the british psychological society
promoting excellence in psychology

British Journal of Educational Psychology (2020), 90, 35–49

www.wileyonlinelibrary.com

Allocation to groups: Examples of Lord's paradox

Daniel B. Wright*
University of Nevada, Las Vegas, Nevada, USA

Background. Educational and developmental psychologists often examine how groups change over time. Two analytic procedures – analysis of covariance (ANCOVA) and the gain score model – each seem well suited for the simplest situation, with just two groups and two time points. They can produce different results, what is known as Lord's paradox.

Aims. Several factors should influence a researcher's analytic choice. This includes whether the score from the initial time influences how people are assigned to groups. Examples are shown, which will help to explain this to researchers and students, and are of educational relevance. It is shown that a common method used to measure school effectiveness is biased against schools that serve students from groups that are historically poor performing.

Methods and results. The examples come from sports and measuring educational effectiveness (e.g., for teachers or schools). A simulation study shows that if the covariate influences group allocation, the ANCOVA is preferred, but otherwise, the gain score model may be appropriate. Regression towards the mean is used to account for these findings.

Conclusions. Analysts should consider the relationship between the covariate and group allocation when deciding upon their analytic method. Because the influence of the covariate on group allocation may be complex, the appropriate method may be complex. Because the influence of the covariate on group allocation may be unknown, the choice of method may require several assumptions.

A common situation in educational and developmental psychology is measuring multiple groups of people at multiple time points with the goal of trying to understand how these groups differ over time. This is a complex situation, and not surprisingly, there are numerous analytic choices. But even its simplest form, two groups at two time points, presents a difficult statistical choice. Lord (1967, 1969) examined two plausible methods of analysis. He used the context of university student weight by gender, with the time points being the beginning and ending of the academic year. The first method is subtracting the initial weight from the prior weight (i.e., how much weight the student lost or gained) and doing a *t*-test on these differences.[1] This is called the gain score approach.

*Correspondence should be addressed to Daniel Wright, Department of Educational Psychology and Higher Education. University of Nevada, Las Vegas. 4505 S. Maryland Parkway, Box #453001, Las Vegas, NV 89154 (email: daniel.wright@unlv.edu or dbrookswr@gmail.com).

[1] An alternative that allows this to be extended to more complex problems is to have a 2 × 2 mixed ANOVA. The interaction between time of the test and group is the same effect as tested by this simpler gain score t-test. The simpler test will be used here.

DOI:10.1111/bjep.12300

An alternative is conducting an analysis of covariance (ANCOVA) on the final weights using the prior weights as a covariate. Because these can lead to different conclusions, Lord called it a paradox, and it fits within a class of paradoxes applicable when comparing relationships among three variables (Pearl, 2014; Wainer & Brown, 2007).

In this section, the gain score and ANCOVA procedures will be briefly described, as well as some extensions to them. The case considered is where there are two groups and two time points. This is for explanatory ease. Both the number of groups and the number of time points can be increased. Increasing these would increase the complexity of the models, but not the fundamental issue considered here.

The gain score model involves subtracting the time 1 scores (following Rubin 1977 called PreTest_i) from the time 2 scores (called PostTest_i), and then performing analyses on these gain scores with a *t*-test. This tests the null hypothesis of $\beta_1 = 0$ in the following equation:

$$\text{gain}_i = \text{PostTest}_i - \text{PreTest}_i = \beta_0 + \beta_1 \text{ group}_i + e_i. \quad (1)$$

This can be re-written as a regression predicting PostTest_i:

$$\text{PostTest}_i = \beta_0 + \beta_1 \text{ group}_i + 1 \cdot \text{PreTest}_i + e_i. \quad (2)$$

Two important assumptions for the gain score model are that PostTest_i and PretTest_i are on the same scale so that $\text{PostTest}_i - \text{PreTest}_i$ makes sense and that this difference has the same meaning for all values of PreTest_i. While in some cases it may be plausible to transform the data to meet these assumptions, sometimes this is not plausible.

The second procedure is ANCOVA:

$$\text{PostTest}_i = \beta_0 + \beta_1 \text{ group}_i + \beta_2 \text{ PreTest}_i + e_i. \quad (3)$$

Sometimes an interaction is included; sometimes it is not included. Here, it will not be included and β_1 therefore estimates an average group effect after conditioning on PreTest_i. The difference between Equations 2 and 3 is that for the gain score approach β_2 is fixed at 1 and for the ANCOVA it is estimated.

Here, I refer to 'gain score' for any model where the variable predicted is the difference between the final scores and previous scores. The word 'ANCOVA' is used for several different types of models (Cox & McCullagh, 1982). Here, it is used for any model where the previous scores are conditioned upon. The conditioning might be linear like Equation 3 or a complex function, like the monotonic splines that are part of the Student Growth Percentile (SGP) models (e.g., Betebenner, 2009) used in many US states to measure student growth (for more discussion of this approach see, e.g., Lockwood & Castellano, 2015; Wright, 2018). Here, linearity is assumed. A related model is where covariates are used to match people with similar scores on a set of covariates. One approach to this is called propensity matching (Rosenbaum, 2002). Matching is related to ANCOVA and is discussed later in this paper.

