Published in final edited form as:

J Abnorm Psychol. 2010 May ; 119(2): 307–319. doi:10.1037/a0017552.

Biased Parameter Estimates and Inflated Type I Error Rates In Analysis Of Covariance (and Analysis of Partial Variance) Arising From Unreliability: Alternatives And Remedial Strategies

Richard E. Zinbarg^{1,2}, Satoru Suzuki¹, Amanda A. Uliaszek¹, and Alison R. Lewis¹

¹Northwestern University

²The Family Institute at Northwestern University

Abstract

Miller and Chapman (2001) argued that one major class of misuse of analysis of covariance (ANCOVA) or its multiple regression counterpart, analysis of partial variance (APV), arises from attempts to use ANCOVA/APV to answer a research question that is not meaningful in the first place. Unfortunately, there is another misuse of ANCOVA/APV that arises frequently in psychopathology studies even when addressing consensually meaningful research questions. This misuse arises from inflated type I error rates in ANCOVA/APV inferential tests of the unique association of the independent variable with the dependent variable when the covariate and independent variables are correlated and measured with error. Alternatives to conventional ANCOVA/APV are discussed as are steps that can be taken to minimize the impact of this bias on drawing valid inferences when using conventional ANCOVA/APV.

Analysis of Covariance (ANCOVA) or its multiple regression (MR) counterpart, Analysis of Partial Variance (APV, Cohen & Cohen, 1983), is commonly used in psychopathology research. The most unambiguous case when conventional ANCOVA/APV has served a legitimate and useful purpose is the one for which conventional ANCOVA was developed in which the dependent variable (*DV*) is correlated with the covariate (*Cov*) but the main independent variable (*IV*) of interest is not. In this case, conventional ANCOVA/APV reduces error in the *DV* and thereby increases statistical power for testing the relationship between the *IV* and the *DV*. Unfortunately, there are many cases in psychopathology research in which conventional ANCOVA/APV has been used in a more controversial fashion.

In these more controversial cases, the researcher uses ANCOVA/APV to test whether a relationship between the *IV* and the *DV* is actually due to a confounder variable. Concern about possible misuses of conventional ANCOVA in these cases has stimulated numerous articles, chapters and books (e.g., Cochran, 1957; Elashoff, 1969; Fleiss & Tanur, 1973; Huitema, 1980; Lord, 1960, ¹⁹⁶⁷, ¹⁹⁶⁹; Maxwell, Delaney & Manheimer, 1985; Porter & Raudenbush, 1987; Reichardt, 1979; Wainer, 1991; Wildt & Ahtola, 1978). Whereas random assignment to conditions often eliminates confounds, thereby obviating the need for these more controversial

Correspondence concerning this article should be addressed to Richard Zinbarg, Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208-2710: rzinbarg@northwestern.edu.

The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/abn.

uses of ANCOVA/APV, random assignment to the different levels of a psychopathology variable represented in a given study “is routinely unfeasible and/or unethical” (Miller & Chapman, 2001, p. 40). Thus, the *Cov* is correlated with the *IV* in a typical psychopathology study. Psychopathology researchers, therefore, need to have a thorough understanding of the possible misuses of conventional ANCOVA/APV and how to avoid or minimize them.

An accessible treatment of some misuses of ANCOVA/APV was provided by Miller and Chapman (2001), who articulate the problems that can arise in conventional ANCOVA when adjusting for the *Cov* may remove part of the effect of the *IV*. They assume a simple design having one *IV*/grouping variable (*Grp*), one *DV* and one *Cov* and they frame their discussion using a MR approach. When *Cov* is entered into the regression, this removes the variance *Cov* shares with *Grp*, leaving a residual portion of *Grp*, *Grp_{res}*, that is not correlated with *Cov*. A problem arises, however, when the variance shared with *Cov* is an important facet of *Grp*, and *Grp_{res}* is used to answer the question of whether the groups would differ on the *DV* if they did not differ on *Cov*. As stated by Miller and Chapman, the problem here is that “*Grp_{res}* is not a good measure of the construct that *Grp* is intended to measure ... (p. 43)”.

Unfortunately, there is another problem in the use of conventional ANCOVA/APV that can arise even when applied to less controversial, consensually meaningful questions. For statistical reasons reviewed below, ANCOVA/APV often generates biased results. Though discussed by several methodologists, even some of the best designed and most conceptually significant, recent psychopathology studies provide little indication that psychopathology researchers are aware of this bias. Indeed, two of us (A.A.U. and A.R.L.) independently coded all 61 articles in three recent issues of this journal (Volume 116, Issue 4 Volume 117, Issue 4, and Volume 118, Issue 1) and agreed that 12 (19.7%) of these articles involved the use of ANCOVA/APV that were likely to be vulnerable to this bias ($\kappa = .60$). Of these 12 articles, we agreed that in 10 (83.3%) of them the researchers provided no indication of awareness of this bias ($\kappa = .63$).

In light of the prevalence of ANCOVA/APV in psychopathology research without indication that researchers are aware of the bias in ANCOVA/APV, this article has three aims. The first is to clarify the nature of the questions a psychopathologist might try to answer with the use of ANCOVA/APV that are more consensually meaningful than the type of question criticized by Miller and Chapman (2001). The second is to raise awareness of the bias in ANCOVA/APV when used to address such consensually meaningful questions. The final aim is to describe alternatives to conventional ANCOVA/APV and discuss how one can strengthen the validity of inferences when using conventional ANCOVA/APV.

Consensually Meaningful Questions an Investigator Might Be Trying to Address When Using ANCOVA/APV For Cases in Which the Groups Differ on the Covariate

Miller and Chapman (2001) use the example of comparing depressed patients with non-patient controls using anxiety as a *Cov* to illustrate their central contention that the questions that some investigators try to address with conventional ANCOVA/APV are not meaningful. Anxiety is higher in patients than controls and Miller and Chapman note that if “we believe that the negative affect that depression and anxiety share is central to the concept of depression, then removing negative affect (by removing anxiety) will mean that the group variance that remains has very poor construct validity for depression (p. 43).” They contend that it is simply not a meaningful question to ask whether depression would relate to another variable if depression did not include a facet widely thought to lie at its core.

Unfortunately, Miller and Chapman (2001) could be read to imply that the only purpose an investigator might have in using anxiety as a *Cov* and depression as the *IV* is understanding the effects of “pure depression”. For example, when discussing the possibility of using a self-report anxiety measure as a *Cov* in an ANCOVA with diagnostic group as the *IV* and the brain-wave measure known as P300 as the *DV*, they state “the hope in such an analysis would be to control anxiety and thus be able to observe the relationship between pure depression (not confounded with anxiety) and P300.” Because Miller and Chapman did not consider any other possible motivations for this analysis, researchers with other motives for such an analysis might be led to wonder whether their questions are consensually meaningful.

There are indeed questions that arise frequently in psychopathology research other than the form criticized by Miller and Chapman that are consensually meaningful and to which ANCOVA/APV is frequently applied. What these questions share in common is that their inferential focus is not on the *Grp* latent variable (*LV*) that an important facet has been removed from it. Rather, either the *Cov* is not thought to measure an important facet of the *Grp* *LV* in the first place or the inferential focus is explicitly on either the *Grp_{res}* *LV* or is not on any *LV* but rather is on the *Grp* observed measure.

Miller and Chapman (2001) illustrate the first class of consensually meaningful questions when noting that if the comorbidity between anxiety and depression were thought to arise because of variance due to factors not central to depression, then ANCOVA might be effective in removing this variance, leaving *Grp_{res}* interpretable as “pure” depression. For example, imagine that anxiety and depression comorbidity arises solely from depression in some people triggering anxiety focused on the worry that their depression will never remit or will recur. In this case, the depression experienced by individuals who do not also experience an elevation in their anxiety is a valid representation of depression.

