Some methods for ranking schools are unfair

Daniel B. Wright

University of Nevada, Las Vegas

Abstract

Value added models (VAMs) have been used in education to measure the effectiveness of schools based on student test scores. Much research has questioned the use of these procedures. Here it is shown that this measure is systematically biased against schools that serve students from historically poor performing groups in some situations. Other situations are shown where some statistical procedures are biased against schools that serve students from historically high performing groups. The main conclusion is that before any statistical procedures are used for high stakes decisions their accuracy should be examined based on how the data are likely to arise.

Keywords: fairness; effectiveness; education

Modern society measures many things (Muller, 2018). Many different societies use student test scores in some way in their accountability systems (see papers in Holloway et al., 2017). Sometimes the statistical procedures used to analyse data favour particular groups of people; what is called fairness within education (American Educational Research Association et al., 2014). Often these favour the group who is in power and who is implementing the measurement (Walter & Anderson, 2013). In education, many jurisdictions estimate the effectiveness of schools (and sometimes teachers) using student test scores. These are often presented as "league tables" in a manner similar to sports team standings or by assigning A to F grades to the schools. These estimates can have serious consequences including school closures.

Using test scores to estimate school effectiveness has been debated in the US and the UK by both statisticians and education policy researchers (e.g., Amrein-Beardsley, 2014; Goldstein & Spiegelhalter, 1996, and many more). For a review focused on history of accountability measures in education focused on the UK see Leckie & Goldstein (2017). The focus here is on a particular model often called the *value added model,* abbreviated

This paper is based on Wright (in press), which examined just the linear Ancova and gain score models (but goes into more depth on Lord's [1967] paradox). If readers do not have access to that paper and wish to read it, please email Dr. Wright.VAM (for review, see Castellano & Ho, 2013), and variant of it. A basic version of this is a multilevel linear Ancova where one of the covariates is the previous scores. The students are nested within the school and the conditional mode for the intercept for the school is used to estimate the effectiveness of the school (for technical details of these models, see Bates et al., 2015; Goldstein, 2014). In simplest terms (and supposing ideal circumstances), suppose all students take an assessment prior to entering the school (maybe the year before at their previous school) and then one at the end of their schooling. Let the students be index by the subscript i and the school that they are in by the subscript *j.* The model (#5 in Aitkin & Longford, 1986, p. 12) is:

$$post_{ij} = \beta_0 + \beta_1 priori_{ij} + u_j + e_{ij} \qquad (1)$$

where u_i and e_{ij} are assumed drawn from normal distributions. The conditional modes (also called conditional means, empirical Bayes estimates, shrunken estimates, and school residuals) estimate the school intercept deviance from the overall β_0.

There are variants of this, for example estimating individual β_0 effects in the model for each school (often called the fixed-effect approach), using longitudinal models, adding further covariates, not using any covariates, allowing a more flexible but still monotonic curve between $prior_{ij}$ and $post_{ij}$, fixing β_1 1, *etc.* The focus here is on eqn. 1, not using any covariates (the status model, equivalent to letting β_1 = 0 in eqn. 1), and fixing β_1 to 1 (the gain score model). Simulations show that these approaches can produce estimates of effectiveness that have small and even negative correlations with true effectiveness (e.g., Wright, 2017, 2018). The focus here will be on fairness: whether these approaches favour one group over another.

It is worth mentioning that demographics are sometimes included in these models. Sometimes politicians argue against their inclusion because there are regulations against doing so for student data, when estimating values for students. This is not relevant here since the student test scores are being used for a different purpose. The group variable is not used here because if it were any group differences would not be apparent. The models would all show no differential effects.

Fairness of Value Added Models

There are many issues using VAMs to estimate school effectiveness, both the statistical procedures and the consequences of these measurements. Examples of negative consequences in the US include:

- Teachers and school administrators changing student responses to increase their schools' scores (Blinder, 2015).
- Parents being concerned that teachers teach for the particulars of the test and not for the students' learning of the subject.
- All stake holders questioning whether the weights given to the different facets within the algorithm are appropriate.