There are several other extensions to the basic gain score and ANCOVA models. One particularly relevant extension for education is that both of these approaches can be used to describe multilevel models where the students are nested within classrooms and schools (e.g., Aitkin & Longford, 1986; Goldstein, 2014). Later in this paper, multilevel versions of gain score and ANCOVA models are used for measuring school effectiveness. This is of much importance in education as many school systems have implemented accountability measures (Muller, 2018). For a historical overview of this approach in the

United Kingdom, see Leckie and Goldstein (2017). The method used in this example is similar to those used in many US states (Amrein-Beardsley, 2014). Therefore, it is worth briefly describing the multilevel versions of the gain score and ANCOVA approaches.

Let *j* refer to the different schools and *i* to the different students. A simple multilevel gain score model would be:

$$\text{PostTest}_{ij} = \beta_0 + 1 \cdot \text{PreTest}_{ij} + u_j + e_{ij}, \tag{4}$$

where the u_j allows for variation around the intercept β_0. The corresponding multilevel ANCOVA is:

$$\text{PostTest}_{ij} = \beta_{0j} + \beta_1 \text{PreTest}_{ij} + u_j + e_{ij}. \tag{5}$$

An estimate for u_j for each school can be calculated from these models. One approach is to estimate the conditional modes. The conditional mode for the *j*th school is the most likely value, given the model, for this school's effectiveness. These are sometimes called *school residuals* (e.g., Goldstein *et al*., 1993, p. 428). This approach is sometimes called the *value-added model* or VAM, though the name is perhaps presumptuous given the debate about what these procedures measure (e.g., American Statistical Association, 2014; Goldstein, 1991; Wright, 2017). The complexity of this procedure has lead to many policy makers to accept over-simplistic explanations and to believe that it accurately measures value-added (see Braun, 2013).

Regression towards the mean

There was a problem with Darwin's theory of evolution. Consider human height. If parents had children whose height naturally varied around the parents' height, and these children had their own children and their height varied in the same way, the variance in a population would increase every generation. Galton collected data showing this did not happen. The variance remained fairly constant across generations, and Galton realized that this was a problem for Darwin's theory. Two very tall parents' offspring will likely be taller than the population average but, on average, not as tall as themselves. This observation has become known as regression towards the mean (RTM). Stigler (2016, Ch. 5) describes how Galton went from proposing a clever (but unnecessary and incorrect) biological mechanism to account for the inter-generational homogeneity of variance, to realizing RTM occurred more generally, and therefore, its cause was a statistical artefact.

RTM can be explained using the central equation of psychometrics:

$$\text{Observed Score} = \text{True Score} + \text{Error}. \tag{6}$$

Assume the True Score and Error are independent. It is expected that those with high Observed Scores are likely to have above average True Scores and are likely to have above average Errors. If a second score is observed, and nothing has been done to change the True Score, it is expected that this True Score will still be above average, but the expected value of Error is zero. The expected value of the observed value will therefore be centred on the True Score, which will tend to be between the original observed value and the population mean. Therefore, the expected value will regress from the original observed value towards the mean of the group.

Following Efron and Morris (1977), a baseball example will be used to illustrate how RTM can affect the estimates. Baseball is a good source of data to illustrate statistical concepts because there is much information on each individual player (Marchi & Albert, 2013). Baseball is popular in North and Central America, Japan, and Korea, and its popularity is growing elsewhere. Further, the confrontation between the pitcher and batter, relevant to the current example, is similar to the one between the bowler and batter in cricket. Enough information is provided in this paper so that the example should be clear to readers not familiar with baseball or cricket. The example is about batting averages. A baseball batting average is, roughly, the proportion of times that the player successfully hits the ball and reaches a base safely (the calculation is slightly more complicated), so high scores are better. Averages are reported to three decimal points, for example, .247 for 24.7% successful. Efron and Morris (1977) took 18 players' batting averages from the first few games of the 1970 season and predicted how well they would hit for the remainder of the season. They showed that the players' averages tended to regress towards the mean of the 18 players. Here, we will consider RMT when there are two groups.

Baseball players are either pitchers or position players (occasionally, a player is both – what is called an all-rounder in cricket – but this is rare enough in baseball to ignore for current purposes). The pitchers are on the team for their defensive (non-batting) skills and therefore tend to have low batting averages: around .150, meaning they are successful about 15% of the time. The position players are chosen because they are good hitters. Their average is about .260, meaning they succeed about 26% of the time. RTM works within these groups, so a pitcher or position player hitting far above or below their groups' average in the first half of the season is likely to regress towards their group mean in the second half of the season. Data from the 2016 season (accessed 30 October 2016, from mlb.mlb.com and the R data file at https://github.com/dbrookswr/VAM-Work/total.Rdata) are shown in Figure 1. The gain scores tended to be lower for players who had high first half averages that those who had low first half averages (the slopes of the regression lines are negative).