A second consensually meaningful class of questions that a researcher might try to address by ANCOVA/APV involve asking whether a specific *component* of a *Grp* *LV* exists that uniquely relates to the *DV* above and beyond the *Cov*. Thus, even if one believes that the negative affect that depression and anxiety share is central to the concept of depression, unless one believes that depression and anxiety are the same *LV*, then one (perhaps implicitly) believes that there is reliable, unique variance in depression and/or anxiety that provides the basis or bases for differentiating them. According to Clark and Watson (1991), for example, the variance in anxiety and depression can be decomposed into: (a) negative affect, which is common to anxiety and depression, (b) physiological hyperarousal, which is specific to anxiety and (c) anhedonia which, is specific to depression. Thus, rather than using anxiety as a *Cov* to observe the relationship between “pure depression” and P300, an investigator might be hoping to observe the relationship between anhedonia and P300. It is important to note that this consensually meaningful question does not take *Grp_{res}* to be a good observed indicator of the *LV* that *Grp* is intended to measure. Rather, this question explicitly recognizes that *Grp_{res}* is just one component of the *LV* that *Grp* is intended to measure. Of course, if the investigator’s hope is to observe the association between anhedonia and P300, a superior design would involve measuring negative affect and anhedonia more directly.

A third class of consensually meaningful research questions frequently addressed via conventional ANCOVA/APV involve differential change in a construct occurring across two time points. For example, there is a great deal of scientific interest in whether some *IVs* (such as anxiety sensitivity) predict increases in various outcome variables (such as fear) in response to a stressor. Such questions are often tested by measuring the *IV* and the outcome prior to the stressor and then re-administering the outcome measure after the stressor. The conventional analysis of the resulting data would be an ANCOVA/APV in which the outcome measured after the stressor is treated as the *DV* and the outcome measured before the stressor is treated

as the *Cov*. Here, the question is primarily focused on the residual portion of *DV*, DV_{res} , that is not correlated with *Cov* (that part of the outcome measure that is independent of its baseline level). When the *IV* and the pre-stressor outcome measure serving as the *Cov* are correlated, however, these analyses will be vulnerable to the bias discussed below.

A final class of consensually meaningful questions that could be addressed via conventional ANCOVA/APV is whether the observed *IV* measure (e.g., a measure of cognitive vulnerability to depression) has unique effects above and beyond the effects of the observed *Cov* measure (e.g., a measure of neuroticism). Such a question might be relevant, for example, in deciding whether to include the *IV* measure in addition to the *Cov* measure in a battery designed to identify individuals for treatment or a preventive intervention. For this class of question, the inferential focus is entirely on the observed measures rather than on the *LVs* that the measures might be purported to be indicators of. (Of course, this class and the second class of questions represent classic applications of hierarchical regression.) As will be discussed in more detail below, ANCOVA/APV generates unbiased answers to this class of question.

Bias in ANCOVA/APV Due to Unreliability in the Service of Addressing Consensually Meaningful Questions When the Independent Variable and the Covariate are Correlated

Given that there are consensually meaningful questions that conventional ANCOVA/APV might be used to answer when the *IV* and the *Cov* are correlated (or, equivalently, when groups that serve as the levels of a categorical *IV* differ on the *Cov*), it becomes important to ask whether conventional ANCOVA/APV provides unbiased answers to such research questions. The answer to this question depends on whether (a) the inferential focus is on the observed indicators versus the *LVs* measured by those indicators, (b) the observed indicators of the *LVs* contain measurement error and (c) there are any unmeasured confounders. In particular, when drawing inferences about *LVs*, if the *Cov* is measured with error and/or there is an unmeasured confounder then the answer is almost certainly no (e.g., Carroll, Ruppert, Stefanski & Crainiceanu, 2006; Huitema, 1980; Kahneman, 1965; Kenny, 1979; Lord, 1960; Maxwell & Delaney, 2004; Sorbom, 1979; Vargha, Rudas, Delaney & Maxwell, 1996). When the *IV* *LV* does not have a unique effect above and beyond the effect of a correlated but less than perfectly reliable *Cov* *LV*, ANCOVA/APV is systematically biased toward underadjusting for the effects of the *Cov* *LV*. Since an unmeasured variable is equivalent to a variable that is measured with zero reliability (Judd & Kenny, 1981), the most extreme version of such underadjustment occurs when a relevant *Cov* *LV* is not included in the analysis (e.g., Kenny, 1979). Though Miller and Chapman (2001) addressed the issue of underadjustment due to unreliability in passing (e.g., p. 42), they were concerned primarily with contexts in which conventional ANCOVA/APV removes too much of the variance in the *IV*. In contrast, this article is concerned primarily with contexts in which ANCOVA/APV does not remove enough of the variance in the *IV* (i.e., it does not remove enough of the shared variance with the *Cov* *LV*).

Figure 1 shows a structural equation model (SEM) representation of this article's main running example. Paths *a*, *b* and *f* represent the standardized loadings of the *Cov* observed indicator (i.e., anxiety), the *IV* observed indicator (i.e., depression), and the *DV* observed indicator on their respective *LVs* and paths *a'*, *b'* and *f'* represent the standardized loadings of alternative, congeneric indicators that could be used to measure the *LVs* (note that, for model identification, three indicators are needed for any *LV* that is uncorrelated with the other *LVs* in a model). Thus, the reliabilities of the three observed indicators are a^2 , b^2 and f^2 and their standardized measurement errors are $1-a^2$, $1-b^2$ and $1-f^2$. Assuming that the model in Figure 1 is valid, including that all errors (not shown in Figure 1) are independent, there are two pathways

originating at the observed measure of the *IV* and ending at the observed measure of the *DV LV*. The first is the unique pathway that begins with the observed measure of the *IV LV* and runs through to the observed measure of the *DV LV* (path *bef* in Figure 1). The second is the pathway resulting from the *IV LV* being correlated with the *Cov LV* which is another cause of the *DV LV* (path *bcd* in Figure 1). The zero-order correlation between the observed measures of the *IV* and the *DV* ($r_{DV,IV}$) is the sum of these two pathways (*bef* + *bcd* or *bf* [*e*+*cd*]). That this estimate of the zero-order correlation between the observed measures of the *IV* and the *DV* is negatively biased is well-known and widely appreciated; the unreliability in the observed indicators attenuate the zero-order correlation that exists between the *IV* and the *DV LVs* (*e* + *cd*). However, unreliability in a *Cov* measure leads to a very different bias that appears to be well known by methodologists but not widely appreciated by psychopathology researchers. When ANCOVA/APV is used to estimate the unique association of the *IV* and *DV LVs* in cases where there is in fact no unique association, unreliability in a *Cov* measure leads to a positive bias and inflated type I error rates (e.g., Bollen, 1989; Kahneman, 1965).

The separate components of the *IV-DV* relationship reflecting the two pathways, (1) the *IV LV* being correlated with the *Cov LV*, which is another cause of the *DV LV*, (path *cd* in Figure 1) and (2) the unique association of the *IV* and *DV LVs* (path *e* in Figure 1), cannot be estimated without at least one observed measure of the *Cov LV*. Conventional ANCOVA/APV uses a single measure of the *Cov LV* and a single measure of the *IV* and the *DV LVs* to do this. The ANCOVA/APV estimate of the unique association of the *IV LV* with the *DV LV* (path *e* in Figure 1) expressed in terms of a standardized partial regression coefficient is

$$\beta_{DV,IV.Cov} = \frac{r_{DV,IV} - r_{DV,Cov}r_{Cov,IV}}{1 - r_{Cov,IV}^2}.$$

In terms of the SEM representation depicted in Figure 1, it is

$$\begin{aligned} \beta_{DV,IV.Cov} &= \frac{(bef + bcd) - (adf + acef)(acb)}{1 - (acb)^2} \\ &= bf \left[e \left(\frac{1 - a^2 c^2}{1 - (acb)^2} \right) + cd \left(\frac{1 - a^2}{1 - (acb)^2} \right) \right]. \end{aligned} \quad (1)$$

Equation 1 is either identical to (e.g., Bollen, 1989) or a more general expression of (e.g. Kenny, 1979) equations presented previously and shows that the conventional ANCOVA/APV estimate of the unique association of the *IV* is quite complicated and subject to multiple biasing influences. The multiple biasing influences may result in underestimation or overestimation of the unique association of the *IV* or, very infrequently, may even completely offset each other to yield an unbiased estimate (Reichardt, 1979).