The focus here is on whether the statistical procedures measure what the policy makers want them measure (and what they have been told that they measure). Do they measure, as the name suggests, the value a school adds to the students' performance? The short answer is often *no.* It will not surprise readers of *Radical Statistics* that a brief bumper sticker label like "value added model" over-simplifies the procedure to the point of being misleading. It also won't surprise readers that the method used to explain these procedures is obfuscation, coupled with a few simple but errant tidbits like "the procedure levels the playing field for all schools" or "using covariates controls for everything (or anything)," that might convince the uncritical. In a brilliant essay, Braun (2013) describes how some people succumb to this propaganda and believe this procedure magically creates accurate measurements. Goldstein (1991, p. 91) brings these beliefs down to the earthly realm: it is "most certainly not a magic wand that will allow us automatically to make definitive pronouncements about differences between individual schools." When peeling away many layers of obfuscation (this is a high frequency word in some government departments), the algorithms remain baffling. When scientists and statisticians at Los Alamos National Laboratory tried to understand the school grading system in New Mexico, they concluded that the procedures remained unclear (Nott, 2013).

Fairness, or bias, refers to a procedure tending to produce higher scores for one group than for another even when their true scores are the same. The focus is on whether the basic model in eqn. 1, and some other models that are often considered, are biased when estimating effectiveness for schools that serve different proportions of students from historically poor performing groups. In the US, States report achievement gaps among groups of students and generally find some groups (e.g., some ethnicities and those from lower socio-economic groups) tend to perform worse. One of the reasons given for this is that statistical methods like VAMs estimate that these schools (and their teachers) are less effective educators than other schools. While these schools may not be as good (and there is some evidence for this using more valid measures), the main point of this paper is to show that the basic value added model is biased against schools that tend to serve historically low performing groups of students. This is shown using simulation, but for discussion of the mathematics for the non-multilevel version see Holland & Rubin (1983). Before presenting the simulation, there is a brief discussion of two data models in the simulation. Next, there is discussion of Lord's paradox (Lord, 1967, 1969) and statistical models usually considered when discussing this paradox. The focus in this paper is on simple data and simple statistical models because this helps in identify the issues. Problems like missing data, students transferring between schools, *etc.* will not be considered.

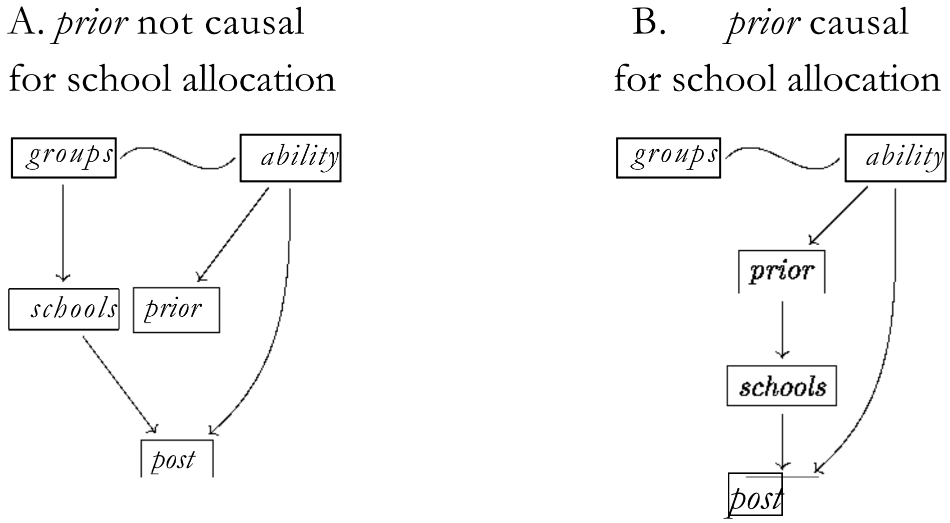
Two Data Models

It is useful both to consider different ways for how your data may have arisen and to create graphs to show these (for discussion of using graphs to identify causes in scientific contexts, see Pearl, 2009; Pearl et al., 2016). Figure 1 shows two ways in which the data might arise.

These models are simple; real data will tend to be messier, but these will illustrate when biases can occur. These models have five variables each: *groups* (groups of students, for example high and low socio-economic status); *schools; prior* (before entering the school, it is more complex if this measurement is while at the school); *post;* and a latent variable *ability* for the ability on academic tests.

Panel **A** of Figure 1 shows that some *groups* tend to go to certain *schools* (because, for example, income is not evenly distributed among school districts and students usually go to school close to where they live), that *groups* vary in *ability,* and that *ability* influences. In panel **B** *prior* does influence *schools.* The curved path between *group* and *ability* denotes a correlation (the achievement gap) without reference to any causal direction. *prior* scores. *schools* and *ability* then influence the *post* scores. The critical edge in the graph is the direct path from *schools→post.* This corresponds to true school effectiveness in this model. More variables and connections between all of these variables could be added, but the simple data models used here allow focus on the effects related to whether the prior scores affect school allocation.