Lord's first statistician would calculate the gain score for each player and then depending on assumptions might conduct a *t*-test. Welch's version of the *t*-test is used here because the standard deviation for pitchers is higher than for position players because pitchers' averages are based on fewer at-bats (other methods could also be used to account for this heteroscedasticity). The mean gain score for the pitcher is: mean gain = .023 and the position players is: mean loss = −.005. The result is as follows: $t(65.493) = 1.95$, $p = .055$. Depending on the α level and if a one- or two-tailed test is being used (two-tailed is used here), the analyst may declare this is a non-significant result (btw, the *p*-value if using Student's rather than Welch's method is: $p = .013$).

Lord's second statistician would conduct an ANCOVA on the second half average, conditioning on the first half average, and look at the coefficient for whether the player is a pitcher or a position player. The finding is a detriment for pitchers of: −.064, or one hit about every 15 at-bats. The residuals are more dispersed for pitchers than the position players so robust standard errors were calculated. The HC1 standard errors from the *estimatr* package (Blair, Cooper, Coppock, Humphreys, & Sonnet, 2019), based on the procedure described by MacKinnon and White (1985), were used. The difference is statistically significant: $t(173) = -3.95$, $p < .001$. Conditioning on first half averages, position players hit higher than pitchers for their second half averages. Sports researchers could probably come up with clever reasons why this may occur, but this difference can be accounted for by regression towards the group mean so requires no further explanation beyond this statistical artefact. The position players are regressing towards a higher mean than the pitchers are regressing towards (see also Wainer & Brown, 2007).

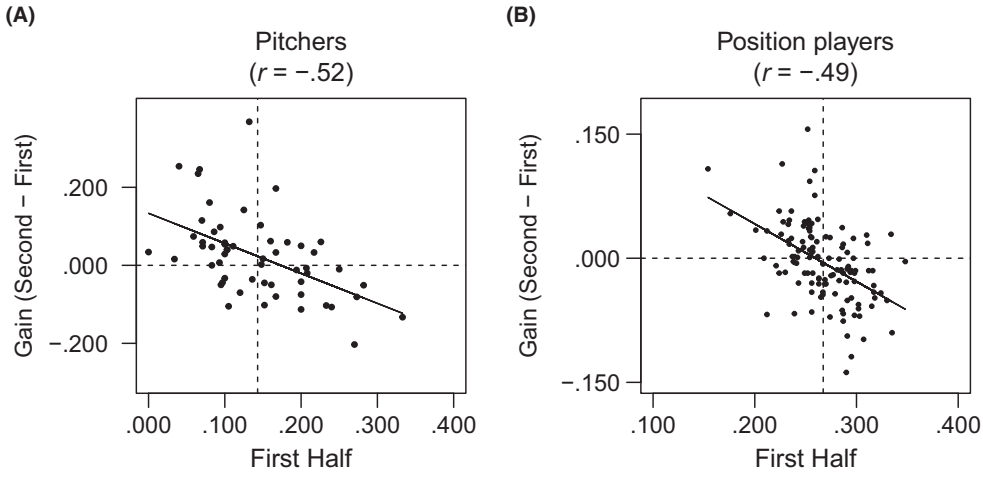

Figure 1. Scatter plots showing, both for pitchers (Panel A) and for position players (Panel B), that those who hit above their group means in the first half (to the right of the vertical dashed lines) tended to have lower second half averages (below the horizontal dashed lines). The opposite is true for those initially hitting below their group mean. The regression lines are in solid.

Wainer (2007) called this regressing towards different means Kelley's paradox, for the influential psychometrician Truman Kelley.[2]

What happens for matching procedures? Methods like propensity matching (Rosenbaum, 2002) are popular. They allow researchers to match on one set of variables and then compare some outcomes by other variables. Suppose pitchers and position players are matched on first half averages for the 2016 season. One pairing in this data set could be pitcher Gerrit Cole, who hit .208 in the first half, and position player Yasmani Grandal, who hit .212 in the first half. This average is high for a pitcher, but low for a position player. Therefore, the prediction is that Cole's average will decrease in the second half, and Grandal's average will increase in the second half. Two players with similar first half averages are expected to have different second half averages if they are from different groups and nothing else is known about them. Consistent with these predictions, in the second half Cole's average decreased to .188 and Grandal's increased to .245. Matching works similar to ANCOVA in this context.

Consider a different sporting example where the appropriate statistical procedure is different. The US magazine *Sports Illustrated* (www.si.com) puts an athlete who performs really well in recent events on its cover. Here, cover athletes are considered a group. There is something called the *Sports Illustrated jinx* where the cover athletes tend to perform worse after being on the cover than before, compared with other athletes (see en.wikipedia.org/wiki/Sports_Illustrated_cover_jinx, accessed 1 December 2018). Here, the gain score approach is inappropriate because the athletes are selected to be on the

[2] *They used the word* paradox *because of the context in which they were studying this. There was the belief that if universities enrolled students with good test scores from schools with low-average scores, that when removed from these low-average environments that the students would excel compared with students with similar scores from schools with high averages. When they looked at the data, they found the opposite occurred; students regressed towards the mean of the group, and for some, this seemed paradoxical. While reasons exists for why these groups may go up or down in comparison with the other (and likely in some contexts these occur), it is important first to account for this regression artefact.*

cover based on their performance. The expectation is that these athletes will regress a bit towards the level of all athletes. The key difference between these examples, discussed more in the next section, is whether the covariate (first half average or performance in recent events) influences which group (pitchers or cover athletes) a person is in. If group membership is not influenced by the covariate, the tendency is to regress towards that group's mean.