Fortunately, Equation 1 simplifies considerably in the special case corresponding to the use of ANCOVA/APV common in psychopathology research and of primary concern here. This use consists of testing the hypothesis that $r_{DV,IV}$ is an artifact of an association between the *IV LV* and the *Cov LV*, with only the *Cov LV* having a unique association with the *DV LV*. That is, this use tests the null hypothesis that the *IV LV* has no unique association with the *DV LV* after accounting for the *Cov LV* ($H_0: e = 0$). In this case (when $e = 0$), equation 1 simplifies to

$$\beta_{DV,IV.Cov} = bcd \left(\frac{1 - a^2}{1 - (acb)^2} \right). \quad (2)$$

Note that the hypothesis that $r_{DV, IV}$ is an artifact of an association between the *IV* and *Cov* LVs would be entertained only when the product of c and d has the same sign as $r_{DV, IV}$ (e.g., if depression correlates positively with the *DV* and with anxiety, but anxiety correlates negatively with the *DV*, then anxiety cannot possibly account for the positive correlation between depression and the *DV*). Assuming that cd is positive, Equation 2 shows that when $e = 0$, $\beta_{DV, IV, Cov} \geq 0$ with equality holding only when a equals 1, or b , f , c , or d equals 0 (and if cd is negative then $\beta_{DV, IV, Cov} \leq 0$ with equality again holding only when a equals 1, or b , f , c or d equals 0). With a few notable exceptions (e.g., when the *Cov* is sex or age), it is unrealistic in a typical psychopathology study to expect the *Cov* to be perfectly reliable ($a = 1$) or either the *IV* ($b = 0$) or the *DV* ($f = 0$) to be perfectly unreliable. Moreover, one needs to test whether the zero-order correlation between the *IV* and *DV* is an artifact of an association between the *IV* and *Cov* LVs only when the *IV* and the *Cov* are correlated; c will therefore not equal 0 when ANCOVA/APV is used in this manner in psychopathology research. Thus, ANCOVA/APV estimates of e and inferential tests of $H_0 : e = 0$ in this case will typically be associated with positively biased type I error rates with the type I error rate bias being an increasing function of the (1) correlation between the *Cov* LV and the *IV* LV (c), (2) unique association between the *Cov* LV and the *DV* LV (d), (3) unreliability of the *Cov* measure ($1 - a^2$), (4) reliability of the *IV* and *DV* measures (b and f) and (5) sample size.

Examples illustrating the size of the underadjustment bias and associated type I error rates given different values of the parameters governing this bias are given in Table 1. The effect size of the bias is small to very small ($f_{effect}^2 \leq .02$, Cohen's $d \leq .20$) in most of the examples. Even with small effect sizes, however, the inflation in type I error rates are large enough to be concerned about in all but one example. This is especially true at the larger sample sizes. Consider example 2 in Table 1, with reliabilities of .81 for the *Cov* ($a^2 = .900^2$) measure and .90 for the *IV* and *DV* measures ($b^2 = f^2 = .949^2$), a correlation of .450 between the *IV* and *Cov* measures ($abc = .949 \times .949 \times .50$), and a unique association of .500 between the *Cov* and the *DV* LVs. In this case, the small bias inherent in $\beta_{DV, IV, Cov}$ as an estimate of e causes type I error rates to be more than double the nominal rate of .05 with a sample size of 200, and more than triple the nominal level with a sample size of 400. Or consider example 3 in Table 1, in which the *Cov* measure has a reliability of .64 (i.e., $a^2 = .800^2$) and is otherwise identical to example 2. Here, with a sample size of 400, there is almost a 50% chance that ANCOVA/APV will lead to the conclusion that there is a unique association between the *IV* and *DV* LVs when there is no such unique association.

Applied contexts in which inferences are focused on the observed measures, such as personnel selection or selection into an intervention program, may be conceptualized as cases in which the *IV*, *Cov* and *DV* measures are perfectly reliable ($a = b = f = 1$). Thus, in such contexts, when $e = 0$, Equation 2 reduces to $\beta_{DV, IV, Cov} = 0$ and (and even when $e \neq 0$, Equation 1 shows that $\beta_{DV, IV, Cov} = e$). That is, ANCOVA/APV is unbiased when inferences are focused on the observed measures (Huitema, 1980).

This discussion reiterates a message that has been all too often ignored by psychopathology researchers. ANCOVA/APV fully adjusts for the *Cov* measure and thus is unbiased when inferences are focused on the observed measures. However, *when there is an association between the IV and the Cov LVs, and inferences are focused on the associations among the LVs, there will usually be underadjustment for the Cov LV and positive bias in the type I error rate of the ANCOVA/APV test of $H_0 : e = 0$.*

Omitted Covariate Bias as an Extreme Form of Underadjustment Due to Unreliability

Omitted variable bias (*OV*) is well known among methodologists (e.g., Kenny, 1979). Figure 2 shows a problematic omitted variable (*OV*) added to the simple model considered above. A problematic *OV* is one that correlates with the *IV* and is a cause of the *DV* (e.g., Judd & Kenny, 1981). In the running depression and anxiety example, a possible example of a problematic *OV* is life stress.

The higher the correlation between the *OV* and *Cov*, the more the *Cov* measure also adjusts for the *OV*. Thus, no correlation between the *Cov* and the *OV* results in the most problematic case. For simplicity, therefore, Figure 2 depicts an *OV* that is uncorrelated with the *Cov* *LV*. Thus, Figure 2 contains two new paths (as the *OV* is, by definition, omitted, there are no observed indicators of it): the correlation between the *IV* and the *OV* *LVs* (*g*) and the unique association between the *OV* and *DV* *LVs* (*h*). Example 6 in Table 1 added an *OV* to the model underlying Example 5. Whereas Example 5 involved a trivial underadjustment bias, the bias is considerable in Example 6. *OV* may be seen as equivalent to the case in which a confounding *Cov* *LV* exists but is measured with a reliability of zero (for a closely related discussion, see Judd & Kenny, 1981, p. 192). That *OV* can be conceptualized as an extreme form of the underadjustment bias due to unreliability is illustrated in Example 7 in which the *Cov* *LV* has identical correlations with the *IV* and *DV* *LVs* as does the *OV* in Example 6. With the reliability of the Example 7 *Cov* indicator equaling only .01 ($a^2 = .1^2$) bias is almost as great as in Example 6.

Potential Alternatives to Conventional ANCOVA/APV

As randomized studies generally support stronger causal inferences than do non-randomized studies (e.g., Bollen, 1989; Shadish, Cook & Campbell, 2002), analogue studies using randomized designs can make important contributions to the study of psychopathology. However, analogue studies can never entirely supplant studies of participants with clinical diagnoses or symptoms, given their limitations in terms of external validity (Sher & Trull, 1996). Therefore, it is important to consider how non-randomized psychopathology studies can minimize underadjustment bias due to unreliability.

