ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

International Journal of Psychophysiology 52 (2004) 277–283

INTERNATIONAL
JOURNAL OF
PSYCHOPHYSIOLOGY

www.elsevier.com/locate/ijpsycho

Analysis of covariance (ANCOVA) with difference scores

John Jamieson*

Department of Psychology, Lakehead University, Thunder Bay, ON, Canada P7B 5E1

Received 3 June 2003; received in revised form 4 December 2003; accepted 4 December 2003

Abstract

When comparing pretest to posttest changes in non-randomized groups, most researchers are correctly avoiding ANCOVA with posttest as the dependent variable and pretest as the covariate. However, there is a widespread use of ANCOVA in which the difference score (posttest minus pretest) is used as the dependent variable, and pretest as the covariate. A computer simulation study is presented which shows that measurement error causes identical, biased conclusions when comparing changes using either the posttest score or the posttest minus pretest difference score as the dependent variable. The reasons for this bias are explained and illustrated.

Keywords: Analysis of covariance (ANCOVA); Difference scores; Change; Measurement error

1. Introduction

Miller and Chapman (2001) have recently directed attention to the widespread but incorrect belief that ANCOVA can be used to control for or eliminate non-trivial group differences. An assumption underlying ANCOVA is that the covariate is independent of group membership, which happens with random assignment to groups, but not with naturally occurring groups. For example, if control and experimental groups are created by random assignment, possible covariates such as body mass index (BMI) will be independent of group membership. However, if groups are naturally occurring, such as male and female, then BMI will not be independent of group membership. ANCOVA is designed to control for covariates when groups were randomly assigned, but not to control for naturally occurring group differences.

One application of ANCOVA of particular relevance to psychophysiologists is to control for baseline (pretest) differences. When groups differ in baseline, ANCOVA may be used to control for these differences. The usual way to do this ANCOVA is to use the posttest score as the dependent variable, and the pretest score as the covariate. By removing the variance explained by the pretest from the posttest, the residual is variation that reflects the change from the pretest. When groups are assigned at random, ANCOVA is an excellent method for comparing changes between groups. However, when groups are naturally occurring, the baseline differences are not due to chance, and ANCOVA will yield biased conclusions.

*Tel.: +1-807-343-8738; fax: +1-807-346-7734.
E-mail address:
john.jamieson@lakeheadu.ca (J. Jamieson).

0167-8760/04/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpsycho.2003.12.009

The fact that ANCOVA (with covariate = pretest and dependent variable = posttest) should not be used to compare changes between naturally occurring groups has been pointed out many times (e.g. Huitema, 1980; Rogosa, 1988; Schafer, 1992). Inspection of recent journals in psychophysiology and health psychology indicates that this use of ANCOVA to compare changes is appearing less frequently. However, a variation of this usage is quite widely used, namely, where the dependent variable is the difference score (posttest minus pretest) and the covariate is the pretest. Because this alternative use of ANCOVA is apparently viewed as superior to the traditional method, computer simulations will be presented to show the two ANCOVA methods are in fact identical.

To understand why ANCOVA produces biased conclusions when the covariate is not independent of group membership, it is useful to examine how ANCOVA adjusts the posttest means for pretest differences. ANCOVA uses two methods for removing the influence of the covariate(s) on the dependent variable. The first method focuses within each group and calculates regression lines for predicting the dependent variable from the covariate in each group. These regression lines are used to find the predicted dependent variable score for each case based on their score on the covariate. The residual scores for each case (observed score on the dependent variable minus the predicted dependent variable score) are pooled to calculate an error term. This within group use of regression is an excellent method for removing the effect of the covariate from error variance, and is not a source of controversy.

The second method of adjustment in ANCOVA is much more problematic. The regression lines from each group are pooled (hence the assumption of homogeneity of regression) to obtain a single regression coefficient (b). This pooled regression coefficient is then used in a formula to adjust the mean on the dependent variable for each group. To facilitate communication, a concrete example will be used, instead of notation. The example is based on the most famous illustration of misuse of ANCOVA, Lord's paradox (Lord 1967). Lord presented a hypothetical example of a group of male and a group of female adolescents each weighed on two consecutive years (called 'year 1' and 'year 2'). Even though both groups gained the identical number of pounds, ANCOVA resulted in the incorrect conclusion that males increased significantly more than did females.