It is worth making clear that statistical artefacts, like RTM, do not preclude other effects. But before speculating on additional or alternative explanations, it is worth ruling out statistical artefacts. For example, in education there is much discussion about Matthew effects (e.g., Stanovich, 1986) or the rich get richer. Evidence for this is increased variance over time (so the type of increase Galton did not observe with inter-generational human height). However, if a researcher used an ANCOVA to predict wealth at time 2, conditioning on wealth at time 1, to show that groups that tend to be rich (higher time 1 values) get even richer at time 2, at least part of this effect could be accounted for by RMT and therefore should be considered prior to hypothesizing some additional mechanism.

Lord's paradox and group allocation

Since Lord (1967) dangled this apparent paradox in front of researchers, numerous authors have taken up the challenge to explain this paradox. Commentators agree that the two analytic approaches are both accurate descriptions of the data (and other possibilities exist), but address different research questions (e.g., Hand, 1994; Wainer, 1991). The gain score approach is 'an unconditional comparison between the gains of the two groups' (Hand, 1994, p. 324) while the ANCOVA 'is a test of an average conditional comparisons, conditioning on initial weight' (Hand, 1994, p. 324). Thus, either can be an accurate description, but of different quantities. Wainer (1991) asks how this relates to making causal inference about the group. This is the focus here. If trying to decide whether some variable is causally efficacious in the absence of random allocation, it may be that only one approach will provide unbiased estimates (and it may not be either of these). The remainder of this section will focus on approaches that show which approach is more suitable depending on the causal relationship between the initial score (weight at the beginning of the year in Lord's case) and the grouping variable (gender in Lord's case).

Figure 2 shows graphical models of two situations where both gain score and ANCOVA procedures might be considered.[3] The figure is adapted from Pearl (2016), and the situations correspond to the examples in §7 of Rubin (1977). Using Rubin's variable names, there are PreTest and PostTest scores and a variable for whether the student was in a computer-aided learning program or the regular program. Panel A corresponds to Lord's (1967) original formulation where the group variable is exogenous to the system. In Lord's original formulation, the weights did not affect gender (similar to how in the baseball example first half averages do not [usually] affect a player's position). In Panel B, the PreTest scores influence which group the student was in. Rubin states that just looking at the data (in his Table 1) it is not possible to know which of these panels is the appropriate causal model and therefore which statistical procedure is appropriate. He describes how assumptions are often necessary to make causal inference.

[3] *There is a specific mathematical sense of the word 'graph' as a set of nodes, some of which are connected by edges. This is an important area of mathematics and for making causal inference in science (Pearl, 2009).*

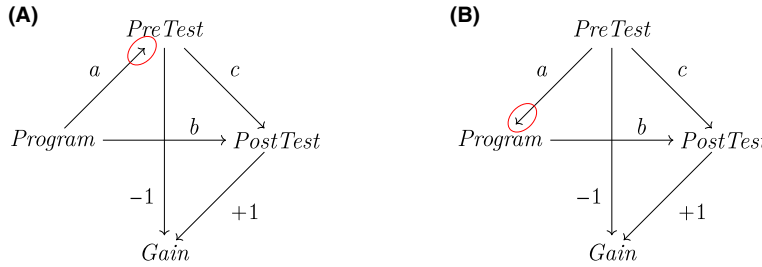

Figure 2. Panel (A) shows the graph based on Lord's original formulation, with *Program* influencing PreTest and PostTest. In Panel (B), the direction of causality is now from PreTest to *Program*. These are based on figures 2b and 5 of Pearl (2016) and table 1 of Rubin (1977).

Table 1. The mean effectiveness values for data created for Figure 3A or B, and whether there was no difference or a small difference between the effectiveness of the two categories of schools (high = historically high performing; low = historically low performing)

		Figure 3A		Figure 3B	
Statistical model	Group	No diff	.2σ diff	No diff	.2σ diff
VAM (multilevel ANCOVA)	High	0.24	2.37	0.00	1.99
$\text{post}_{ij} = \beta_0 + u_j + \beta_1\ \text{pre}_{ij} + e_{ij}$	Low	−0.24	−2.37	−0.00	−1.99
	Diff	0.47	4.74	0.00	3.98
Multilevel gain score	High	−0.01	4.12	−6.11	−1.53
$\text{post}_{ij} = \beta_0 + u_j + \text{pre}_{ij} + e_{ij}$	Low	0.00	−0.01	6.10	5.86
	Diff	−0.01	4.13	−12.21	−7.40

Note. Standard errors are approximately .01.

Several researchers have shown that which of these panels is more appropriate informs whether the gain score or ANCOVA model is more appropriate. Holland and Rubin (1983) use Rubin's potential outcomes model for causation (for review of this model, see Holland, 1986). For Panel A, each approach is applicable if different untestable assumptions are made (Holland & Rubin, 1983, table 1.2). The need to make assumptions about how some aspects of the data arise in order to reach conclusions about other aspects is nicely summarized by Cartwright's motto: 'no causes in, no causes out' (Cartwright, 2014, p. 312). For the gain score model, the assumption is that without any group effects the expected value for the PostTest is the PreTest (Holland & Rubin, 1983, equation 3.7).