Structural equation model analysis or Aggregated Measures analysis

Though researchers will never be able to entirely eliminate the effects of measurement error in their analyses, they can minimize its impact via the judicious incorporation of multiple observed indicators of their *IV*, *DV*, and (especially) *Cov* *LVs*. The resulting data could then be analyzed in one of two ways. The first option would be to attempt to explicitly model measurement error using SEM analyses (e.g., Huitema, 1980; Maxwell & Delaney, 2004; for an excellent example, see Aiken, Stein & Bentler, 1994). The second option would be to aggregate the multiple measures of each construct into composites and use ANCOVA/APV. The latter option might be called an Aggregated Measures ANCOVA/APV to distinguish it from a conventional ANCOVA/APV in which there is a single measure of each *LV*. The potential advantage of this approach is that if the multiple measures are properly selected then error variance would tend to be smaller in a composite measure of the *Cov* compared with a single indicator measure of the *Cov* *LV* (e.g., Cronbach, 1951). In terms of comparing SEM versus an Aggregated Measures ANCOVA/APV, SEM would require larger samples but, when sample size is adequate, would have the advantage of allowing a formal assessment of the goodness of fit of one's measurement model.

Of course, incorporating multiple measures of the *LVs* will not automatically ameliorate bias (e.g., DeShon, 1998). Imagine that the multiple indicators of the *IV LV* share method variance and this shared method variance is also associated with the *DV* measure(s). In this case, the shared method variance would constitute an *OV*. Though random error in the measurement of the *Cov LV* will be reduced, *OV*B will certainly not be reduced and may even be exacerbated (given that the *OV* likely accounts for a larger proportion of the variance common to the set of *IV* indicators than in any single member of that set).

Theory should play a central role in guiding the selection of measures whenever possible (Little, Lindenberger & Nesselroade, 1999). Indeed, blind reliance on selecting those measures that are most highly correlated can increase bias under some conditions. For example, if choosing between two self-report measures of depression or one self-report and one other report measure, it is likely that the highest correlation will be the one between the two self-report measures. Thus, reliability would be maximized by using the two self-report measures. However, use of the two self-report measures would also be more likely to create what Cattell (1978) and Little et al. (1999) would call a “bloated specific” factor and possibly exacerbate *OV*B due to shared method variance. That is, if the shared method variance is also shared with the *DV* measure then that method variance would constitute an *OV*. Selecting one self-report measure and one other-report measure is likely to produce a smaller increase in the reliability of measurement of depression but might reduce the potential for *OV*B due to shared method variance. Little et al. (1999) provide a discussion of four key dimensions of indicator selection that many psychopathology researchers should find helpful. Of course, when measures are clustered (e.g., several measures of each of several methods are included), it is also important to follow DeShon’s (1998) recommendation to take this clustering into account (DeShon’s recommendation could also be generalized to an Aggregated Measures ANCOVA/APV; instead of forming a single *Cov* composite the researcher would form several *Cov* composites with one per method/cluster).

Gain score analysis

When a longitudinal design includes two time points and the research question concerns differential change over that interval, the conventional analysis is an ANCOVA/APV analysis of “regressed change”. That is, the time 2 measure of the outcome variable is entered as the *DV* and the time 1 measure of the outcome variable is entered as the *Cov* to “control” for the association between the *IV* and the time 1 outcome measure (or for group differences at time 1). When the time 1 outcome *LV* is correlated with the *IV LV*, these analyses will generate biased estimates of e when this path in fact equals 0 with a corresponding inflation in the type I error rate of the test of $H_0 : e = 0$.

Whereas gain scores have been much criticized, some methodologists have refuted these criticisms and/or articulated the advantages of gain scores (e.g., Allison, 2005; Rogosa, 1995; Rogosa, Brandt, & Zimowski, 1982; Willett, 1988; Williams & Zimmerman, 1996). Maxwell and Delaney (2004) conclude that an ANCOVA is often preferable to an analysis of gain scores for randomized designs; they also conclude (p. 448) that “in intact group studies, then the ANOVA of gain scores is to be preferred.” Thus, ANCOVA/APV will often be preferable in randomized psychotherapy studies. However, gain scores should be seriously considered in longitudinal, two wave studies of psychopathology as subtracting the time 1 outcome measure from the time 2 outcome measure rather than using the time 1 outcome measure as a *Cov* produces an unbiased estimate of true change (Willett, 1988). Of course, gain scores are only interpretable when the measures of the outcome variable demonstrate factorial temporal invariance (e.g., Horn & McArdle, 1992). Raw gain score analysis further assumes that the variance of the outcome measures also demonstrate temporal invariance and

when this assumption is violated standardized gain score analysis should be used (e.g., Judd & Kenny, 1981).

Among those who argue that gain scores are more appropriate than an analysis of “residualized change” for designs with two measurement waves, many recognize that such designs are limited in the first place (e.g., Rogosa, 1995; Willett, 1988). When possible, three or more waves of data should be collected when studying change, and autoregressive structural equation models, hierarchical linear modeling, conventional growth curve analysis, latent growth curve analysis or survival analysis should be used (e.g., Bollen & Curran, 2004; Hertzog & Nesselroade, 2003; Singer & Willett, 2003).

Propensity score analysis

Propensity score analysis (PSA) was developed by Rosenbaum and Rubin (1983) for analyzing data from quasi-experimental research with many confounder *Cov* *LVs* so as to “control for naturally occurring systematic differences in background characteristics between the treatment group and the control group” (Rubin, 1997, p. 757). There are two steps to PSA. First, all available *Covs* are used to predict group membership in a logistic regression. Plugging a participant’s values on the *Covs* into the logistic regression equation yields their expected probability of being in the treatment (psychopathology) group rather than the control group. This expected probability is the person’s propensity score. In the second step, participants across the two groups are then matched or stratified on the basis of their propensity scores. The propensity score could also be used as a *Cov* in ANCOVA. Though PSA may have some advantages over conventional ANCOVA (e.g., Rubin, 1997; Shadish et al., 2002), reliability has been a largely neglected topic in the PSA literature (Glynn, Schneeweiss & Sturmer, 2006). When the multiple *Covs* involved in a PSA are correlated, it seems likely that PSA will minimize bias due to unreliability (analogous to an aggregated measures ANCOVA). However, the logic of PSA does not call for the *Covs* to be correlated. Thus, it is not clear PSA is less vulnerable than conventional ANCOVA/APV in general to underadjustment bias due to unreliability.

Minimizing the Impact of Bias Due to Unreliability on Drawing Valid Inferences When Using Conventional ANCOVA/APV and the Independent Variable Correlates with the Covariate

When it is impractical to include multiple measures of the *LVs* under study, there is often little choice but to conduct a conventional ANCOVA/APV. At least eight recommendations can be offered to minimize the impact of bias due to unreliability on the validity of the inferences drawn from such results.

The first recommendation is that one could estimate each of the parameters in Equation 2 to evaluate by how much one’s estimate of the unique association between the IV and the DV *LVs* and type I error rate are inflated given the sample size and parameter estimates. That is, based on the estimated value of the regression coefficient and assuming normal distributions, one could compute an effect size estimate and (using power tables or calculators) estimate the probability of rejecting the null hypothesis given that effect size and the sample size in a given study. In the case of non-normal distributions, one could conduct a Monte Carlo study to estimate the type I error rate. One could then use this information to adjust the regression coefficient and the nominal type I error rate of the test of $H_0: \beta = 0$ such that the actual type I error rate equaled the desired level. For instance, in example 2 in Table 1, if the nominal type I error rate in a study with a sample size of 400 were set to .00719, the actual type I error rate would be .05 (rather than the actual type I error rate of .181 if one used a nominal type I error rate of .05). Of course, the extent to which this approach would succeed depends on the accuracy

of the estimates of the parameters in Equation 2. It is likely that the reliability of the *IV*, *Cov* and *DV* measures can be readily estimated in one's sample (and, when that is not the case, may be available from other studies), and one could then use these reliability estimates to estimate the correlation between the *IV* and *Cov* *LVs*. However, empirical estimates of the unique association between the *Cov* and *DV* *LVs* may not be readily available. Relatedly, one could use the reliability estimates of the *IV*, *Cov* and *DV* measures to fix their measurement errors in a SEM model with single indicators and simultaneously estimate each of the parameters in the model including the unique association between the *Cov* and *DV* *LVs* (e.g., Hayduk, 1987; McDonald, Behson & Seifart, 2005; also see the somewhat related approach of simulated extrapolation, Carroll, Ruppert, Stefanski & Crainiceanu, 2006).

