



## Comparing groups in a before–after design: When $t$ test and ANCOVA produce different results

Daniel B. Wright\*

University of Sussex, UK

**Background.** Researchers often test people before and after some treatment and compare these scores with a control group. Sometimes it is not possible to allocate people into conditions randomly, which means the initial scores for the two groups may differ. There are two main approaches:  $t$  test on the gain scores and ANCOVA partialling out the initial scores. Lord (1967) showed that these can lead to different conclusions. This is an often-discussed paradox in psychology and education.

**Aims.** The reasons why these approaches can lead to different conclusions, the assumptions that each approach makes and how the approaches relate to group allocation, are discussed

**Methods.** Three sets of simulations are reported that investigate the relationships among effect size, group allocation, measurement error and Lord's paradox.

**Conclusions.** Recommendations are given that stress careful examination of the research questions, sampling and allocation of participants and graphing the data. ANCOVA is appropriate when allocation is based on the initial scores,  $t$  test can be appropriate if allocation is associated non-causally with the initial scores, but often neither approach provides adequate results.

A common design in psychology and education is where two groups are measured before and after some treatment. Ideally, people should be randomly allocated into the groups, but this is not always possible. This paper focuses on the situation when random allocation is not used, and therefore the groups are often assumed to be non-equivalent. The pre-treatment measurement takes place so that any initial between-group differences can, in some way, be taken into account in determining whether any post-treatment difference is due to the treatment, pre-existing differences, or some combination of these. If random allocation is used this is not problematic because the groups' initial scores should not differ systematically.

The two most common statistical approaches are doing a  $t$  test on the gain scores (post-score minus pre-score) and an analysis of covariance (ANCOVA) partialling out the initial score. Lord (1967) showed that these alternatives can lead to different

\*Correspondence should be addressed to Daniel B. Wright, Department of Psychology, University of Sussex, Falmer BNI 9QH, UK (e-mail: danw@sussex.ac.uk).

conclusions. Both approaches are valid descriptions of the data and they address very similar research questions; thus the apparent paradox. However, the questions they address are different (Hand, 1994) and subsequent conclusions require different assumptions (Wainer, 1991; Wainer & Brown, 2004).

Many postgraduate textbooks on statistics discuss before-after designs and Lord's paradox. For example, Maxwell and Delaney (2004, pp. 444-448) describe the paradox and say that an ANCOVA is usually the preferred approach, but that there are situations where analysis of the differences is the preferred approach. They describe how applied researchers are often interested in whether each group has increased in scores, and therefore analysis of the differences can sometimes be more informative than the ANCOVA.

In Wright (2003), I made several recommendations for conducting and reporting statistics for this and other journals. Due to the nature of that paper each topic was only discussed briefly. For before-after designs I said:

The only procedure that is always correct in this situation is a scatterplot comparing the scores at time 2 with those at time 1 for the different groups. In most cases you should analyse the data in several ways. If the approaches give different results . . . think more carefully about the model implied by each. (p. 130)

Since the publication of those guidelines I have been asked numerous times to expand both upon this advice for different situations and also to describe this interesting paradox in more detail. These are the aims of the current paper.

In the following section, Lord's paradox is illustrated with an example and the two main approaches ( $t$  test and ANCOVA) are compared. In addition, an ANCOVA allowing different slopes is also conducted (Rutherford, 2000). Following several authors, Rubin's model of causality is used to help explain the different approaches, although notation is kept to a minimum and the model is described at a level appropriate for non-statisticians. Next, a series of simulations are conducted that show some of the situations where the two approaches are likely to yield different results. These simulations are based upon the theoretical derivations of Maris (1998).

### Lord's (1967) paradox

Suppose a researcher wanted to examine the value of a new method for teaching arithmetic. Five children from one intact group are allocated to a control group and five children from another group are allocated to a group who are provided with some supplementary instruction (SI). As often occurs in educational research it is not practical to allocate people randomly into groups, so differences before the SI takes place are likely. The researcher gives all the children an arithmetic test before the SI occurs (*pre*) and then again after the SI (*post*). The hypothetical data are shown in Table 1. The means are:  $pre = 30$  for the control group,  $pre = 70$  for the SI group,  $post = 30$  for the control group, and  $post = 70$  for the SI group.

