle JJ (1994): Appropriate questions about causal inference from "Direction of Causation" analyses. Genet Epidemiol 11:477–482.
Molenaar PCM, Boomsma DI, Neeleman D, Dolan CV (1990): Using factor scores to detects $g \times e$ interactive origin of "pure" genetic or environmental factors obtained in genetic covariance structure analysis. Genet Epidemiol 7:83–100.
Neale MC (1991): "Mx: Statistical Modeling." Richmond: Department of Human Genetics.
Neale MC (1994): Factorial invariance in studies of relatives. Unpublished manuscript.
Neale MC, Cardon LR (1992): "Methodology for Genetic Studies of Twins and Families." The Netherlands: Kluwer Academic Publishers.
Neale MC, Eaves LJ (1993): Estimating and controlling for the effects of volunteer bias with pairs of relatives. Behav Genet 23:271–278.
Neale MC, Stevenson J (1989): Rater bias in the EASI temperament scales: A twin study. J Pers Soc Psychol 56:446–455.
Neale MC, Eaves JL, Hewitt JK, MacLean CJ, Meyer JM, Kendler KS (1989a): Analyzing the relationship between age of onset and risk to relatives. Am J Hum Genet 45:226–239.
Neale MC, Eaves LJ, Kendler KS, Hewitt JK (1989b): Bias in correlations from selected samples of relatives: The effects of soft selection. Behav Genet 19:163–169.

Neale MC, Hewitt JK, Heath AC, Eaves LJ (1989c): The power of multivariate and categorical twin studies. Presented at the 6th International Congress of Twin Studies, Rome.

Neale MC, Walters EW, Heath AC, Kessler RC, Pérusse D, Eaves LJ, Kendler KS (1994a): Depression and parental bonding: Cause, consequence or genetic covariance? Genet Epidemiol 11:503–522.

Neale MC, Eaves LJ, Hewitt JK, Kendler KS (1994b): Multiple regression with data collected from relatives. Multivariate Behav Res 29:33–60.

van Eerdewegh P (1982): Statistical selection in multivariate systems with applications in quantitative genetics. Unpublished doctoral dissertation, St. Louis: Washington University.

Wohlwill JF (1973): "The Study of Behavioral Development." New York: Academic Press.