In Panel B, group membership is influenced by the initial scores. Holland and Rubin (1983, pp. 21–22) describe this situation in §A.4 of their appendix and show (assuming linearity and parallel slopes for the groups) that the ANCOVA approach yields appropriate estimates. Pearl (2016) uses graphical models and reaches a similar conclusion: in Panel A, both approaches can be correct depending on the research question and assumptions, but for Panel B, 'one [statistician] was right (ANCOVA) and one [statistician] was wrong'. Wright (2006) reached similar conclusions, but using simulation methods. When the group is not influenced by the covariate, and the assumptions for gain score model in Holland and Rubin (1983) hold, the gain score approach provides unbiased estimates and ANCOVA does not. The converse is true when the covariate influences group membership.

An alternative way of conceptualizing this is whether to expect regression towards the group means, as with Panel A of Figure 2, or regression towards the mean for the whole sample, as with Panel B. The distinction is whether group membership is dependent on the covariate. In Panel A, the group exists without the measurement of that covariate. In the baseball example, the players are pitchers or position players without having to have a first half batting average. In Panel A, students opt for the experimental program without being influenced by their marks on the PreTest. However, membership of the group, *Sports Illustrated* cover athletes, is due to their recent performance. The group would not exist without excellent recent performances. Similarly in Panel B of Figure 2, the group exists because of the measurement. In these cases, the group distinction can be thought of as just recoding the covariate. As such, people will tend to regress towards the overall sample mean.

Measuring educational equity

The purpose of this section is to show how the choice of statistical model – using covariance or gain scores – is relevant to a controversial topic in education: measuring school effectiveness. This was chosen because of the its importance in education. It is a complex example and therefore illustrates how Lord's paradox and regression towards the mean can be applied to important real-world problems.

In education, many jurisdictions estimate the effectiveness of schools and teachers using student test scores. These estimates can have serious consequences including school closures, loss of employment for teachers, and have enticed educators to change student test answers (Blinder, 2015). Use of these to measure educational equity is the focus here, but policy makers should be cautious in general using this approach (e.g., American Statistical Association, 2014; Goldstein, 1991; Wright, 2017).

Achievement gaps in education refer the differences between scores for historically low-performing groups (e.g., those from low socio-economic status [SES] groups) and scores for historically high-performing groups. In the United States, Chief State School Officers are required to submit their plans for how they will try to lessen these gaps (see www2.ed.gov/programs/titleiparta/resources.html, accessed 1 December 2018). One reason given for achievement gaps is that some analyses show that historically low-performing groups of students are more likely to be enrolled in less effective schools than other groups of students. This is called the educator equity gap. For example, New Mexico's 2015 report finds minority students and economically disadvantaged students attend schools that the state rates as less effective than other schools (https://www2.ed.gov/programs/titleiparta/equitable/nmequityplan060115.pdf, accessed 1 December 2018). The argument is that if states can ensure equal access to quality schools, then the achievement gaps should shrink. This is a laudable aim, and programmes that provide incentives for good teachers to work in less effective schools are encouraged. The goal here is to explore whether the educator equity gaps are being measured appropriately. The method used is based on the method used in New Mexico for rating schools prior to 2019.

How jurisdictions measure effectiveness varies, but many include measures based on student test scores often using ANCOVA or ANCOVA-like procedures (like the VAM and SGP methods discussed earlier). The statistical question is in which situations does the VAM (Equation 5) or the multilevel gain score model (Equation 4) perform better than the other. The results of Holland and Rubin (1983), Pearl (2016), and Wright (2006) reinforce

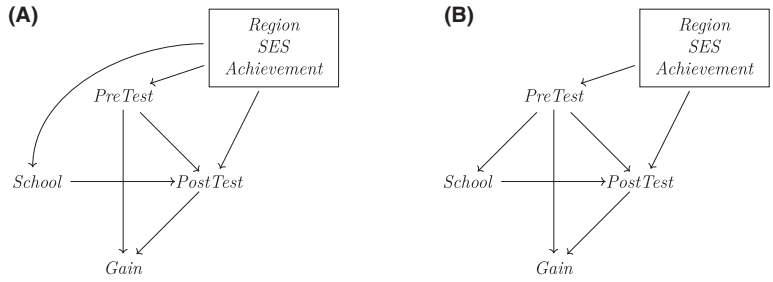

Figure 3. The causal models used to create data for the simulation. The node Region/SES/Achievement stands for these three associated variables. In Panel A, these directly influence which school the student attends (in K–12 situations usually based on region). In Panel B, school attendance is influenced by PreTest (e.g., a university could use ACT or 'A' level scores as part of their admissions process, and this could be conditioned upon in the analyses).

the notion that: 'knowledge of the assignment process is critical to drawing inferences about the effect of the treatments' (Rubin, 1977, p. 22). A simulation will be conducted to show how the way in which students are assigned to schools should affect the choice of statistical method.