Because the means for both groups are the same before and after the treatment it seems reasonable to assume that the SI was not beneficial. Many researchers would use a  $t$  test on the gain scores:  $post_i - pre_i$ . For comparison with an alternative approach it is useful to describe this approach using a regression format. Let  $group_i$  be 0 for the control group and 1 for the group that receives the SI. Written as a regression this is:  $post_i = pre_i + \beta_1 group_i + \beta_0 + e_i$ , and the null hypothesis is  $\beta_1 = 0$ . Solving this for the data in Table 1 yields:

**Table 1.** Data for a hypothetical example

	Pre	Post	Difference
Control instruction	10	20	10
	20	25	5
	30	30	0
	40	35	–5
	50	40	–10
Supplementary group	50	60	10
	60	65	5
	70	70	0
	80	75	–5
	90	80	–10

$$E(post_i) = pre_i + 0 \text{ group}_i + 0,$$

where  $E(post_i)$  means the expected value for  $post_i$ . Therefore, the hypothesis  $\beta_1 = 0$  is not rejected. The research question being asked here is whether any change that occurs is on average greater for the control group or the SI group.

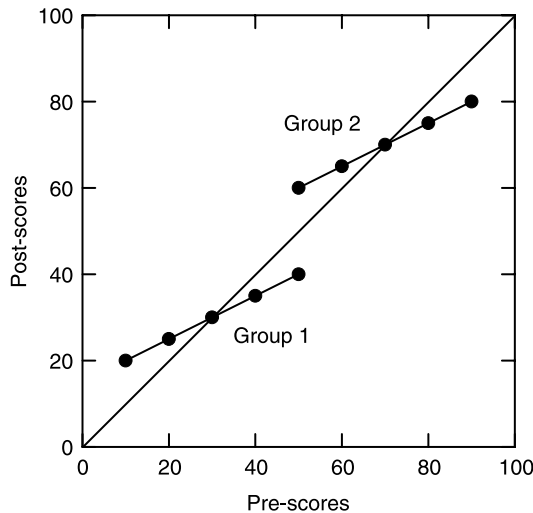
Consider a different question. Suppose that you are a parent and you want to decide whether your child should be given the SI. A teacher gives your child the pre test and the score is 50. You want to see for this score if giving the SI predicts a higher post-score. You run an ANCOVA with the form,  $post_i = \beta_2 pre_i + \beta_1 group_i + \beta_0 + e_i$ , and find  $E(post_i) = 0.5 pre_i + 20 group_i + 15$ . The expected value for your child if s/he is in the control group is 40. The predicted value if in the SI group is 60. Therefore, you would choose that your child had the SI. In fact, regardless of your child's pre-score, the expected post-score is 20 points higher with the SI than without it. But the previous analysis showed that the SI did not make a difference with respect to the amount of improvement. Hence, the paradox.

When framed as regressions, one way to conceptualize the differences between the approaches becomes clear:

$$t \text{ test: } post_i = pre_i + \beta_1 group_i + \beta_0 + e_i$$

$$\text{ANCOVA: } post_i = \beta_2 pre_i + \beta_1 group_i + \beta_0 + e_i.$$

Conducting a  $t$  test on the gain scores is equivalent to the ANCOVA approach if the slopes of the regression lines ( $\beta_2$ ) are equal to 1. The approaches are graphically depicted in Figure 1. The diagonal line ( $post_i = pre_i$ ) shows where there are no differences between the gain scores at the two points in time. The vertical distance between this line and the individual points shows the amount of improvement or decline for each individual. The  $t$  test on the gain scores compares these distances for the two groups. For the data in Table 1, overall neither group improves so there is no difference in their improvement. However, the ANCOVA shows that the regression line for the control group is higher than the regression line for the SI group. For all pre-scores, the expected post-score is 20 points higher for children with the SI than without it. Identifying the difference in these conclusions shows that Lord's paradox can occur, but it does not show which approach should be used or when it is likely that these approaches lead to different conclusions.



**Figure 1.** The data from Table 1 presented graphically. Group 2 received the SI; Group 1 is the control group.

There are two ways that have been used to address which of these, if either, is the appropriate approach. The first is about how best to describe the data. The second is about making causal inference. Hand (1994) describes how the *t* test and ANCOVA are asking different research questions. The first (*t* test) asks whether the average gain in score is different for the two groups. The second (ANCOVA) asks whether the average gain, partialling out pre-scores, is different between the two groups. In one sense this is simply translating the regression equations into something that resembles English, but by doing so it makes clear that the pre-scores play two roles in the ANCOVA approach if the interest is on the improvement. They are used in calculating the gain score and they are used in predicting the gain score. This becomes clearer if the regression forms of the two approaches are rewritten as:

$$t \text{ test: } post_i - pre_i = \beta_1 group_i + \beta_0 + e_i$$

$$ANCOVA: post_i - pre_i = (\beta_2 - 1)pre_i + \beta_1 group_i + \beta_0 + e_i$$

The main idea behind Hand's paper was that researchers have to be precise about the research questions that they ask. As Lord's paradox shows, subtle differences between questions can produce different answers.