Figure 1. Two models of how the data may arise.



A. *prior* not causal for school allocation

B. *prior* causal for school allocation

In Panel **A** prior does not influence schools. In Panel **B** of Figure 1 *prior* influences which school a student attends. This could be if *prior is* used for an entrance exam. As with the Panel A, *schools* and *prior* influence *post,* and it is the direct effect from *schools* to *post* that the statistical models attempt to estimate for school effectiveness. While *prior* does not influence *schools* for many public elementary and high school systems, it is more common for universities where measures (e.g., high school GPA, ACT or SAT scores, 'A' levels) are often used for admissions and are often used as covariates predicting outcomes like university GPA and graduation. In practice, some situations will be a mixture of these data models.

Lord's Paradox

Lord (1967) described a situation where two statisticians offer different "solutions" for the same set of data and come to different conclusions. A large university is interested in the effects of university dining on the health of its students, and if there are any differences in these effects by gender. The university measures weight in September and June. The first statistician shows that the females weighed about the same in September and June, and so did the males, so not only was their no overall gain for either group, there was no gender difference in the gain. This is often called the gain score approach. The second statistician ran an Ancova (let *gender* = 0 for males and 1 for females): $June_i = \beta_0 + \beta_1 gender_i + \beta_2 September_i + e_i$, and found $\beta_1 < 0$ showing that conditioning on September weights, males had higher expected June weights. Many researchers, with before and after scores, wishing to examine the differential effect of some manipulation on two groups would be choosing between these two methods. Given that either of these may seem plausible, but that they lead to difference conclusions, led Lord to labeling this a paradox.

Since Lord described this paradox several papers have described when one procedure would be preferred over the other (e.g., Hand, 1994; Holland & Rubin, 1983; Pearl, 2016; Wainer, 1991; Wright, 2006). The focus here is on whether the covariate influences group allocation. Within the Lord's paradox context, if the covariate (September weight in his example, prior score in the example here) influences group allocation (gender in his example, school here), then the Ancova can yield unbiased results for the difference in the causal effect. If it does not, then the gain score approach may yield unbiased results, though some other assumptions are also necessary (like the two scores are on the same scale). The gain score model for the school effectiveness example can be written as:

$$post_{ij} = \beta_0 + 1.prior_{ij} + u_j + e_{ij} \qquad (1)$$

Because this equation appears so similar to eqn. 1, just constraining β_1 to be equal 1, it can surprise people that the two procedures often produces different results. If *post* and *prior* are measured on the same scale and each with measurement error, β_1 will tend to be estimated as less than 1 (this is regression towards the mean). If there are no differences for the gain scores, it will mean that the predicted value from the Ancova approach will be higher for the group with the higher initial mean.

In addition to the gain score model, the status model will be examined. Like the sports league tables, it just using the final outcomes for accountability. Some education accountability systems use measures like proportion reaching proficient or percentage graduating without taking into account any prior information. It is widely recognized that this there are problems using this approach to assess the causal impact of the schools. To make this as similar to eqns. 1 and 2, this model will be that same as eqn. 1 just fixing β_1 at 0.

$$post_{ij} = \beta_0 + u_j + e_{ij} \qquad (3)$$

This is a simple variance component model.

Estimating Effectiveness in Education

Many schools are evaluated based in part on models like eqn. 1 that use students' scores on standardized tests. The gain score and status models are presented for comparison. More complex models exist, but at their core is using previous scores (and sometimes other variables) to predict later scores, finding how much above or below the students' scores are from their predicted scores, and then aggregating these differences by school (see Castellano & Ho, 2015, for issues regarding different aggregation methods within the context of educational effectiveness).

Simulation Methods

Simulation is a useful method to test how the properties of different statistical procedures vary by changing how the data are constructed. For most simulations the true effects are known so it is relatively easy to assess the accuracy of estimates. Further, the simulations can be repeated as many times the researcher wants in order to achieve the desired degree of precision.

Only two conditions will be used here, corresponding to the two data models in Figure 1. Let there be 10,000 students divided among 100 schools. For both of these there will be no difference in the effect of schools by which group of students they tend to serve. This makes it easy to show if there is a bias (other data models were used in additional simulations, but these show the main finding). The groups will be called high and low, for historically high and low performing groups.