Simulation methods

Two different data models are used for the simulation. Figure 3A shows how most US public K–12 education districts operate. There is variability in socio-economic status (SES) by region and achievement gaps associated with these variations. These influence where students usually go to school (though in some locations there is limited parent choice) and the PreTest and PostTest scores. The school impacts the PostTest, but not the PreTest scores (the PreTest being 'pre' the schools' influence). Figure 3B is appropriate if the covariate (here PreTest scores) has a causal influence on which school the student attends. This is appropriate for some universities and some selective primary and secondary schools (e.g., if using ACT scores both for admissions and as a covariate in the model). It is also applicable for teacher evaluation within schools where students may be streamed into a class based on previous performance (Paufler & Amrein-Beardsley, 2014). This is important because sometimes these methods are used to estimate teacher effectiveness (Amrein-Beardsley, 2014).

The simple data models in Figure 3 are used to focus on the importance of knowing how students are assigned to schools. It is important to stress that these are simple models. If the statistical models exhibit problems with these simple data models, it will allow the cause of the problems to be more easily identified than if complex (more realistic) models were used. In applied settings, other variables will impact each of these variables and it is likely that for some schools the covariate causally impacts which school the student attends, but not for others. Appropriate statistical models should take these aspects into account.

An advantage of using simulation methods is that the researcher knows the true values of all variables. Here, the critical variable – the direct effect of schools on students' scores – is unobserved with empirical data, but it is known in simulations. The primary question here is whether these scores differ between groups of schools. This would suggest

educator inequity. With simulations, data sets can be constructed where it is known whether there is a difference between the true effectiveness between schools that serve predominantly historically low-performing groups of students and those serving predominantly historically high-performing groups of students. An additional second advantage of simulations is that the procedure can be repeated thousands of times to provide precise estimates of the procedure's performance.

The simulation has a 2×2 design. The first factor is whether the data are constructed according to Figure 3A or B. The second is whether there is no difference in the true effectiveness of schools serving historically high- and low-performing groups of students or a difference of .2 of the standard deviation in the true variability of school effectiveness. Cohen (1988) calls this a 'small' difference, though in this context it would be considered substantial. Lipsey *et al*. (2012) discuss how to interpret effect sizes in different educational contexts. While some educators search for extremely large effects (e.g., Bloom, 1984, talked about searching for effects ten times larger than this), Chetty *et al*. (2011) show how much smaller effects can produce large outcomes when they persist over time. This factor allows both Type I and Type II errors to be examined. The R code for this simulation is at the end of this document and is available at https://github.com/dbrookswr/VAM-Work/2020LordSimul.r.

For each replication, the sample is divided into two equally sized groups (e.g., these might be those above and below the median on household income, or different ethnicities). There is a latent variable, called *Achieve*, for individual differences among students that influence test scores. The latent variable for both groups is normally distributed, but the mean for the higher performing group is $.2\sigma$ higher (Cohen's *small* effect). The PreTest scores are based on this variable and normally distributed random error. Within each group, *Achieve* and the random error have the same standard deviation. Half the schools are labelled *high* and half labelled *low* for which types of students they tend to serve (historically high- or low-performing groups of students). For Figure 3A, students from the historically high-performing group have an 80% probability of being assigned to an *upper* school and students from the historically low-performing group have an 80% probability of being assigned to a *lower* school. For Figure 3B, students who score above the median on PreTest have an 80% probability of being assigned to a *upper* school and those who score below the median have an 80% probability of being assigned to a *lower* school.

The true school effects are all drawn from normal distributions. For the no difference conditions, the true effectiveness scores for all schools are drawn from a distribution with a mean of zero; there are no systematic differences in effectiveness between the two sets of schools. For the 'small' difference conditions, the *high* and *low* schools' effectiveness scores are drawn from normal distributions, but the mean for *high* schools is $.2\sigma$ greater than the mean for the *low* schools. The PostTest scores are based on *Achieve*, the true school value-added, and random error. These data are created to adhere to some common statistical assumptions so that lack of fit cannot be attributed to, for example, skewness of effectiveness scores. There were 2,000 replications for each condition so the standard errors are small. There are 500 schools each with 100 students.

The VAM (Equation 5) and the multilevel gain score (Equation 4) are used to estimate effectiveness (other statistical models were also in the simulation and are available from the author). The conditional modes are calculated and used to estimate the effectiveness for the individual schools.

Simulation results

Table 1 shows the mean effectiveness scores for the two groups of schools, and the differences, in each of the four conditions of the 2 × 2 design and for the two statistical models. Starting with the data models from Figure 3A with no true differences between groups (first column), the VAM (multilevel ANCOVA) estimates that the schools that predominantly serve historically high-performing students are more effective (Type I errors). Of these estimates, 1,756 of the 2,000 (88%) were in this direction, and of these, 439 (25%) were statistically significant at $\alpha = 5\%$. Only 2 of the 242 (0.83%) estimates in the other direction were statistically significant. The gain score model correctly shows no overall bias. There were 962 (48.1%) trials showing an advantage for schools serving predominantly groups of historically high-performing students and 1,038 (51.9%) in the opposite direction. These are not significantly different from 50% ($p = .094$). In total, 102 of these (5.1%) were statistical significant, which is not statistically significantly different from the nominal level of 5% ($p = .878$). In the next column, where there is a true effect to detect, both models show 100% statistically significant differences in the correct direction.