Here the *t* test approach finds that there are no differences between the groups on much they change, which suggests that it does not matter which group you are in. The ANCOVA finds that the supplementary instruction is valuable, and therefore suggests that it is better to be in the supplementary instruction group. Because the two approaches can lead to very different conclusions, arbitrarily choosing one approach is not good. The problem is magnified by there being more than just two possible research questions. For example, it might make more sense to look at the ratio of  $post_i$  and  $pre_i$  than the difference, or some other function of these (Wright, 1997). Given the number of possibly valid research questions, in many cases it is unlikely that the typical psychologist will have considered many of them, and even less likely that it will be clear which, if any, is the most appropriate. It would be wrong to think that there is always

a single right answer, which is why my advice in Wright (2003) was to use multiple approaches to describe the data if you are unsure about the research question.

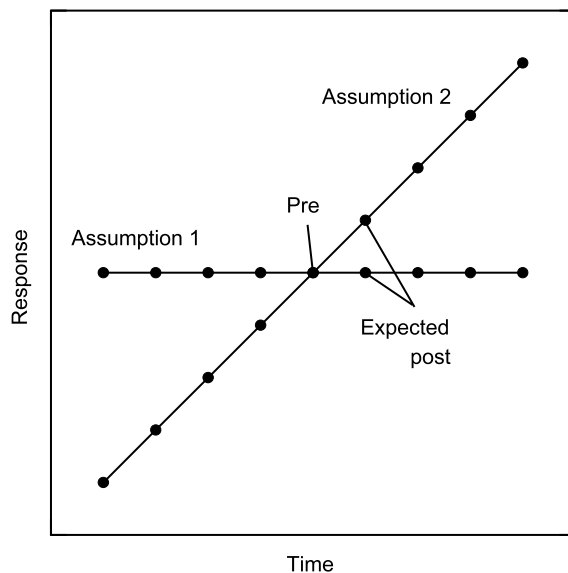
Often there is a more specific question that is relevant in these situations. It is not about describing the data, but about inferring causality (Wright, 2006b). While there is not a right answer for how best to describe the data, there can be a best way to test if a manipulation has an effect. Researchers often want to infer whether one teaching method on average caused a greater difference. Wainer (1991) and Maris (1998) have used Rubin's model of causality (see Holland, 1986) to explore when the *t* test or ANCOVA approach is better for measuring the average gain caused by the treatment. These papers focus on different aspects so both will be reviewed after a brief description of Rubin's model (see Holland, 1986, for details).

If you want to measure the effect of some treatment, ideally you would want to measure the same people, in the exact same situations, both with and without the treatment and compare the scores in these two conditions. This may be possible in computer simulations, but in much psychology research it is not possible to have people in both an experimental and a control condition. A between-subjects design is necessary. Using the current example, the researcher would want to compare people's scores in the SI group with the scores that they would have received if they had not been given any SI. Similarly, the effect can be measured by comparing the scores from the control group with the scores that they would have received if they had been given the SI. The problem is that both of these ways of measuring the effect involve a term that cannot be directly measured. It is necessary to predict how the SI group would have performed without the SI and how the control group would have performed with the SI. This is the basis for Rubin's model.

In a randomized study how the SI group would have performed without the SI can be estimated by the control group (and vice versa for how the control group would have performed with the SI). The difficulties arise when random assignment is not used. The initial or baseline scores can be used to estimate the unobserved conditions, but this requires some assumptions (Wainer, 1991). The before–after design involves measurements at two points in time. To explain the assumptions Wainer asks readers to imagine that the study occurred over several measurements (see Figure 2, based on Wainer, 1991, Figure 2). Imagine two situations. In the first if there was no treatment people's scores would tend to continue in this way. Thus, the predicted score for the SI group without the SI would be their initial baseline score. The size of the effect would be the difference between the mean  $post_i$  and  $pre_i$  scores for each group. If you make this assumption (labelled Assumption 1 in Figure 2), then the *t* test between these gains scores would be the appropriate way to measure if the SI had an effect.

The second situation is where scores follow a linear trend. The line for Assumption 2 shows that scores before treatment are steadily increasing. If there was no intervention, the expectation is that scores would increase. This increase would mean that you would expect  $\beta_2 \neq 1$ . The expected post-score would be a linear function of the form  $\beta_2 pre_i + \beta_0$ . It is assumed that the slope ( $\beta_2$ ) is the same for both groups. This is the way that ANCOVAs are normally conducted, although often this assumption is not valid (Rutherford, 2000). The average effect is the difference between the two intercepts, which is the  $\beta_1$  in the equations presented earlier. Thus, the ANCOVA approach would be appropriate for Assumption 2.