For each replication, the sample is divided into two approximately equally sized groups (e.g., these might be those above and below the median on household income, or different ethnicities).

$$StudGr_i \sim Bernoulli(n = 10,\ 000,\ p = .5)$$

There is a latent variable, in the R code in the Appendix called *achieve,* for individual differences among students that influence test scores. The latent variable for both groups is normally distributed, but the one for the higher performing group is 0.2б higher (Cohen's *small* effect).

$$achieve_i \sim Normal(n = 10,\ 000,\ \mu = -2 + 4\ (StudGr),\ \text{б}\ \ = 20)$$

The *prior* scores are based on this variable and normally distributed random error. Within each group, *achieve* and the random error have the same standard deviation *(*б = 20).

$$prior_i \sim Normal(n = 10,\ 000,\ \mu = achieve,\ \text{б}\ ^{-} = 20)$$

Half the schools are labelled *high* (1-50) and half labelled *low* (51-100) for which types of students they tend to serve. The subscript *j* will be used for schools.

For the Panel A *(prior → school)* students from the historically high performing group have an 80% probability of being assigned to a *high* school *(StSchGr* = 1) and students from the historically low performing group have an 80% probability of being assigned to a *low* school *(StSchGr* = 0). The prior test score plays no role in this allocation.

$$StSchGr_i\ Bernoulli(n = 10,\ 000,\ 0.8(group) + 0.2(1 — group))$$

For Panel B *(prior → school),* students who score above the median on *prior* have an 80% probability of being assigned to a *high* school and those who score below the median have an 80% probability of being assigned to a *low* school. Let $ab_i = 1$ if *prior > median(prior)* and 0 otherwise, then

$$StSchGr^{ti}\ Bernoulli(n = 10,000, 0.8(ab_i) + 0.2(1 — ab_i))$$

A normally distributed variable for school effect was created:

$$SchEffect_j \sim Normal(n = 100, \mu = 0, б = 5)$$

Importantly, this variable does not differ systematically by the make-up of the school (or anything else). Therefore it is known that, on average, an unbiased procedure should find no differential effect for schools that tend to serve historically high performing groups of students than those that tend to serve historically low performing groups of students. The *post* scores are based on *achieve,* the school effect, and random error $(б = 20)$.

$$post_{ij} \sim Normal(n = 10,000, \mu = achieve_{ij} + SchEffect_j, б = 20)$$

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.

The statistical procedures use functions from the R package **lme4** (Bates et al., 2015). The **lmer** function estimates the multilevel model and the **ranef** function estimates the conditional modes (the deviation from the etimated β_0). The three models are those in eqn. 1-3. The relevant code is in the Appendix.

Simulation Results

Table 1 shows the asymptotic 95% confidence intervals and means for the 2,000 replications for the data models for Panel A and Panel B, for models of eqns. 1-3. It is easiest to start with the status model. For both data situations, the effectiveness of schools that serve predominantly students from historical low performing groups is underestimated and the opposite for the other schools. Some people argue that like how a football team that wins all of its matches is considered good, a school that gets good results is doing well. The problem with this argument is that these measures are meant to tap the causal effect of the schools on the students, not the students' achievements (yes, the two should be related, but they are not the same). As shown here, they do not do this and they are systematically biased in both situations tested here (and in most reasonable situations that could be considered). The reminder of the results and discussion will focus on the other two statistical procedures.

The result for the VAM for Panel A is that it estimates that the schools that tend to serve groups of historically high performing students are better than those that tend to serve groups of historically low performing students. This despite that we know there is no difference. This result is important because Panel A is a better approximate for most K-12 (or primary and secondary schools) than Panel B and the VAM (or other types of Ancova) is a common method. The gain score model does not show this bias. A different result emerges if the data arise in the manner depicted in Panel B. Now the VAM produces unbiased results for the difference between the groups of schools, but the gain score method produces a large difference in favor of the schools that tend to serve low performing groups of students.

Summary

The statistical reasons for these effects are simple and based on Galton's (1886) regression towards the mean (Wright, in press). For Panel A, the two groups of students each regress towards the mean of their own group. Wainer & Brown (2007) describe this as Kelley's paradox. They call it a paradox because of one application of it. They showed that if you matched students on standardized test scores (in the US, ACT and SAT are the main examples) and then compared university output between students from groups that differ on these scores, those students from the low performing groups tended to do worse than their matched counterparts from the high performing groups. The reason is the initial scores are a combination of true score and error, and if you are scoring higher than your classmates it is likely that both of these components are positive. When you are retested, if the error is random, it is likely that it will not be as high. How this result is used in education is controversial, but the statistical explanation is straight-forward. Efron & Morris (1977) use a sports example to show this shrinkage towards the group mean, which is a less controversial context.

