Abuse of Statistics – Linking Student Test Scores to Teacher Accountability

Margaret Wu

Victoria University, Melbourne, Australia

Email address: wu@edmeasurement.com.au

Postal address:

PO Box 3011
Strathmore
Victoria
Australia, 3041

Keywords: test scores, accountability, teacher performance

Introduction

The recent publication of the names of teachers and their "performance scores" by the New York Post[1] (February, 2012) is so disturbing that I feel statisticians must take actions to stop such abuse of statistics, and consequent abuse of people. Despite repeated warnings by academics[2] about the inappropriate use of student test scores for teacher accountability purposes, some press still prefer sensationalizing news than being responsible. I now believe that unless there are dire consequences for the press involved, they will continue to abuse statistics to sell newspapers. I hope this article will put fears in abusers of statistics so that such abuse will never happen again. It is my intention to demonstrate that, beyond reasonable doubt, the likes of New York Post have a case to answer for regarding defamation of innocent people. I will establish my case using *proper* uses of statistics in a non-technical way, so, after all, the abusers of statistics will have a taste of their own medicine, and statistics will turn on its head and triumph over its misuse. So the battle line is drawn.

In the case of the New York Post article, the so-called teachers' performance scores are based on students' exam scores in mathematics and reading. Irrespective of how these performance scores are computed using student test scores, I will make two unequivocal statements:

(1) Teacher accountability cannot be established by student test scores

(2) Statistical inferences alone cannot be used for any high-stakes decision making.

Making Inferences

To link student test scores to teacher performance, an *inference* needs to be made, since we haven't directly observed teachers' teaching (performance). Inferences are conjectures made by people, not proven by statistics. For example, when we see a person driving an old and battered car we might make an inference that the person is not that well-off. This inference may be correct some of the time, or even most of the time, but there are always exceptions. So inferences can never be used as proofs. There is always a margin of error when inferences are made.

In the case of using student test scores to judge teacher performance, the inference made has a huge margin of error, simply because there are so many factors impacting on student test scores and their gain scores (value-added scores). Even if we control for students' socio-economic status (SES), there are many other factors that have large impacts on the academic growth of students, such as parental support, natural academic ability, motivation, interests, personality, and cultural and ethnic differences. Above all, there are always many exceptions where there are individual students with high SES background performing poorly and students with low SES background performing well, due to factors completely unrelated to schools and teachers.

So how can the New York Post claim that in some cases there is no margin of error? I believe this is the result of a misunderstanding about margins of error regarding test scores and making inferences. Let's suppose we have a student who has a learning difficulty. It will not be surprising that the student has very low gain scores because of learning difficulties. The margin of error surrounding the student's score may be very small, since we have reliably measured that the student could do little in academic tests. But the margin of error in making an inference about teacher performance based on this test score is large, in fact, totally invalid, since the low gain

score has nothing to do with teacher performance. This is just one of many examples that demonstrate that there is a margin of error surrounding the test scores, but there is a larger margin of error surrounding the inferences made regarding teacher performance.

This is the first point why student test scores cannot be used for teacher accountability purposes. That is, when student test scores are used as teacher performance measures, inferences are made. The margin of error surrounding such inferences is generally large, as we shall see in the next section.

Measuring Group Effect Versus Measuring Individual Effect

Many education research studies use value-added models to estimate the effect of factors contributing to students' academic success, including teacher effect. With a large amount of data, we can estimate a ballpark figure for overall how teachers make a difference to student achievement. Specifically, teacher effect is found to be around one year of growth[3]. That is, a very effective teacher can bring students up one more year than an ineffective teacher. Say, on average, the growth of students in one calendar year is defined as "one year of growth". A very effective teacher can bring students up one and a half years, while an ineffective teacher brings students up only half a year. These figures are somewhat plausible. Any larger teacher effect would not be credible. I have not yet seen any teacher who can make a whole class of students finish six years of primary schooling in three or fewer years, except for a handful of gifted students. To claim that a teacher can make a class of students grow two or more years in one calendar year is simply not very likely. I know the press like to report success stories and people like to hear success stories. But these are mostly exceptions rather than the norm.

The fact that we can get a handle on an overall teacher effect leads many people to think that we can measure *individual* teacher's effect accurately. This is a misperception. We can often measure a group effect without being able to measure individuals accurately. For example, consider a weight loss program that claims people will lose 0.5 kg per week. If my scale only shows weight in whole numbers of kgs, I will not be able to accurately measure my weight loss in fractions of kgs. But if we use the same scale with 1000 people in the weight loss program, we may find that 500 people lost 0 kg, and 500 people lost 1 kg, so, on average, we can still establish a ballpark average weight loss for the group. That is, an overall effect of weight loss can be estimated, but not for an individual persons' weight loss. The same argument applies to measuring teacher effectiveness: While we can provide a rough ballpark figure for overall teacher effect, there is little chance that individual teacher effectiveness can be measured through student test scores. Since the range of teacher effect is about one year of growth, to separate effective from ineffective teachers within this one year of growth, our measures need to be accurate to fractions of a month's growth. Achievement tests, like my scale for measuring weight, just don't provide that kind of accuracy.