Because there are only two measurement points, it is not possible in many cases to determine which of these assumptions, if either, is appropriate. Maris (1998) examines some of the reasons that might affect this decision. In particular, Lord's paradox arises



**Figure 2.** Applying Rubin's model of causality to Lord's paradox.

because of non-random assignment. Maris examined the relationship between how people are allocated to the groups and the initial baseline scores. If the baseline scores are used to allocate people to groups, then the ANCOVA approach is preferred. Otherwise, he describes how there is no general preference between the approaches. In the next section these situations, and others, are illustrated using Monte Carlo (simulation) methods.

One final consideration is the power of the two methods for detecting differences between the groups (see Baguley, 2004, for discussion of power in general, and Oakes & Feldman, 2001, with reference to the designs discussed here). One theme of this paper is that researchers should be aware of the specific research questions, the assumptions that they are making, and how the allocation into groups is made, and this should lead them to conduct certain statistical analyses. However, in practice, researchers may not be able to address these issues and may simply want to know which method is more likely to discriminate among the groups. The ANCOVA is the more powerful procedure when groups are allocated randomly (Oakes & Feldman, 2001). This can be explained in several ways. If we consider the difference between the two previous equations, because the ANCOVA allows an extra parameter to be fit, this will make the error sum of squares smaller. With non-random allocation, however, allowing the extra parameter can in certain situations make the group effect smaller and therefore adversely affect the power. In general, neither procedure is more powerful than the other (Oakes & Feldman, 2001). This is explored in the current simulations. The focus is on the variability of the unstandardized effect sizes.

In summary, when using a before-after design with two groups, researchers often use either a *t* test on the differences or an ANCOVA partialling out the initial scores. These analytic techniques address different research questions and are based on different assumptions, and therefore they sometimes produce different results. This has become known as Lord's paradox. It is an apparent paradox because researchers think that both research questions are relevant and often they have difficulty deciding which, if either, is appropriate.

### Simulation studies

Three sets of simulations are reported. Others were run, but these three were chosen because they highlight the most interesting differences. The following research situations were explored:

- (1) when the baseline measurement is used to allocate people to groups,
- (2) when the baseline measurement is associated non-causally with group membership,
- (3) when the slopes of the regression lines are not parallel.

Situations both with and without effects, and with and without measurement error, are simulated.

The simulations were based on 1,000 replications for each condition unless specified. All were run in S-Plus. The code is available from the author's website, [www.sussex.ac.uk/Users/danw/s-plus/Lordcode.htm](http://www.sussex.ac.uk/Users/danw/s-plus/Lordcode.htm). The focus is on the estimate of the group effect ( $\beta_1$ ) for the  $t$  test and ANCOVA.  $pre_i$  and  $post_i$  are related to some true ability (sometimes with measurement error) and  $group_i$  can be related to  $pre_i$  and the ability. As the measurement error increases, this will produce smaller correlations between the  $pre_i$  and  $post_i$  scores. For all the simulations reported there are 100 cases in each study or replication.

#### When the baseline score determines group allocation

In some applied situations an intervention is designed specifically for a subset of the population. For example, some interventions are designed for children who perform poorly. Suppose one such intervention was designed. The researcher gave a pre-test to pupils and assigned the lower 50% of children the SI group and the remaining children to a control condition. The assignment is clearly not random, but based on the pre-test. It is likely that the pre-test measures some ability, but also includes measurement error:

$$pre_i = ability_i + ME\ error1_i$$

For these simulations,  $ability_i$  and both  $error_i$  terms are normally distributed variables with means of 0 and standard deviations of 1.  $ME$  determines the amount of measurement error and is  $ME = 0, 0.5, 1$  or  $2$ . The effect size ( $EF$ ) is also varied:

$$post_i = ability_i + EF\ group_i + ME\ error2_i,$$

where  $EF = 0, .5, 1$  or  $2$ .  $Group_i$  is 1 for the experimental group and 0 for the control group. Measurement error at the two times is independent, but of the same magnitude. Monte Carlo methods are not necessary for this study, as the equations are fairly simple, but they are done to allow comparisons across all simulations and for illustration.