The final two columns of Table 1 correspond to when students are allocated into the two groups of schools on the basis of the covariate (Figure 3B). The VAM (multilevel ANCOVA) correctly estimates that there is no group bias (column 3). Nine-hundred and ninety-three estimates (49.65%) were in the direction of favouring schools serving predominantly high-performing students and 1,007 (50.35%) in the opposite direction. This is not statistically significantly different from 50% ($p = .771$). Overall, 86 of these (4.30%) were statistical significant, which is not statistically significantly different from the nominal level of 5% ($p = .166$). The gain score model estimated statistically significant effects, for all trials, with an advantage for those schools serving mostly historically low-performing groups of students. The fourth column shows when there is an effect. The two models give 100% statistically significant estimates, but in opposite directions. The VAM correctly shows the advantage for schools serving predominantly groups of historically high-performing students. The gain score estimates that those schools are less effective. Finding an effect in the opposite direction of the true effect is what Gelman and Carlin (2014) call a Type S error, for an error in the sign of the effect.

Simulation discussion

The VAM (multilevel ANCOVA) provided biased results when school allocation was based on Figure 3A, when the PreTest did not causally influence which school the student attended. This is because student scores regressed to their group means. This is analogous to how the pitchers and position players each regressed towards their group means in the second half of the baseball season. Because ANCOVA, in some sense, compares students with similar initial scores, this means this statistical artefact will lead the VAM to estimate schools that serve mostly high-performing students to be more effective than those that serve mostly lower-performing students. The schools that serve groups that predominantly historically high achieving will appear more effective even when there are no true differences (column 1 of Table 1). The gain score model performs better in this situation, though caution is urged using the gain score model for this purpose. These data were created to be consistent with the assumptions Holland and Rubin (1983) list for this model. Before implementing a gain score model in any situation, it is important to address whether the scores on the two tests are comparable so that their differences make sense to compare.

The VAM (ANCOVA) procedure performed better than the gain score model when school allocation was based on Figure 3B. The gain score model provides biased estimates here because students regress to the overall sample mean. This is analogous to the *Sports Illustrated jinx* discussed above. The fourth column shows when there is an effect. The ANCOVA correctly shows the effect. The gain score estimates that those schools serving low-performing students do better: the opposite direction of the true effect.

This is an important conclusion for places using these types of models for estimating school (and teacher) effectiveness. Analysts should consider whether the covariates, usually some set of prior test scores, can influence the schools students attend. Wright (2017) argues that those using these methods for high-stakes decisions should construct multiple causal models of how they believe the data may arise, simulate data using these models, and examine the accuracy of their statistical methods. This would allow them to show their assumptions and demonstrate the performance of their analytic method.

The substantive conclusion is that because most K–12 school allocation is not based on the covariate in these models, that the ANCOVA methods (including VAM, SGP, propensity matching, *etc*.) are likely to yield biased estimates and should not be used in these situations. These will often show large educator equity effects (e.g., the New Mexico findings discussed above), but as shown here at least part of these effects are a statistical artefact.

GENERAL DISCUSSION

For over 50 years, Lord's (1967, 1969) paradox has been used to illustrate how the choice of statistical methods is not simple, even when there are only three variables. Part of the reason for this is that when asked what ANCOVA does statisticians sometimes give the misleading short-hand response that 'it controls for the covariates' and that it somehow allows causal inference for the other variables. This advice is given despite methodologists describing the many limitations of the procedure for decades (e.g., Meehl, 1970). The result is that many non-statisticians come to believe that the procedure does more than it actually does. With reference to educational effectiveness, Braun (2013) describes how some policy makers attribute almost magical power to these methods.

Several authors (e.g., Holland & Rubin, 1983; Pearl, 2016; Wright, 2006) discuss that it is important to consider whether the PreTest scores influence the grouping variable. These authors use of mathematical, graphical, and simulation methods, and their texts will appear quite technical to many readers. It is hoped that choosing a sports example and presenting limited mathematical details that the concepts underlying regression towards the mean are clear. The critical consideration is whether group allocation is determined by the covariate or just associated with it. The mechanism to explain the different outcomes is regression to the group mean, if the group membership was not influenced by the initial score.

The examples and simulations describe situations where a single covariate either does or does not have a causal impact on group membership. In some real-world situations, there will be multiple covariates some of which may have a causal influence, some not, and some may have an influence only on group membership for some in the sample. To complicate matters further, the researcher may not know which variables influence group membership and for whom. If the researcher is unsure, several different simulations can be conducted in order to evaluate how sensitive the different statistical methods are to whichever aspects of the model that the researcher is uncertain about. This can create a

very complex situation. It is hoped that this paper helps researchers to take steps towards these types of situations.