The critical feature of ANCOVA, which leads to the incorrect adjustments, is that it incorrectly assumes both males and females in the population actually have the same average weight at year 1. ANCOVA obtains an estimate of this average weight in the population by averaging the year 1 weights for both males and females. For example, if males averaged 130 lb at year 1 and females averaged 110 lb, the average of these two weights, 120 lb, would be used as an estimate of the population mean weight at year 1. If the groups had been assigned at random, such an estimate would make sense, but when the groups are naturally occurring, it does not.

ANCOVA then uses a formula to adjust the observed year 2 weight for each of males and females. This formula takes into account both the difference between the actual year 1 weight for each group and the estimated population average year 1 weight, as well as the pooled regression coefficient. For example, if males weighed 140 lb at year 2 and females weighed 120 lb at year 2 (both groups gained the identical weight of 10 lb), the ANCOVA would adjust the average year 2 weights using the following formulas:

Adjusted year 2 weight = year 2 weight − b(year 1 weight − population year 1 weight).

For males: adjusted year 2 weight = 140 − b(130 − 120).
For females: adjusted year 2 weight = 120 − b(110 − 120).

If there is a moderate amount of measurement error, b might be 0.5, which would yield the following adjusted values:

For males: adjusted year 2 weight = 140 − 0.5(130 − 120) = 140 − 5 = 135.
For females: adjusted year 2 weight = 120 − 0.5(110 − 120) = 120 + 5 = 125.

Under the null hypothesis that males and females will show identical weights at year 2, after adjusting for year 1 weight (i.e. there is no difference between the changes in weight for males and females), the adjusted year 2 weights for each group should be equal. But because 135 is not equal to 125, ANCOVA will result in the conclusion that there is a significant difference in changes, and that males changed more (since their adjusted year 2 mean is higher than the adjusted year 2 mean for females). This is Lord's paradox: a significant difference in changes found by ANCOVA, when the actual difference in changes was identical for both males and females (10 lb).

However, it can readily be seen that this paradox is simply due to measurement error. If there is no measurement error, the weights on year 1 and year 2 will be perfectly correlated and b will equal 1.0 (assuming homogeneity of variance from year 1 to year 2). When $b=1.0$, the adjusted year 2 weights for both males and females will be identical (130 lb), and ANCOVA will yield the correct conclusion that the two groups showed identical changes in weight from year 1 to year 2. As measurement error increases (as b approaches zero), the formulas produce less adjustment, which results in more divergent adjusted means, and hence more bias. When b is less than 1.0, the year 1 difference in weight between males and females is not totally included in the adjusted year 2 weights. This results in a difference between the adjusted year 2 means, which is interpreted by ANCOVA as a significant difference in changes.

This example illustrates how measurement error causes ANCOVA to result in misleading conclusions, when comparing naturally occurring groups. If measurement error is present, b will be less than 1, and there will be less adjustment to the dependent variable means (year 2 weights) based on the covariate differences (year 1), which will result in an under-adjustment of the predicted mean year 2 weights. This under-adjustment will yield differences in the adjusted means for each group, which will be interpreted by ANCOVA as a significant difference. However, if the groups had been assigned at random, the group with the higher year 1 mean would contain more positive errors, which would regress on retesting at year 2. The b coefficient takes into account this regression. In contrast, when the groups are naturally occurring, there is absolutely no reason to expect the means to regress at year 2. On the contrary, it is much more reasonable to assume the year 1 differences are 'real' (unbiased estimates) and will also be present at posttest. So, with naturally occurring groups, ANCOVA produces an incorrect adjustment. It can also be seen that this incorrect adjustment is caused by measurement error.

A set of computer simulations is presented to show that using ANCOVA to compare changes from pretest (covariate) to posttest (dependent variable) produces a bias, simply as a result of measurement error. Because of the widespread usage of ANCOVA with the pretest minus posttest difference score as the dependent variable (in place of the posttest score), the following computer simulations also include this analysis, to show that using difference scores as the dependent variable is identical to using posttest as the dependent variable, and results in equally biased conclusions.