A drawback to the use of the reliability estimates in either Equation 2 or in a single indicator SEM is that these approaches are likely to be very sensitive to the reliability estimates and these estimates are known to under-estimate reliability in some conditions and over-estimate reliability in others (e.g., Zinbarg, Revelle, Yovel and Li, 2005). It is also well known that reliability estimates are sample specific. Therefore, reliability estimates obtained from other studies may also fail to accurately represent the reliabilities of the measures in the sample in hand. If the *IV*, *Cov* and *DV* measures are composite scores derived from multiple items, these issues could be addressed by conducting item-level SEMs with careful measurement modeling. As item-level SEM can be problematic when items have few response options and non-linear relationships with their factors (e.g., Bernstein & Teng, 1989; Little, Cunningham, Shahar & Widaman, 2002; Waller, Tellegen, McDonald & Lykken, 1996), such analyses will often benefit from using SEM approaches for categorical data (e.g., Muthén, 1984; also see Bauer & Curran, 2004) or from grouping items into parcels (e.g., Little et al., 2002). Given our earlier discussion of indicator selection in SEM and that the most commonly used reliability estimates are often inflated by correlated residual variance (e.g., Judd & Kenny, 1981), a limitation common to using reliability estimates in Equation 2, single indicator SEM, item-level SEM and parcel-level SEM is that each of these approaches will typically be vulnerable to bias arising from correlated residuals. Thus, thoughtful design involving multiple indicators carefully chosen to be heterogeneous with respect to residual variance should often lead to greater bias reduction than will the choice of data analytic approach.

A second recommendation is to conduct a sensitivity analysis to assess the extent to which biases of various sizes would change the results of the study when empirical estimates of the unique association between the *Cov* and *DV* *LVs* are not available (e.g., Marcus, 1997; Rosenbaum, 2002; Rosenbaum & Rubin, 1983). That is, one can determine by how much the nominal type I error rate would need to be adjusted to achieve an actual type I error rate of .05 for each of a plausible range of values of the unique association between the *Cov* and *DV* *LVs* (and/or for each of a plausible range of values of the reliabilities as suggested by Judd & Kenny, 1981, p. 114). The observed result might remain significant at the adjusted levels for all but the most extreme estimates of the unique association between the *Cov* and *DV* *LVs*. This would indicate that a conclusion that the *IV* has a unique association with the *DV* would not be biased by underadjustment for the *Cov* unless the unique association between the *Cov* and *DV* *LVs* is very large. Alternatively, the observed result might only remain significant at the adjusted levels associated with small estimates of the unique association between the *Cov* and *DV* *LVs*. This pattern would indicate that a conclusion that the *IV* is uniquely related to the *DV* would be warranted at conventional type I error rates only if the unique association between the *Cov* and *DV* *LVs* is small.

The pattern of adjusted significance levels and the size of the unique association between the *Cov* and *DV* *LVs* required to produce them will vary over studies. This is illustrated in Table 2 and Table 3 that provide examples of sensitivity analyses of the results from two hypothetical studies. In both studies, the reliabilities of the *IV* and the *DV* measures equal .90 and the

correlation between the *IV* and *Cov* *LVs* equals .50 (thus the observed correlation between the *IV* and the *Cov* measures equals .45?). In the hypothetical study presented in Table 2, the reliability of the *Cov* measure equals .88 whereas it equals .72 in the hypothetical study presented in Table 3. In addition, the sample size in the hypothetical study presented in Table 2 is 140 whereas it equals 300 in the one presented in Table 3. Clearly the study presented in Table 3 is much more sensitive to underadjustment bias than the one presented in Table 2.

To make these examples even more concrete, imagine that the test of the regression coefficient of the unique association between the *IV* and the *DV* using a nominal type I error rate is associated with a *p* value of .020 in both studies. From the results in Table 2 we would infer that the test of the unique association between the *IV* and the *DV* in that study would remain significant unless the unique association between the *Cov* and the *DV* *LVs* is larger than .7. This does not entirely rule out the presence of underadjustment bias in this study as an explanation for the significant result at the nominal type I error rate of .05; bias would be present if the unique association between the *Cov* and the *DV* was greater than .7. If such a large unique association would be implausible however, it would make underadjustment bias implausible as an explanation for the significant result obtained at the unadjusted type I error rate of .05. In contrast, from the results in Table 3 we would infer that the test of the unique association between the *IV* and the *DV* in the study presented in that table would remain significant only if the unique association between the *Cov* and the *DV* *LVs* were smaller than .3. Unless one could compellingly argue that it is plausible to assume that the unique association between the *Cov* and the *DV* *LVs* is smaller than .3, underadjustment bias would remain a plausible alternative explanation for the result that was significant at the unadjusted type I error rate of .05.

A third recommendation is specific to ANCOVA and is to always closely examine the group means and standard deviations (*sds*) on the *Cov* and the *DV*. In the classic example presented by Lord (1967) in which ANCOVA indicates a sex difference in residualized post-test weight when the sexes did not differ in average weight gain from pre-test to post-test, an examination of the group means and *sds* would have made very clear to the analyst who chose to analyze the data via ANCOVA that the sex difference in weight was no greater at post-test than at pre-test and that the mean weight at post-test was the same as the mean weight at pre-test for both sexes. It would have therefore been clear that there could not have been a sex difference in average weight change over time. Thus, close examination of the group means and *sds* would have suggested that the ANCOVA result was misleading.

The fourth recommendation is to use great care when selecting observed indicators of their *LVs* (Little et al., 1999). This point underscores the importance of careful psychometric assessment of our measures. Errors in the *IV* and *DV* measures will attenuate the estimate of the zero-order relation between their *LVs*. Such attenuation obviously reduces statistical power which is typically rather poor in most psychological research (Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). Even when we have sufficient power to detect a zero-order association between the *IV* and the *DV*, however, error in the *Cov* measure will positively bias tests of $H_0: \rho = 0$. Thus, when using conventional ANCOVA/APV to test $H_0: \rho = 0$, it is especially important to select a highly reliable *Cov* measure that does not share method variance with the *DV* (see example 5 in Table 1).

A fifth recommendation is to not dichotomize a continuous *Cov*. Dichotomization produces underadjustment bias (Vargha et al., 1996), a result consistent with the effects of unreliability focused on here because dichotomization is a source of measurement error.

A sixth recommendation concerns a design feature to strengthen the validity of inferences based on a conventional ANCOVA/APV, the nonequivalent dependent variable (Shadish et al.,

2002). A great deal of inferential leverage can be gained by incorporating a second *DV* (*DV2*) in a study that is expected to show a unique association with *Cov* but not with the *IV*, as depicted in Figure 3. If the expected pattern of results is obtained in which the *IV* shows a reliable unique association with the first *DV* but not *DV2* and the zero-order correlation of the *Cov* with the two *DVs* are comparable in magnitude, then one can be more confident that the unique association of the *IV* with the first *DV* is not merely the result of underadjustment due to unreliability of the *Cov*. That is, the *Cov* is reliable enough in this population to account for a relationship for which one would expect the estimate of the unique association to be at least as biased as one would expect the estimate of the unique association of the *IV* and the first *DV* to be. Thus, it could be concluded with more confidence that the *IV* does have a unique association with the first *DV* above and beyond the effects of the *Cov* than would be the case in a study that did not include *DV2*.