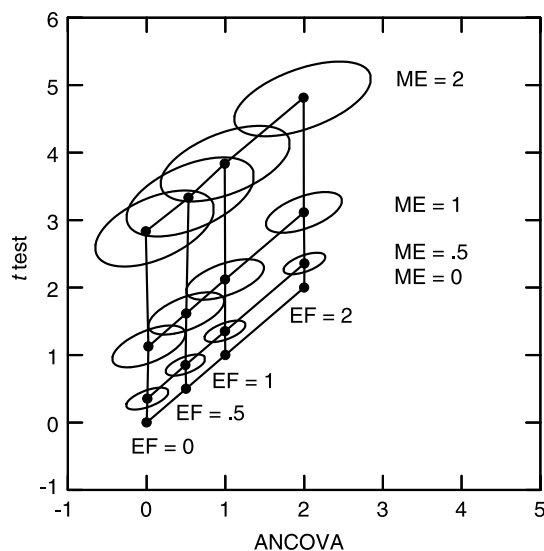
The results are shown in Figure 3. This is a scatterplot comparing the estimates for  $\beta_1$  for both approaches. The condition means are shown as dots. Each estimate is very precise (all  $SE < 0.03$ ) because it is based on 1,000 replications. Lines are drawn between the condition means to show the relationships between the experimental variables (effect size,  $EF$ , and measurement error,  $ME$ ) and the outcome variables (the estimates of  $\beta_1$ ). If the estimates were unbiased estimates of the effect size, then they should increase with the effect size. If the ANCOVA produces unbiased estimates there should be vertical lines for the different measurement errors all at the appropriate point on the  $x$ -axis for each effect size. If the  $t$  test produces unbiased estimates there

should be horizontal lines for the measurement error at the appropriate level on the  $y$ -axis. The ellipses show the spread of the 1,000 replications. Each ellipse includes approximately 50% of the replications, assuming bivariate normal distributions of the estimates. The importance about the shape of the ellipses is discussed below.

For the data in Figure 3, the size of the ellipses increases with measurement error as expected. The most important feature of the figure is that the lines for the different effect sizes are vertical at the true effect size. The ANCOVAs provide unbiased estimates (each estimate is within 2  $SE$  of the expected value). When there is no measurement error the  $t$  test also provides an unbiased estimate. However, as the measurement error increases the  $t$  test produces biased results. With 100 participants,  $t$  values over approximately 2 are statistically significant. The size of this bias is large. Thus, if the sampling is based completely on the initial scores and these are measured with error, the  $t$  test does not provide an adequate test. The  $t$  test will often show that the treatment was effective when it was not. Similarly, if the treatment was aimed and applied to the group that originally performed better, then it would have appeared that the treatment was detrimental.

Because many of the estimates for the  $t$  test are biased, calculating power by the proportion of times the null hypothesis is rejected for a given effect size would not be appropriate. Instead, the size of the interquartile range (IQR) of the estimates is used. If the IQR is small this means that the variability of the estimates is small. In general, it is best to have unbiased estimates with small interquartile ranges. IQR does not take into account the error residual. An alternative is to standardize the IQR of the mean by dividing by the standard error within group. The focus here is on any bias in the estimate so the actual IQR is more appropriate.

The sizes of the IQRs in this simulation are not affected by the effect size manipulation, but they are affected by the amount of measurement error. For the  $t$  test the IQRs for 0, .5, 1 and 2 levels of measurement error are 0, 0.18, 0.36 and 0.66. For the ANCOVA they are 0, 0.30, 0.56 and 0.98. Thus, the variation in coefficient estimates is



**Figure 3.** Results for the first simulation where group allocation is based on initial scores. The ellipses show approximately 50% of the data. The results show that the ANCOVA approach provides unbiased estimates.

greater for the ANCOVA. This can be seen in Figure 3 by the shape of the ellipses, they are wider than they are tall, which means that there is more variability in the ANCOVA estimates than in the  $t$  test estimates.

Returning to Rubin's model and Wainer's (1991) analysis, measurement error was involved in the initial scores. Because these are used for group allocation, some people with above-average ability will be allocated to the below-average group because of measurement error, and vice versa. Because the error term for the post measurement is independent of the error term for the pre-score, the expectation is that if there are no other effects the post-scores for those in the lower group will go up, and those in the higher group will go down. This produces regression towards the mean (Galton, 1886; Morton & Torgerson, 2003) and is consistent with Maris' (1998) findings.

Because regression towards the mean implies the two groups should move in opposite directions if there were no effects, it is reasonable to ask what occurs if the slopes of the regression lines for the two groups are allowed to differ. When the ANCOVA model with the interaction term was estimated, the interaction term was centred on zero. Thus, it makes no difference.