Table 1

The means and upper and lower bounds for the 95% confidence intervals for the two data models, for the three statistical models, and for schools that tend to serve historical high and low performing groups of students. All the true effects (known, because this is a simulation) are zero.

		Panel A		Panel B	
Status	Mean	-0.92	0.91	6.15	6.19
	95% CI	(-0.95, -0.89)	(0.88, 0.94)	(-6.18, 6.11)	(6.15, 6.22)
VAM	Mean	-0.48	0.48	0.01	0.01
	95% CI	-0.50, -0.46)	(0.46, 0.50)	9-0.02, 0.01)	(-0.01, 0.02)
Gain	Mean	-0.00	0.01	6.12	-6.06
	95% CI	(-0.03, 0.02)	(-0.03,0.02)	(6.08, 6.15)	(-6.10, -6.03)

For Panel B, the students also regress. Here however those chosen to be in school because of their high test scores are likely to regress downward. An extreme example is if all students performing above the median were assigned to School X and all below the median were assigned to School Y. If the students are re-tested at the end of the year, in terms of percentage in the top half, School X cannot improve so can only appear less effective than School Y. Another example is the *Sports Illustrated jinx* (https: //en.wikipedia.org/wiki/Sports_Illustrated_cover_jinx). This is a US magazine that usually puts a photo of an athlete who has recently performed well on its cover. The "jinx" is that, after being on its cover, the athletes tend to perform less well than they did before being on the cover. This is a simple demonstration of regression towards the mean. The main statistical concepts for the results in Table 1 have been around for over one hundred years. An obvious question is: Why some of these procedures have been used for high stakes decisions when these biases are known? I do not have an answer, obvious or otherwise.

It is important to stress that this paper has not shown nor is the author arguing for either the VAM or gain score models to be used. While the gain score model does perform better for the simulated data in Panel A, it requires other assumptions. Further, in practice the way that the data arise will be much more complex than depicted in either Panel. While the data models used will always be just approximations, it is important for those analysing such data to develop more thorough approximations of the data models based on knowledge of their particular context. The analysts should then conduct simulations of different statistical models, like done here, to decide if these statistical models produce valid, reliable, and fair estimates (Wright, 2017).

In relation to education policy, it is important to improve school effectiveness for all schools including those serving groups that historically perform poorly. There are many initiatives for this. Some involve moving principals and teachers who have received high VAM scores from schools that serve predominately historically high performing groups to schools with low VAM scores that serve predominately historically low performing groups. The difficulty is that if school allocation for students is like that in Panel A, then it may be that the effectiveness scores of these principals and teachers are too high (so they may not be as effective as thought) and the locations that they are being assigned to may be more effective than thought.

For any education policy to be useful it is important that it is based on accurate (valid, reliable, and fair) estimates. There are many controversial issues about using student test scores to evaluate schools. This paper is not addressing whether, in principle, test scores should be used. There are arguments for and against this. This paper addresses one aspect of the accuracy of a common statistical method used and shows that it is biased in systematic ways.

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. doi: 10.2307/ 2981881

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software,* 67(1), 1-48. doi: 10.18637/jss.v067.i01

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

Braun, H. I. (2013). Value-added modeling and the power of magical thinking. *Ensaio: Evaluation of Public Policies in Education [Brazil], 21,* 115-130. doi: 10.1590/SO104 -40362013000100007

Castellano, K. E., & Ho, A. D. (2013). A *practitioner's guide to growth models.* Council of Chief State School Officers. Retrieved from http: //scholar .harvard. edu/andrewho/ publications/practitioners-guide-growth-models

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics, 40,* 35-68.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236,* 119-127.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland, 15,* 246-263. Retrieved from www.jstor.org/stable/2841583

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. doi: 10.1146/annurev -statistics-022513-115615

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society. Series A (Statistics in Society), 159(3),* 385-443.

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

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.