To make this inference, it is crucial that the zero-order correlation between the *Cov* and *DV2* is at least as great as that between the *Cov* and the first *DV*, so that the estimated unique association between the *IV* *LV* and the *DV2* *LV* is equivalently or more vulnerable to underadjustment bias than that between the *IV* *LV* and the first *DV* *LV*. That is, assuming that there is no unique association between the *IV* *LV* and the *DV2* *LV*, the bias in the estimate of

this association would equal $bhcg\left(\frac{1-a^2}{1-(acb)^2}\right)$. Dividing this quantity by Equation 2 to compare the size of the bias in the two estimates assuming that both unique associations equal zero

yields $\frac{hg}{fd}$. Further, if *DV2* correlates at least as highly with the *Cov* as does the first *DV* then $hg \geq fd$. That is, if *DV2* correlates at least as highly with the *Cov* as does the first *DV*, then the estimate of *DV2*'s unique association with the *IV* would be associated with an even more positively biased type I error rate. Thus, the lack of significant unique association with the *DV2* can't be attributed to the test of this association having a smaller inflation in its type I error rate. Of course, the conclusion would be strengthened further by showing that the *IV*'s unique association is significantly stronger with the first *DV* than with *DV2* such that the conclusion is not dependent on accepting a null hypothesis. That is, it would have then been demonstrated that the unique association of the *IV* with the first *DV* is significantly larger than an association with at least as much underadjustment bias, allowing one to rule out the possibility that underadjustment bias entirely accounts for the unique association of the *IV* with the first *DV*. A strength of the *DV2* approach is that it can be used when *a*, *b* and *f* cannot be estimated with confidence (such as when using single-item measures).

A seventh recommendation stems from the recognition that though the ANCOVA/APV estimate will be positively biased, the ANCOVA/APV estimate will be less biased than the zero-order correlation as an estimate of the unique association of the *IV* latent variable with the *DV* latent variable when in fact no such unique association exists. In the case in which there is no unique association of the *IV* with the *DV* ($e = 0$), the zero-order correlation equals $bfdc$ and Equation 2 shows that in this case $bfdc \geq \beta_{DV,IV.Cov}$ with equality holding only in the unrealistic case in which the *Cov* is entirely unreliable ($a = 0$). Thus, we recommend that psychopathology researchers remind reviewers and readers that, even though it doesn't eliminate bias, ANCOVA/APV does reduce bias when one's question is whether the *IV*-*DV* association could be entirely due to a potential confounder.

To illustrate this point, consider Example 3 in Table 1 in which the ANCOVA/APV estimate of the unique association of the *IV* with the *DV* is positively biased with a type I error rate of nearly 50% when sample size equals 400. The zero-order correlation between the *IV* and *DV* measures (.225) in this example would be more than twice as positively biased (Cohen's $d = .46$) with a type I error rate of 100% when sample size equals 400 if it were taken as an estimate

of the unique association of the IV latent variable with the DV latent variable. That is, the ANCOVA/APV estimate in this case does lead to substantial bias reduction and could be useful. To increase the usefulness of ANCOVA/APV in such cases, however, we recommend focusing less on the significance of the ANCOVA/APV estimate and more on the fact that inclusion of the Cov did result in a substantial reduction in the effect size estimate. Along these lines, it might be useful to test the significance of the Cov in accounting for at least a portion of the association between the IV and DV using techniques developed for the testing of mediation such as those developed by Mackinnon and colleagues (e.g., Mackinnon, Lockwood, Hoffman, West & Sheets, 2002), Preacher and Hayes (2004) or Shrout and Bolger (2002). When ANCOVA/APV results in a substantial reduction in effect size with the Cov accounting for a significant portion of the association between the IV and DV measures and sensitivity analyses suggest that underadjustment bias remains a plausible explanation for the significant ANCOVA/APV result, we recommend that researchers should acknowledge that the results might be taken as evidence that the zero-order correlation between the IV and DV measures is spurious and due to the confound of the IV with the Cov.

A final recommendation is to refrain from using the language of control – such as claiming an effect of the *IV* after “controlling for” the *Cov* - when discussing ANCOVA/APV results (Miller & Chapman, 2001). Phrases such as “after partialing” the *Cov* measure or “after covarying” it have less potential to foster overconfidence in ANCOVA/APV results.

Minimizing the Impact of Omitted Variable Bias on Drawing Valid Inferences

In practice *OV*B may often be unavoidable as all of the relevant variables in many areas are not known and the inclusion of some, but not all, relevant variables does not necessarily reduce *OV*B (Clarke, 2005; Rubin, 2006). *Thus, there is no simple solution to OV*B *when randomization is unfeasible or unethical as is often the case in psychopathology research.* Rather, minimizing *OV*B is facilitated by the iterative process of articulating specific *OV*s that might have confounded a given result and then designing studies less vulnerable to that confound or that otherwise allow predictions derived from the original explanation to be pitted against those derived from the confounder explanation. As noted earlier, even the inclusion of a relatively unreliable *Cov* can result in a substantial reduction in bias compared with omission of the *Cov*. To illustrate this point, again consider Example 7 in Table 1 in which the reliability of the *Cov* is nearly zero and therefore approximately equivalent to the case in which the *Cov* had been omitted. Inclusion of a *Cov* with a reliability of .50 would have resulted in substantial bias reduction ($\beta_{DV,IV,Cov} = .076$ with Type I error rates ranging from .074 when *n* equals 60 to .239 when *n* equals 400).

Another practice that might help to minimize the impact of *OV*B on drawing valid inferences is to follow the recommendation of Blalock (1964) and Bollen (1989) to restrict the language of unique effects to those variables within a specific model. As *OV*s are identified and included in subsequent studies as additional covariates, the results would begin to clarify whether uniqueness could then be claimed with respect to the expanded set of covariates (Rosenbaum, 1999; Shadish & Cook, 1999).

Finally, SEM fit indices are sensitive to *OV*B in many cases (Tomarken & Waller, 2003). There are also tests of model misspecification (e.g., Long & Trivedi, 1993) and sensitivity analyses (e.g., Marcus, 1997; Rosenbaum, 2002; Rosenbaum & Rubin, 1983) that can be helpful in suggesting the extent of *OV*B.

Limitations and Conclusion

One limitation of the approach taken here is that there are potentially important problems with ANCOVA/APV that were not addressed here including nonlinear associations among the

LVs and heteroscedasticity. Another limitation is that our approach assumes a classical measurement error model. Techniques for handling other error models such as multiplicative error have been developed and may prove useful in some areas of psychopathology research (e.g., Browne, 1984; Carroll, Ruppert, Stefanski & Crainiceanu, 2006; Marsh, 1989). Almost all extant measurement error models, however, make the assumption that shared method variance exerts a positive bias on correlations. In contrast, Campbell and O'Connell (1982) have raised the provocative possibility that hetero-method correlations may have an attenuating effect and mono-method correlations may be unbiased (in a fashion analogous to that in which differential skew attenuates associations relative to associations among measures that are similarly skewed). This possibility warrants further study. In addition, we only considered type I error rate inflation when there is no unique association between the *IV* and *DV LVs*. When there is a unique association between the *IV* and *DV LVs*, underadjustment for the *Cov LV* can lead to underestimation of this unique association and inflated type II error rates (e.g., Reichardt, 1979). Such effects will arise under different circumstances than those confronting the psychopathologist concerned that a simple correlation between the *IV* and *DV* measures is due to a confounder. These circumstances may be relevant to some psychopathology research, however, and type II error rate inflation in ANCOVA/APV also warrants greater attention than it has received by psychopathologists. Finally, designs in psychopathology studies are often more complex than those involving a single *IV*, a single *Cov* and a single *DV*. The potential for bias in more complex designs is at least as great as in the simpler design considered here and at least as much caution is therefore required for interpreting ANCOVA/APV analyses of more complex designs.