2. Method

The SAS generator RANNOR (SAS Institute, 1990) was used to generate pseudorandom variates, with means of zero and standard deviations (see below) selected to yield realistic effects. One thousand simulations were computed for each of the 12 conditions.

For comparison of changes between two groups, pretest and posttest scores were created for two groups, each of $n=25$, by adding a different error component ($\mu=0$) to the same true score component ($\mu=0$, $\sigma=10$). The standard deviation of the error component was varied (values = 1, 5 and 20) to create conditions with different amounts of measurement error. Baseline differences between groups were created by adding a constant to the pretest and posttest values of one group. The constants were 0, 5, 10 and 20. These baseline differences were also present at posttest. In this way the data simulated differences in naturally occurring groups that were not due to chance, and did not regress towards zero at posttest.

Difference scores for the repeatedly assessed variables were computed by taking the difference

Table 1
Rate of Type 1 errors for ANOVA on difference scores as a function of baseline differences and measurement error

Error S.D.	Baseline difference			
	0	5	10	20
1	0.052	0.044	0.048	0.048
5	0.055	0.047	0.060	0.051
20	0.047	0.049	0.044	0.055

between the pretest and posttest scores. ANOVA was used to compare the mean difference scores of the two groups. Two ANCOVAs were conducted. In both cases, pretest was the covariate. In one ANCOVA the dependent variable was the posttest, while in the other ANCOVA the dependent variable was the posttest minus pretest difference score. The proportion of significant ($P<0.05$) samples (Type 1 errors) are reported for each condition.

3. Results

Table 1 presents the Type 1 error rates for ANOVA comparing mean changes in the two groups from pretest to posttest for each of the 12 conditions. In all conditions, the Type 1 error rate was approximately 0.05.

Tables 2 and 3 present the Type 1 error rates for the two ANCOVAs. Both ANCOVAs gave identical answers, showing that ANCOVA with posttest as the dependent variable and pretest as the covariate is identical to ANCOVA with the difference score as the dependent variable and pretest as the covariate. When there was no baseline difference (first column), ANCOVAs also yielded approximately 5% Type 1 errors. However, when there were baseline differences, the rate of Type 1 errors increased. Except for the no baseline difference conditions, as measurement error increased, the Type 1 error rate also increased. The highest error rates (63%) appeared when the highest level of measurement error was combined with the largest baseline difference.

4. Discussion

These simulations clearly show that using ANCOVA with a difference score (posttest minus pretest) as the dependent variable and pretest as the covariate is identical to using ANCOVA with posttest as the dependent variable and pretest as the covariate. The simulations yielded identical answers in each condition. This result shows that using difference scores as the dependent variable provides no advantage over the use of posttest scores. ANCOVA on difference scores shares with ANCOVA on posttest scores the well-documented problems for measuring change with naturally occurring groups (Huitema, 1980; Rogosa, 1988; Schafer, 1992).

The simulations also show that the problem with using ANCOVA to measure change (Lord's paradox) results from error variance. As the size of the error variance approached zero (error S.D. of 1 was the smallest examined in the simulations), the rate of Type 1 errors approached 0.05. However, when the largest error variance condition (error S.D. of 20) was combined with the largest baseline difference condition (20), the highest rate of Type 1 errors was found (63%). The present simulations manipulated error in both the pretest and posttest to mimic realistic data. Additional simulations, not presented here, showed that error, only in the covariate, is sufficient to produce the

Table 2
Rate of Type 1 errors for ANCOVA on posttest, controlling for pretest, as a function of baseline differences and measurement error

Error S.D.	Baseline difference			
	0	5	10	20
1	0.051	0.045	0.059	0.068
5	0.044	0.073	0.175	0.326
20	0.044	0.075	0.225	0.630

Table 3
Rate of Type 1 errors for ANCOVA on difference scores, controlling for pretest, as a function of baseline differences and measurement error

Error S.D.	Baseline difference			
	0	5	10	20
1	0.051	0.045	0.059	0.068
5	0.044	0.073	0.175	0.326
20	0.044	0.075	0.225	0.630

bias, while error, only in the dependent variable, just reduces power. It is well documented that error in the covariate causes imperfect removal of the covariate in both ANCOVA (Huitema, 1980) and multiple regression (Cohen and Cohen, 1983).