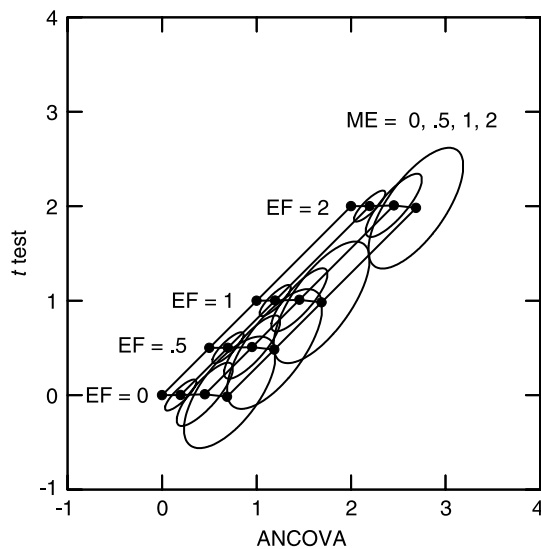
**When the baseline score is associated with, but does not determine, group allocation**

Often the baseline condition is associated with group allocation but does not determine it. Suppose that students are allocated to groups on the basis of 'ability' rather than the baseline score. Let the probability of being in the control group be Bernoulli distributed with the probability of  $1/(1 + \exp(-ability_i))$ . Bernoulli distributed means that the outcome can be one of two possibilities, like a flip of a coin. However, while it is usually assumed that a coin has a 50-50 chance of landing heads, the 'coin' here is weighted so that it has a probability, based on the variable  $ability_i$ , to 'land' in the control group. The above simulations were repeated with just this one difference in group allocation.

The results are in Figure 4. Once again, as measurement error increases, the size of the ellipses increases. The remaining results are very different from Figure 3. Now, the lines for different amounts of measurement error are horizontal and at the appropriate height on the  $y$ -axis for effect size. This means that the  $t$  test provides unbiased estimates (all means within  $2 SE$  of predicted values) and the ANCOVA provides biased estimates. The bias increases as the amount of measurement error increases. The lines are closer together than those in Figure 3. The bias is not large and often would not produce a significant effect when none was present. The parallel slope assumption was not what caused this bias. The simulation was repeated allowing the groups to have different slopes and the estimated parameter for the interaction was centred on zero (condition means within 0.02 of zero), thus showing that it makes little or no difference.

The sizes of the IQRs show the opposite pattern as in the first simulation, though the difference is only noticeable when the measurement error is large. As can be seen in Figure 4, the large ellipses (where the measurement error is 2) are taller than they are wide: the IQRs are bigger for the  $t$  test. The IQRs for effect sizes of 0, .5, 1 and 2 for the  $t$  test are: 0, .20, .37 and .71. For the ANCOVA they are: 0, .20, .35 and .57.

Returning to Figure 2 and Wainer's (1991) analysis, it is expected the pre-scores should not systematically increase or decrease for either group because the predicted values without any intervention would still be ability plus a random error centred on zero. Therefore, Assumption 1 is valid and the  $t$  test provides an unbiased estimate of the gain caused by the intervention. However, it is less powerful as the IQRs are larger and the error sum of squares will be larger for the  $t$  test than for the ANCOVA.



**Figure 4.** Results for the second simulation, when group allocation is based on ability. The  $t$  test provides unbiased estimates.

For completeness it is worth describing what occurs for the  $t$  test and ANCOVA if effect size and measurement error are varied but random assignment is used. The result is that the estimates for the group effect are centred on the diagonal and provide unbiased estimates of the effect size for both approaches. As measurement error increases the variability around the condition mean (i.e. the size of the ellipse) also increases but the estimates remain unbiased. In this situation the ANCOVA is the more powerful test.

In the simulations discussed so far either the  $t$  test or the ANCOVA or both approaches gave unbiased estimates. If this were always the case, then researchers could use both approaches and if they provided the same result then this result would be unbiased. However, this is not the case. In fact, a likely scenario is one that combines aspects of the first two simulations. People are assigned to a condition based on measures of ability and these measures are based in part on test scores that have the same measurement biases as the pre-test and some random variation. Let the probability of being in the control group be  $1/(1 + \exp(-ability_i + .5 pre_i))$  with Bernoulli error. If  $pre_i = ability_i + error1_i$  and  $post_i = ability_i + error2_i$ , then on the basis of 10,000 replications, both a  $t$  test and an ANCOVA produce a mean estimate for the group effect to be about 0.45. However, both are in error as there is no group effect. If a researcher found these effects, on the basis of the pre- and post-scores alone they would not be able to judge whether or not there was an effect.