Kenny (1975, p. 360), in likening the difference between true experiments and quasi-experiments to that between testimony from a sighted person and a blind person wisely noted that "when we have only the blind man, we would not dismiss his testimony, especially if he were aware of his biases and had developed faculties of touch and hearing that the sighted man could have developed but has neglected." Unfortunately, psychopathologists rarely give evidence of the awareness of the underadjustment bias in ANCOVA/APV let alone of having made use of approaches that might help to compensate for that bias. Thus, we do not propose to "dismiss testimony" from an ANCOVA/APV. Rather, we hope to raise awareness that as psychopathologists we experience partial blindness due to our inevitable reliance on non-random assignment and we further hope that our recommendations might encourage more widespread use of strategies that can help compensate for our partial blindness.

Acknowledgments

We thank J. Michael Bailey, Emily Durbin, Lewis R. Goldberg, Michael B. Gurtman, Lynne M. Knobloch-Fedders, William Revelle, and the students in Zinbarg's graduate seminar in clinical research methods for their comments on earlier drafts of this article and/or their discussion of the ideas contained in this article. Preparation of this article was supported by the Patricia M Nielsen Research Chair of the Family Institute at Northwestern University and by National Institutes of Health Grants R01- MH65652-01 to Richard E. Zinbarg and R01- EY014110 and EY018197 to Satoru Suzuki, and National Science Foundation grant BCS0643191 to Satoru Suzuki.

References

- Aiken LS, Stein JA, Bentler PM. Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology* 1994;62:488–499. [PubMed: 8063975]
- Allison, PD. Fixed effects regression methods for longitudinal data using SAS. Cary, NC: SAS Institute; 2005.
- Bauer DJ, Curran PJ. The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods* 2004;9:3–29. [PubMed: 15053717]

- Bernstein IH, Teng G. Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin* 1989;105:467–477.
- Blalock, HM. Causal inferences in nonexperimental research. Chapel Hill: University of North Carolina; 1964.
- Bollen, KA. Structural equations with latent variables. New York: Wiley; 1989.
- Bollen KA, Curran PJ. Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods and Research* 2004;32:336–383.
- Browne MW. The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology* 1984;37:1–21. [PubMed: 6733052]
- Campbell, DT.; O'Connell, EJ. Methods as diluting trait relationships rather than adding irrelevant systematic variance. In: Brinberg, D.; Kidder, L., editors. *New Directions for methodology of social and behavioral science: Forms of validity in research*, no. Vol. 12. San Francisco: Jossey-Bass; 1982. p. 93–111.
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement error in nonlinear models: A modern perspective*. 2nd edition. Boca Raton: Taylor & Francis; 2006.
- Cattell, RB. The scientific use of factor analysis in behavioral and life sciences. New York: Plenum; 1978.
- Clark LA, Watson D. Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology* 1991;100:316–336. [PubMed: 1918611]
- Clarke KA. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 2005;22:341–352.
- Cochran WG. Analysis of covariance: Its nature and uses. *Biometrics* 1957;44:261–281.
- Cohen J. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal & Social Psychology* 1962;65:145–153. [PubMed: 13880271]
- Cohen, J.; Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences*. 2nd edition. Hillsdale, NJ: Erlbaum; 1983.
- Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297–334.
- DeShon RP. A cautionary note on measurement error corrections in structural equation models. *Psychological Methods* 1998;3:412–423.
- Elashoff JD. Analysis of covariance: A delicate instrument. *American Educational Research Journal* 1969;6:383–401.
- Fleiss, JL.; Tanur, JM. The analysis of covariance in psychopathology. In: Hammer, M.; Salzinger, K.; Sutton, S., editors. *Psychopathology: Contributions from the social, behavioral and biological sciences*. New York: Wiley; 1973. p. 509–527.
- Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology* 2006;98:253–259. [PubMed: 16611199]
- Hayduk, LA. *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: Johns Hopkins Press; 1987.
- Hertzog C, Nesselroade JR. Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging* 2003;18:639–657. [PubMed: 14692854]
- Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research* 1992;18:117–144. [PubMed: 1459160]
- Huitema, B. *Analysis of covariance and alternatives*. New York: Wiley; 1980.
- Judd, CM.; Kenny, DA. *Estimating the effects of social interventions*. New York: Cambridge University Press; 1981.
- Kahneman D. Control of spurious association and the reliability of the controlled variable. *Psychological Bulletin* 1965;64:326–329. [PubMed: 5851540]
- Kenny DA. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin* 1975;82:345–362.
- Kenny, DA. *Correlation and causality*. New York: Wiley; 1979.
- Little TD, Cunningham WA, Shahar G, Widaman KF. To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling* 2002;9:151–173.

- Little TD, Lindenberger U, Nesselroade JR. On selecting indicators for multivariate measurement and modeling with latent variables: When “good” indicators are bad and “bad” indicators are good. *Psychological Methods* 1999;4:192–211.
- Long, JD.; Trivedi, PK. Some specification tests for the linear regression model. In: Bollen, KA.; Long, JS., editors. *Testing structural equation models*. Newbury Park, CA: Sage; 1993.
- Lord FM. Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association* 1960;55:307–321.
- Lord FM. A paradox in the interpretation of group comparisons. *Psychological Bulletin* 1967;68:304–305. [PubMed: 6062585]
- Lord FM. Statistical adjustments when comparing preexisting groups. *Psychological Bulletin* 1969;72:336–337.
- Mackinnon DP, Lockwood CM, Hoffman J, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods* 2002;7:83–104. [PubMed: 11928892]
- Marcus SM. Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational and Behavioral Statistics* 1997;22:193–201.
- Maxwell, SE.; Delaney, HD. *Designing experiments and analyzing data: A model comparison approach*. 2nd edition. Mahwah, N.J: Erlbaum; 2004.
- Maxwell SE, Delaney HD, Manheimer JM. ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational Statistics* 1985;10:197–209.
- McDonald RA, Behson SJ, Seifert C. Strategies for Dealing with Measurement Error in Multiple Regression. *Journal of Academy of Business and Economics* 2005;53:80–97.
- Miller GA, Chapman JP. Misunderstanding analysis of covariance. *Journal of Abnormal Psychology* 2001;110:40–48. [PubMed: 11261398]
- Muthén BO. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 1984;46:115–132.
- Porter AC, Raudenbush SW. Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology* 1987;34:383–392.
- Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers* 2004;36:717–731.
- Reichardt, CS. The statistical analysis of data from nonequivalent group designs. In: Cook, T.; Campbell, D., editors. *Quasi-Experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin; 1979. p. 147–205.
- Rogosa, D. Myths and methods: “Myths about longitudinal research” plus supplemental questions. In: Gottman, J., editor. *The analysis of change*. Mahwah, NJ: Erlbaum; 1995. p. 3–65.
- Rogosa D, Brandt D, Zimowski M. A growth curve approach to the measurement of change. *Psychological Bulletin* 1982;90:726–748.
- Rosenbaum PR. Choice as an alternative to control in observational studies. *Statistical Science* 1999;14:259–278.
- Rosenbaum, PR. *Observational studies*. 2nd edition. New York: Springer; 2002.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
- Rossi J. Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting & Clinical Psychology* 1990;58:646–656. [PubMed: 2254513]
- Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997;127:757–763. [PubMed: 9382394]
- Rubin, D. *Matched Sampling for Causal Effects*. New York: Cambridge Univ. Press; 2006.
- Sedlmeier P, Gigerenzer G. Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 1989;105:309–316.
- Shadish W, Cook T. Comment-design rules: More steps toward a complete theory of quasi-experimentation. *Statistical Science* 1999;14:294–300.
- Shadish, W.; Cook, T.; Campbell, D. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin; 2002.