In the present simulations, pretest differences were also present at posttest. If the baseline differences had been due to random assignment, they would have regressed, and ANCOVA would not have produced the elevated level of Type 1 errors. Instead, these simulations were created to reflect the sort of data that are found with non-randomized (naturally occurring) groups, namely, group differences at baseline will also be present at posttest. These differences thus reflect real, not chance, differences between the groups.

To understand the reasons for this bias, it is useful to consider how the data simulations were created in the present study. First, it was assumed that individual differences at pretest would also be present at posttest. Thus, the same true score term was included in both the pretest and posttest scores, along with the error variance. As well, for one group, the scores on both pretest and posttest were ‘shifted’ by the addition of a constant. This mimicked naturally occurring group differences, which would appear at both pretest and posttest. These differences did not regress, but the ANCOVA adjustment predicted they would regress (the *b* coefficient was less than 1.0). The failure of the group differences to be diminished at posttest led to the finding of differences in the changes (Type 1 error).

It is important to note that measurement error always produces under-adjustment, which results in a directional bias. Jamieson (1999) described in some detail the directional bias that results from the misuse of ANCOVA to measure change with naturally occurring groups. Basically, the directional bias is for the group with the higher baseline mean to be found to increase more. This is the finding of Lord’s paradox, and Baldwin et al. (1984) have also pointed out that ANCOVA with naturally occurring groups is biased, to find greater change in the group with the higher baseline mean. To illustrate this directional bias, it is perhaps useful to consider another example. When studying the effect of parental history of hypertension on cardiovascular reactivity, it will generally be found that those participants with positive parental histories of hypertension have significantly higher baseline cardiovascular levels. Because these baseline differences are not due to chance, they will also appear at posttest (e.g. after a stressor). Therefore, ANCOVA using these resting levels as covariates will be biased to find that the group with a positive parental history of hypertension shows greater cardiovascular reactivity to stress. This finding would be a statistical artifact of the misuse of ANCOVA. The group with the higher baseline mean will be found by ANCOVA to increase significantly more than the group with the lower baseline mean, simply as a result of measurement error.

It is possible to assess the likelihood of this bias due to measurement error by examining the pretest–posttest correlation. Unless the correlation is very close to 1.0, there will be measurement error, which will create the bias. For less reliable measures, the bias will be of greater magnitude. When measuring change, such measurement error will invariably be present, which has led to the generally accepted position that ANCOVA should not be used to measure change unless the groups were created by random assignment (e.g. Schafer, 1992).

While the traditional usage of ANCOVA to measure change (posttest as the dependent variable) has become rare, the present findings show that the widespread usage of ANCOVA to measure change using the difference score as the dependent variable should also be avoided. This raises the question of what analysis is appropriate for comparing changes in naturally occurring groups. If sufficient sample size and multiple measures are available, structural equation models of change (Raykov, 1992) are not susceptible to the ANCOVA bias caused by measurement error (Cribbie and Jamieson, 2000). However, for the two-phase design with a single dependent variable, it is most appropriate to simply compare the posttest minus pretest difference scores using *t*-tests or ANOVA. Early concerns about the unreliability of difference scores (Cronbach and Furby, 1970) are no longer seen as obstacles (Llabre et al., 1991; Williams and Zimmerman, 1996). Difference scores are unaffected by the presence of non-randomized differences between the groups, as can be seen from the present simulations. As well, difference scores yield unbiased estimates of change so long as the data are not skewed (Collins, 1996). Cribbie and Jamieson (submitted) have described some of the problems with the use of difference scores and structural equation models of change when data are skewed, for example by floor/ceiling effects.