Further, it is important to show how the choice of statistical models can have important policy implications. There is much debate about using test scores to evaluate schools and teachers (e.g., Amrein-Beardsley, 2014; Foley & Goldstein, 2012). This is why this particular example was chosen. It is important to consider the causal models underlying the data before deciding how to analyse them. Examining these clearly shows when the statistical method often used is inappropriate.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A*, *149*(1), 1–43. https://doi.org/10.2307/2981882

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment (Tech. Rep.)*. Retrieved from https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf.

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2019). *estimatr: Fast estimators for design-based inference [Computer software manual]*. Retrieved from https://CRAN.R-project.org/package=estimatr (R package version 0.16)

Blinder, A. (2015). *Atlanta educators convicted in school cheating scandal*. Retrieved from https://www.nytimes.com/2015/04/02/us/verdict-reached-in-atlanta-school-testing-trial.html?_r=0

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Braun, H. I. (2013). Value-added modeling and the power of magical thinking. *Ensaio: Evaluation of Public Policies in Education [Brazil]*, *21*, 115–130. https://doi.org/10.1590/S0104-40362013000100007

Cartwright, N. (2014). Causal inference. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of social science* (pp. 308–326). Oxford, UK: Oxford University Press.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from project STAR. *Quarterly Journal of Economics*, *126*, 1593–1660. https://doi.org/10.1093/qje/qjr041

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cox, D. R., & McCullagh, P. (1982). Some aspects of analysis of covariance. *Biometrics*, *38*, 541–561. https://doi.org/10.2307/2530040

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127. https://doi.org/10.1038/scientificamerican0577-119

Foley, B., & Goldstein, H. (2012). *Measuring success: League tables in the public sector*. London, UK: British Academy.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651. https://doi.org/10.1177/1745691614551642

Goldstein, H. (1991). Better ways to compare schools? *Journal of Educational Statistics*, *16*(2), 89–91.

Goldstein, H. (2014). Using league table rankings in public policy formation: Statistical issues. *Annual Review of Statistics and its Application*, *1*, 385–399. https://doi.org/10.1146/annurev-statistics-022513-115615

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education*, *19*, 425–433. https://doi.org/10.1080/0305498930190401

Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*, 317–356. https://doi.org/10.2307/2983526

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960. https://doi.org/10.2307/2289064

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 3–35). Hillsdale, NJ: Erlbaum.

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'progress 8'. *British Educational Research Journal*, *43*, 193–212. https://doi.org/10.1002/berj.3264

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms. (ncser 2013–3000)*. Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncser/

Lockwood, J. R., & Castellano, K. E. (2015). Alternative statistical frameworks for student growth percentile estimation. *Statistics and Public Policy*, *2*(1), 1–9. https://doi.org/10.1080/2330443X.2014.962718

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305. https://doi.org/10.1037/h0025105

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, *72*, 336–337. https://doi.org/10.1037/h0028108

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*, 305–325. https://doi.org/10.1016/0304-4076(85)90158-7

Marchi, M., & Albert, J. (2013). *Analyzing baseball data with R*. Boca Raton, FL: Chapman & Hall/CRC.

Meehl, P. E. (1970). Nuisance variales and the *ex post facto* design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science*, Vol. *IV* (pp. 373–402)., Analysis of theories and methods of physics and psychology Minneapolis, MN: University of Minnesota Press.

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton, NJ: Princeton University Press.

Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Education Research Journal*, *51*, 328–362. https://doi.org/10.3102/0002831213508299

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J. (2014). Understanding Simpson's paradox. *The American Statistician*, *68*, 8–13. https://doi.org/10.1080/00031305.2014.876829

Pearl, J. (2016). Lord's paradox revisited (Oh Lord! Kumbaya!). *Journal of Causal Inference*, *4*(2). https://doi.org/10.1515/jci-2016-0021

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-3692-2

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1–26. https://doi.org/10.3102/10769986002001001

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360–407. https://doi.org/10.1598/RRQ.21.4.1

Stigler, S. M. (2016). *The seven pillars of statistical wisdom*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674970199

Wainer, H. (2000). *Visual revelations: Kelley's paradox. CHANCE*, *13*, 47–48.https://:doi.org/10.1080/09332480.2000.10542192.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, *109*, 147–151. https://doi.org/10.1037/0033-2909.109.1.147

Wainer, H., & Brown, L. M. (2007). Three statistical paradoxes in the interpretation of group differences: Illustrated with Medical School Admission and Licensing Data. In C. Rao & S. Sinharay (Eds.), *Handbook of statistics* (vol. 26: Psychometrics, pp. 893–918). North Holland: Elsevier B.V.

Wright, D. B. (2006). Comparing groups in a before-after design: When *t* test and ANCOVA produce different results. *British Journal of Educational Psychology*, *76*, 663–675. https://doi.org/10.1348/000709905X52210

Wright, D. B. (2017). Using graphical models to examine value-added models. *Statistics and Public Policy*, *4*, 1–7. https://doi.org/10.1080/2330443X.2017.1294037

Wright, D. B. (2018). Estimating school effectiveness with student growth percentile and gain score models. *Journal of Applied Statistics*, *45*, 2536–2547. https://doi.org/10.1080/02664763.2018.1426742

Received 14 February 2019; revised version received 20 May 2019

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. R Code for the simulations