#### **When the size of the effect is related to ability**

In the example that has been used throughout these simulations it is assumed that the SI was designed for students not performing well. This suggests that the benefit of the instruction may depend on the person's ability level. If random allocation was used this would be shown by an interaction between group and the pre-scores. Three sets of simulations are conducted where the effect size is allowed to vary with ability. In the first set (labelled A in Table 2), group allocation is based on the pre-scores (as in Figure 3), in the second set (labelled B), group allocation is random and in the third set

(labelled C), group allocation is based on ability (as in Figure 4). The error variables used in the simulations have standard deviations of 2 and the effect size is either 0 or 1. The equation to create the data sets is:

$$post_i = ability_i + (0, 1)group_i + (-2, -1, 0, 1, 2)ability_i group_i + error2_i$$

The models fit are:

$$t \text{ test: } post_i = pre_i + \beta_1 group_i + \beta_0 + e_i$$

$$ANCOVA \ 1: \ post_i = \beta_2 pre_i + \beta_1 group_i + \beta_0 + e_i$$

$$ANCOVA \ 2: \ post_i = \beta_3 pre_i group_i + \beta_2 pre_i + \beta_1 group_i + \beta_0 + e_i$$

**Table 2.** The means for the third simulation for the parameters for a t test, ANCOVA without an interaction term, and ANCOVA with an interaction term. Three allocation mechanisms are used (A – based on pre-scores, B – random and C – based on ability). The effect sizes vary with ability

	ef	Slope	t test		ANCOVA 1			ANCOVA 2			
			Intercept	Group	Intercept	Group	Pre	Intercept	Group	Pre	Interaction
A	0	-2	-.57	2.25	.55	.01	.00	-.02	.01	.51	-1.01
		-1	-.55	1.68	.29	-.02	.24	.01	-.01	.49	-.49
		0	-.55	1.11	.01	-.02	.49	.02	-.03	.49	.01
		1	-.57	.56	-.29	.01	.75	.00	.01	.49	.53
		2	-.56	-.02	-.56	-.01	1.00	.01	-.02	.50	1.01
	1	-2	-.56	3.25	.56	1.01	.00	-.01	1.02	.51	-1.01
		-1	-.56	2.67	.28	.99	.25	.00	.99	.50	-.51
		0	-.54	2.10	.02	.98	.50	.02	.98	.50	.00
		1	-.55	1.54	-.27	.99	.75	.02	1.00	.49	.52
		2	-.56	.99	-.56	1.00	1.00	.00	.99	.50	1.00
B	0	-2	-.01	.00	-.01	.00	.00	-.01	.00	.49	-.99
		-1	.01	.00	.00	.00	.25	.00	.00	.50	-.49
		0	.00	.01	.00	.00	.50	.00	.00	.49	.01
		1	.00	.00	.00	.00	.75	.00	.00	.50	.50
		2	.00	-.01	.01	-.01	1.00	.00	.00	.50	.99
	1	-2	-.01	1.01	.00	1.01	.00	-.01	1.01	.50	-.99
		-1	.00	1.01	-.01	1.01	.25	-.01	1.01	.50	-.50
		0	.00	1.00	.00	1.00	.50	.00	1.00	.50	.00
		1	.00	1.00	.00	1.00	.75	.00	1.00	.50	.50
		2	.01	.99	.01	.99	.99	.01	.99	.51	.97
C	0	-2	.00	-.83	-.41	-.01	.01	-.23	-.01	.45	-.90
		-1	.00	-.40	-.32	.23	.23	-.22	.23	.45	-.45
		0	.01	-.01	-.22	.45	.45	-.22	.45	.45	.00
		1	-.01	.42	-.14	.69	.68	-.23	.69	.45	.45
		2	.01	.81	-.03	.88	.91	-.22	.89	.45	.91
	1	-2	-.01	.17	-.42	1.00	.00	-.23	1.00	.45	-.91
		-1	.00	.58	-.33	1.23	.23	-.23	1.23	.45	-.45
		0	-.01	1.02	-.23	1.47	.45	-.23	1.47	.45	-.01
		1	.01	1.40	-.12	1.66	.68	-.22	1.66	.46	.45
		2	-.01	1.85	-.04	1.92	.91	-.23	1.92	.45	.91

The design is a 3 (group allocation method)  $\times$  2 (effect size)  $\times$  5 (slope differential). One thousand replications for each of the 30 conditions were estimated for the three models. The means are shown in Table 2. For method A, where group allocation is based on pre-scores, the  $t$  test produces biased estimates for the group effect, but the ANCOVA, both with and without the interaction term, produces unbiased estimates for the group effect. For method B, where group allocation is random, all the approaches produce unbiased estimates for the group effect regardless of the slopes being different. The most interesting results are for method C, where group allocation is based on ability. When the slopes are the same the  $t$  test produces unbiased estimates of the group effect (the same as in Figure 4). When the slopes are allowed to differ the  $t$  test produces biased results. The ANCOVA produces biased results both when parallel regression lines are assumed and when this is not assumed. Thus, both approaches provide biased estimates in this situation.