However, there is another reason why researchers may incorrectly choose to use ANCOVA. Researchers frequently want to control for other confounding variables in addition to baseline differences. For example, researchers may want to equate the groups not just for baseline differences, but also for differences on other measures such as BMI, personality dimensions, or socioeconomic measures. This rationale for using ANCOVA is also flawed when groups are not randomly assigned. As Miller and Chapman (2001) point out, the problem with using ANCOVA to remove non-trivial group differences applies to any use of ANCOVA with naturally occurring groups. Again, the problem is that ANCOVA assumes the group differences on the covariates are due to chance, and will regress. Since these differences are not due to chance, ANCOVA will produce a biased adjustment.

Ideally, ANCOVA should be used to remove covariates that have a high correlation with the dependent variable, but have no relationship to the independent variable (no group differences). ANCOVA is an excellent method for increasing power through removing the within group variation explained by the covariate. However, when naturally occurring groups have non-trivial differences on the covariate, ANCOVA does not provide appropriate adjustments to the dependent variable. One clear recommendation is to avoid use of ANCOVA when naturally occurring groups have large differences on the covariate. As the magnitude of the differences on the covariate increases, the likelihood of incorrect adjustments also increases. Attempts to control for or eliminate large, naturally occurring group differences on the covariate will invariably yield highly questionable conclusions. Unfortunately, the size or significance of group differences on the covariate is often not reported. Perhaps reviewers and editors should insist on the inclusion of such information, so the reader might assess the likelihood that findings from ANCOVA with naturally occurring groups may be incorrect. A rough guideline might be to avoid using covariates with naturally occurring groups, unless the relationship between the covariate and the dependent variable is much larger than the relationship between the covariate and the independent variable.

In summary, the present study indicates that there are no advantages of using ANCOVA on difference scores over ANCOVA on posttest scores. Both methods should be avoided when comparing changes in naturally occurring groups, which differ in baseline, since both methods yield biased and misleading conclusions.

References

Baldwin, L., Medley, D., MacDougall, M., 1984. A comparison of covariance to within-class regression in the analysis of non-equivalent groups. J. Exp. Educ. 52, 68–76.

Cohen, J., Cohen, P., 1983. Applied Multiple Regression/Correlation Analyses for the Behavioral Sciences. 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Collins, L.M., 1996. Is reliability obsolete? A commentary on 'Are simple gain scores obsolete?' Appl. Psychol. Meas. 20, 289–292.

Cribbie, R.A., Jamieson, J., 2000. Structural equation models and the regression bias for measuring correlates of change. Educ. Psychol. Meas. 60, 893–907.

Cribbie, R.A., Jamieson, J., Decreases in posttest variance and the measurement of change, (submitted).

Cronbach, L.J, Furby, L., 1970. How should we measure 'change'—or should we? Psychol. Bull. 105, 68–80.

Huitema, B.E., 1980. The Analysis of Covariance and Alternatives. Wiley, New York.

Jamieson, J., 1999. Dealing with baseline differences: two principles and two dilemmas. Int. J. Psychophysiol. 31, 155–161.

Llabre, M.M., Spitzer, S.S., Saab, P.G., Ironson, G.H., Schneiderman, N., 1991. The reliability and specificity of delta vs. residualized change as measures of cardiovascular reactivity to behavioral challenges. Psychophysiology 28, 701–711.

Lord, F.E., 1967. A paradox in the interpretation of group comparisons. Psychol. Bull. 68, 304–305.

Miller, G.A., Chapman, J.P., 2001. Misunderstanding analysis of covariance. J. Abnorm. Psychol. 110, 40–48.

Raykov, T., 1992. Structural models for studying correlates and predictors of change. Aust. J. Psychol. 44, 101–112.

Rogosa, D.R., 1988. Myths about longitudinal research. In: Schaie, K.W., Campbell, R.T., Meredith, W., Rawlings, S.C. (Eds.), Methodological Issues in Aging Research. Springer, New York, pp. 171–209.

SAS Institute, 1990. SAS Procedures Guide. 3rd ed. SAS Institute, Cary, NC Version 6.

Schafer, W.D., 1992. Analysis of pretest–posttest designs. Meas. Eval. Couns. Dev. 25, 2–4.

Williams, R.H., Zimmerman, D.W., 1996. Are simple gain scores obsolete? Appl. Psychol. Meas. 20, 59–69.