Holloway, J., Sorensen, T. B., & Verger, A. (2017). Global perspectives on high-stakes teacher accountability policies: An introduction. *Education policy analysis archives, 25(85).* Retrieved from http : //dx . doi . org/10. 14507/epaa . 25.3325

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 19922016: 'Contextual value-added', 'expected progress' and 'progress 8'. *British Educational Research Journal, 43,* 193-212. doi: 10.1002/berj.3264

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

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

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

Nott, R. (2013, December 16). Los Alamos scientists: School grading system is unclear. *The New Mexican.* Retrieved from https : //www. santaf enewmexi can . corn/ news/education/los-alamos-scientists-school-grading-system-is-unclear/ article_Oc2103fa-a7b9-538b-b401-e56317d6c310.html

Pearl, J. (2009). *Causality: Models, reasoning, and inference (2nd ed.).* New York: Cambridge University Press.

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

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer.* Chichester, UK: Wiley.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin, 109,* 147-151.

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. 893918). Elsevier B.V.

Walter, M., & Anderson, C. (2013). *Indigenous statistics: A quantitative research methodology.* New York, NY: Routledge.

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. doi: 10.1348/000709905X52210

Wright, D. B. (2017). Using graphical models to examine value-added models. *Statistics and Public Policy,* 4, 1-7. doi: 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. doi: 10.1080/02664763 .2018.1426742

Wright, D. B. (in press). Allocation to groups: Examples of Lord's paradox. *British Journal of Educational Psychology.*

Department of Educational Psychology and Higher Education. Box 453001. 4505 S. Maryland Pkwy., Las Vegas, NV 89154-3001. Email: daniel.wrightOunlv.edu or dbrookswragmail.com. Dr. Wright is the Dunn Family Endowed Chair and Professor of Educational Assessment. There was no funding beyond this endowment for this paper.

R Code for the simulations

The function used to run the simulation is below. It allows the user to choose the number of replications, the number of schools and students, which of the two graphs in Figure 1 to use, the probability that students are assigned to their group's school, and the name of a data file to create. Readers are encouraged to adapt the code for their own needs and contact the author with any questions or comments.

```
makedata <- function(reps=1000,nschool=100,nstud=10000,
  dag=1,ratio=.8,name="sim",writeit=FALSE){

effvals    matrix(nrow=reps,ncol = 2*3)
for (x in l:reps){

  StudGr   rbinom(nstud,l,.5)

  achieve <- rnorm(nstud,-2+StudGr*4,sd=20)

  prior <- achieve + rnorm(nstud,sd=20)

  PreGr   as.numeric(prior > median(prior))
  school <- vector(length=nstud)

  SchGr  rep(0:1,each=nschoo1/2)
  ifelse(dag==1,

    StSchGr    rbinom(nstud,l,ratio*StudGr+(l-ratio)*(1-StudGr)),

    StSchGr    rbinom(nstud,l,ratio*PreGr+(l-ratio)*(1-PreGr)))

  school [StSchGr == 0] <-

    sample(1:(nschoo1/2),sum(StSchGr==0),replace=TRUE)

  school [StSchGr == 1] <-

    sample((nschool/2 + 1):nschool,sum(StSchGr==1),replace=TRUE)

  seff     rnorm(nschool,sd=5)

  SchData   cbind(1:nschool,seff,SchGr)

  colnames(SchData)  c("school","TrueVA","SchGr")

  StudData   cbind(StudGr,achieve,prior,school,StSchGr)

  SimlData    merge(StudData,SchData,by="school",all.x=TRUE)

  SimlData$post    SimlData$achieve + rnorm(nstud,sd=20) +
    SimlData$TrueVA
  AvePrex    aggregate(Sim1Data$prior,by=list(Sim1Data$school),mean)
  colnames(AvePrex)   c("school","AvePre")
  Siml    merge(Sim1Data,AvePrex,by="school",all.x=TRUE)
  ml    unlist(ranef(lmer(post-prior+(lIschool),data=Sim1)))
  effvals[x,1:2]       tapply(ml,SchGr,mean)
  m2 <- unlist(ranef(lmer(post - prior - 0 +
      (lIschool),data=Sim1)))
  effvals[x,3:4]       tapply(m2,SchGr,mean)
  m3 <- unlist(ranef(lmer(post - 0 +
      (lIschool),data=Sim1)))
  effvals[x,5:6]       tapply(m3,SchGr,mean)
    }
if(writeit) write.csv(effvals,paste0(name,".csv"))
return(effvals)}
```

The following code runs the two simulations.

```
set.seed(815) #no meaning for these seeds
effvalsl <- makedata(writeit=TRUE,
         name="priornotinfluenceallocation",reps=2000)
set.seed(825)
effvals2 <- makedata(writeit=TRUE,
         name="priorinfluenceallocation",dag=2,reps=2000)
```