A caveat is necessary for this simulation. The outcome measure that is being examined is the average gain for some treatment. If this treatment depends on existing ability, and if this is expected, then researchers are unlikely to be specifically interested in average gain. However, in practice, many will still begin by looking at average gain.

## Discussion

Lord (1967) presented a paradox where two statisticians, each testing a very similar model, came up with different answers. Lord created his data so that they produced this effect. When lecturers describe Lord's paradox to students, the norm is also to create the data (like those in Table 1) to illustrate the paradox, but data that illustrate the paradox also occur with real data (Wainer & Brown, 2004). An appealing aspect of Lord's paradox is that it occurs with a simple design. With more than two measurements other statistical issues come into play, and some form of MANOVA is often appropriate (Maxwell & Delaney, 2004).

The simulations reported here show the situations where it is likely that the different approaches will lead to different conclusions. Differences between the approaches should not occur when there is random allocation, but often random allocation is impractical. Often there are differences in the initial scores for different groups, particularly within an educational context. These are the situations where researchers should consider whether Lord's paradox may occur. Figure 3 shows that if the allocation is based on the initial score, the ANCOVA approach produces unbiased estimates for the effectiveness of the treatment. However, if the allocation is only associated with the initial scores, the choice of statistical method is more complex. If the initial and final scores are related to some ability but are otherwise unrelated, and this ability is used for group allocation, and the effect of the treatment is the same across all levels of ability, then the  $t$  test is appropriate. However, this is a fairly unlikely situation. When these assumptions are not met, both approaches gave biased estimates.

The simplest advice is to use random allocation. It requires fewer assumptions to make causal inference and both the  $t$  test and ANCOVA generally produce good estimates. As Cook and Campbell (1979, p. 5) put it: 'random assignment is the great *ceteris paribus* that is, other things being equal - of causal inference'. However, in much research this is not possible and methods for non-equivalent groups are necessary. It is always important to graph the data, to be very clear about which research questions are being asked, and which assumptions are being made. The ANCOVA is appropriate more often than a  $t$  test on the differences, so should be used more often. The  $t$  test approach

is preferred when the interest is more in the amount of gain in either of the conditions, rather than explicitly on comparing why there may be differences between the effects.

What if you are not sure which, if either, approach is preferred? One of the goals of this paper is to allow readers to make an informed choice about which approach to use. However, sometimes researchers will be interested in both types of questions. If so, it is best to conduct both types of analysis and to report both. Statistics is about discovering and communicating patterns in the data (Wright, 2006a), not about restricting your search to some specific hypotheses. This is not an invitation to go fishing for a significant finding, quite the opposite. If you find different results from the different methods you should describe this difference, and be more cautious in your interpretation.

## References

- Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, *35*, 73–80.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, *15*, 246–263.
- Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society: A*, *157*, 317–356.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of American Statistical Association*, *81*, 945–960.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *72*, 304–305.
- Maris, E. (1998). Covariance adjustment versus gain scores – revisited. *Psychological Methods*, *3*, 309–327.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analysing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.
- Morton, V., & Torgerson, D. J. (2003). Effect of regression to the mean on decision making in health care. *British Medical Journal*, *326*, 1083–1084.
- Oakes, J. M., & Feldman, H. A. (2001). Statistical power for non-equivalent pretest–posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review*, *25*, 3–28.
- Rutherford, A. (2000). *Introducing ANOVA and ANCOVA: A GLM approach*. London: Sage.
- Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, *109*, 147–151.
- Wainer, H., & Brown, L. M. (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *American Statistician*, *58*, 117–123.
- Wright, D. B. (1997). Football standings and measurement levels. *Journal of the Royal Statistical Society: D*, *46*, 105–110.
- Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73*, 123–136.
- Wright, D. B. (2006a). The art of statistics: A survey of modern statistics. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 879–901). Mahwah, NJ: Erlbaum.
- Wright, D. B. (2006b). Causal and associative hypotheses in psychology: Examples from eyewitness testimony research. *Psychology, Public Policy, and Law*, *12*, 190–213.