- Sher KJ, Trull TJ. Methodological issues in psychopathology research. *Annual Review of Psychology* 1996;47:371–400.
- Shrout PE, Bolger N. Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods* 7:422–445. [PubMed: 12530702]
- Singer, JD.; Willett, JB. *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press; 2003.
- Sorbom, D. An alternative to the methodology for analysis of covariance. In: Joreskog, KG.; Sorbom, D., editors. *Advances in factor analysis and structural equation models*. Lanham, MA: University Press of America; 1979.
- Tomarken AJ, Waller NG. Potential problems with “well fitting” models. *Journal of Abnormal Psychology* 2003;112:578–598. [PubMed: 14674870]
- Vargha A, Rudas T, Delaney HD, Maxwell SE. Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics* 1996;21:264–282.
- Wainer H. Adjusting for differential base rates: Lord’s Paradox again. *Psychological Bulletin* 1991;109:147–151. [PubMed: 2006226]
- Waller NG, Tellegen A, McDonald RP, Lykken DT. Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality* 1996;64:545–576.
- Wildt, AR.; Ahtola, OT. *Analysis of covariance*. Beverly Hills, CA: Sage; 1978.
- Willett J. Questions and answers in the measurement of change. *Review of Research in Education* 1988;15:345–422.
- Williams RH, Zimmerman DW. Are simple gain scores obsolete? *Applied Psychological Measurement* 1996;20:59–69.

Figure 1.

Path diagram of a model with latent variables corresponding to an independent variable (*IV*; depression), a covariate (*Cov*; anxiety) and the dependent variable. Circles represent latent variables and squares represent observed indicators. Paths a , b and f represent the standardized loadings on their respective latent variables of the *Cov*, *IV* and *DV* observed indicators used in a conventional ANCOVA/APV. Paths a' , b' and f' represent the standardized loadings on their respective latent variables of alternative indicators that could be used to measure the *Cov*, *IV* and *DV* latent variables. Paths c , d and e represent the structural relations among the latent variables. See the text for more details.

Figure 2.

Path diagram of a model with latent variables corresponding to an independent variable (*IV*; depression), a covariate (*Cov*; anxiety), an omitted variable (*OV*; life stress) and the dependent variable. Circles represent latent variables and squares represent observed indicators. Paths a , b and f represent the standardized loadings on their respective latent variables of the *Cov*, *IV* and *DV* observed indicators used in a conventional ANCOVA/APV. Paths a' , b' and f' represent the standardized loadings on their respective latent variables of alternative indicators that could be used to measure the *Cov*, *IV* and *DV* latent variables. Paths c , d , e , g and h represent the structural relations among the latent variables. See the text for more details.

Figure 3.

Path diagram of a model with latent variables corresponding to an independent variable (*IV*; depression), a covariate (*Cov*; anxiety) and first dependent variable hypothesized to have a unique association with the *IV* (*DV1*) and a second dependent variable hypothesized to have a unique association with the *Cov* only (*DV2*). Circles represent latent variables and squares represent observed indicators. Paths a , b , f and h represent the standardized loadings between each of the observed indicators used in a conventional ANCOVA/APV with a nonequivalent dependent variable and its latent variable. Paths a' , b' , f' and h' represent the standardized loadings on their respective latent variables of alternative indicators that could be used to measure the *Cov*, *IV*, *DV1* and *DV2* latent variables. Paths c , d , e and g represent the structural relations among the latent variables. See the text for more details.

Table 1

Examples of underadjustment bias and actual type I error rates when the unique association of the independent variable (IV) is zero with a nominal alpha level of .05 as a function of the loadings of the covariate (a), IV (b) and dependent variable (f) measures on their latent variables, the correlations of the latent IV with the latent covariate (c) and omitted variables (g), and the standardized unique association of the latent covariate (d) and omitted variable (h) with the latent dependent variable.

	Example						
	1	2	3	4	5	6	7
a	.900	.900	.800	.800	.938	.938	.100
b	.900	.949	.949	.949	.949	.949	.949
c	.600	.500	.500	.500	.500	.500	.500
d	.600	.500	.500	.300	.300	.300	.300
f	.900	.949	.949	.949	.949	.949	.949
g	—	—	—	—	—	.500	—
h	—	—	—	—	—	.300	—
$r_{DV,IV}$	.292	.225	.225	.135	.135	.270	.135
$\beta_{DV,IV:Cov}$	.073	.052	.095	.057	.020	.189	.134
sr	.063	.047	.088	.053	.018	.136	.134
Effect Size	.00529	.00274	.00904	.00292	.00035	.02119	.01827
Cohen's d	.146	.105	.190	.108	.038	.291	.270
n			Type I	Error	Rates		
60	.086	.068	.112	.070	.052	.198	.177
80	.098	.075	.134	.076	.053	.251	.223
100	.111	.081	.156	.083	.054	.302	.268
200	.176	.114	.267	.118	.058	.535	.477
400	.306	.181	.475	.190	.066	.828	.769

Note. $r_{DV,IV}$ = zero-order correlation between the IV and DV measures; sr = semi-partial correlation between the IV and the DV measures with the Cov measure partialled;

Effect size = $f^2_{effect} = \frac{sr^2}{1 - R^2_{DV|IV,Cov}}$; n = sample size. Cohen's d values assume the IV to be dichotomous with samples of the same size.

Hypothetical results of a sensitivity analysis of a study in which $a = .938$, $b = f = .949$, $c = .50$ and sample size is 140.

Table 2

D	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
$\beta_{DV,IV,Cov}$	0.007	0.013	0.020	0.027	0.034	0.040	0.047	0.054	0.061
sr	0.006	0.012	0.018	0.024	0.030	0.036	0.042	0.048	0.054
Effect									
Size	0.00004	0.0002	0.0004	0.0007	0.0011	0.0018	0.0029	0.0047	0.0083
Cohen's d	0.0121	0.0245	0.0376	0.0516	0.0674	0.0857	0.1081	0.1377	0.1820
α	0.051	0.052	0.056	0.061	0.068	0.079	0.097	0.128	0.188
α'	0.049	0.048	0.045	0.041	0.035	0.029	0.022	0.014	0.007

Note: a = the standardized loading of the covariate measure on its latent variable; b = the standardized loading of the independent variable measure on its latent variable; c = the correlation between the covariate and independent variable latent variables; f = the standardized loading of the dependent variable measure on its latent variable; d = the standardized unique association of the latent covariate with the latent dependent variable; $\text{Effect size} = f^2_{effect}$; sr = semi-partial correlation between the independent variable and the covariate; α = actual type I error rate using a nominal level of .05; α' = nominal type I error rate corrected such that the actual type I error rate will equal .05. Cohen's d values assume the IV to be dichotomous with samples of the same size.

Hypothetical results of a sensitivity analysis of a study in which $a = .850$, $b = f = .949$, $c = .50$ and sample size is 300.

Table 3

d	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
$\beta_{DV,IV,Cov}$	0.015	0.030	0.045	0.060	0.075	0.089	0.104	0.119	0.134
Sr	0.014	0.027	0.041	0.055	0.068	0.082	0.096	0.109	0.123
Effect									
Size	0.0002	0.0008	0.0018	0.0034	0.0057	0.0090	0.0138	0.0212	0.0334
Cohen's d	0.0276	0.0558	0.0852	0.1165	0.1508	0.1895	0.2349	0.2909	0.3654
α	0.051	0.077	0.114	0.171	0.256	0.373	0.527	0.709	0.884
α'	0.049	0.030	0.017	0.008	0.003	0.001	0.0003	0.00004	0.000002

Note: a = the standardized loading of the covariate measure on its latent variable; b = the standardized loading of the independent variable measure on its latent variable; c = the correlation between the covariate and independent variable latent variables; f = the standardized loading of the dependent variable measure on its latent variable; d = the standardized unique association of the latent covariate with the latent dependent variable; $\text{Effect size} = f^2_{effect}$; sr = semi-partial correlation between the independent variable and the covariate; α = actual type I error rate using a nominal level of .05; α' = nominal type I error rate corrected such that the actual type I error rate will equal .05. Cohen's d values assume the IV to be dichotomous with samples of the same size.