Error Rates in Measuring Teacher Performance using Student Test Score Gains

A report[4] from the National Center for Education Evaluation, U.S.A, concludes that the error rate in labelling an average teacher as high- or low-performing is 1 in 3, when one year of student test gain scores are used. When 3 years of data are used, the error rate is 1 in 4. Note that one cannot

do any worse than an error rate of 1 in 2, since that is the outcome of tossing a coin. Essentially, the accuracy of labelling teachers as effective or ineffective based on students' test gain scores is not much better than tossing a coin. The reason for this is there is a huge variation in students' test gain scores because students are very different in their academic abilities. We only need to look into one school, we will notice that the students are not all the same, and there is a large variation in academic abilities within a school. It is the luck of the draw for teachers whether they have a good class or a bad class, since teachers generally do not choose students – students are assigned to teachers. Even without changing the way a teacher teaches, a teacher's class results can differ greatly, as noted by Darling-Hammond et al (2011, page 5)[5]:

> ***As one teacher noted:***
> *I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back. [My] second year got me kicked in the backside. And for year three my scores were off the charts. I got a huge bonus, and now I am in the top quartile of all the English teachers. What did I do differently? I have no clue.*[6]
> ***Another teacher classified her past three years as "bonus, bonus, disaster."***

The large variation in students' academic abilities in a class, together with the unreliability in test scores (since tests are typically quite short), result in random fluctuations of a teacher's class test scores, which contributes to the misidentification of teachers as ineffective or effective.

When we have many years of student data (e.g., over 10 years), the error rate of misidentifying a teacher as effective or ineffective will be lower. So what error rates will be acceptable? One in 10? It depends on how high-stakes the results are used. If a teacher is publicly named and shamed, or even sacked based on student test scores, an error rate of one in 10 or even one in 100 will still be too large. That is, one out of 10 teachers will be labelled as ineffective with the consequence of the loss of reputation and job, by mistake! Surely no court of justice will allow that! So how about error rates like 1 in 3 or 1 in 4, as is typically the case for many data sets collected by education authorities? We will never allow such error rates in medical diagnosis or other professions, so why does the law allow the data to be published when the error rate of mislabelling teachers is so high?

Some Bad News for Abusers of Statistics

The New York Post reported "the rankings have an average error range [in percentile points] of 35 in math and 53 in English". This means that a teacher deemed in the bottom 1 percent of teacher rankings may not be shown to be different from a teacher at the 50th percentile. In fact if we randomly assign ranks to teachers, the largest margin of error is 100 (when a bottom teacher is assigned the top rank, or vice versa). The lowest margin of error is 0 (spot-on rank). So the average margin of error is 50 percentile points - a better result than the 53 percentile points reported. Try to figure this one out! (I am beginning to think this is one of Martin Gardner's *aha* moments for mathematicians, or Darrell Huff's *How to Lie with Statistics*.)

Let's have some fun with statistics. There are 12,000 teachers in the case of the New York rankings, so one percent is 120 teachers. That is, by definition of percentile rank, 120 teachers are in the bottom 1 percent. Suppose there is a probability of 1 in 5 that a teacher in the bottom 1 percent should not be there (I am being generous to the abusers of statistics here. I think the error

rate is far greater). So we expect around 24 teachers being wrongly labeled as "in the bottom 1 percent".

To make the case that we are confident that at least some teachers in the bottom 1 percent have been wrongly placed, we will compute that probability. So, the probability that some of the 120 teachers have been misidentified as "in the bottom 1 percent" is

$$1 - probability\ (none\ of\ the\ 120\ teachers\ is\ misidentified) = 1 - 0.8^{120} = 1$$

EXCEL returns the value 1 for this computation. So, as far as EXCEL is concerned, there is certainty that some teachers have been wrongly labeled in the bottom 1 percent of teacher ranks. Surely this probability is beyond reasonable doubt, and we now have a case to file for defamation.

In fact, if the error rate of identifying teachers is 1 in 10, the probability is 0.99999. If the error rate is 1 in 20, the probability is 0.99788. So, reporters, judges and proponents of value-added models, please note that statistics can demonstrate a defamation case beyond reasonable doubt.

Conviction by Numbers

To illustrate the misuse of statistics in labeling teachers as low or high performing and applying sanctions or rewards, I will bring your attention to the Sally Clark case[7]. Sally Clark was a lawyer in the United Kingdom. Sally's two sons were thought to have died of SIDS (Sudden Infant Death Syndrome). But in 1999 the court convicted Sally of murdering her two sons based on statistical evidence that the chance of having two children die of SIDS is extremely small. Her conviction was later overturned in 2003 after she served more than three years of her sentence. One of the reasons for the overturn of Sally's conviction was that statisticians argued that some people may be predisposed to certain medical conditions, and the chance of having two SIDS in one family may not be so unlikely, as there may be genetic or other reasons for SIDS that run in the family. Sadly, Sally never quite recovered from this ordeal, and she died in 2007, aged 42.

Misusing statistics has dire consequences and people's lives may be irrevocably ruined. I would like you to consider three issues in relation to the Sally Clark case and compare them to the current abuse of student test scores for judging teachers.

First, an inference was made. Sally's conviction was not the result of direct evidence, but the result of an inference made based on statistics. Similarly, for making judgments on teacher performance using student test scores, inferences are made based on statistics, not based on direct evidence.

Second, in computing statistics, many assumptions are made. In the Sally Clark case, incorrect assumptions were made. That is, the two SIDS deaths were not independent events, but they were assumed to be. In the case of using value-added models, many assumptions are also made, such as the assumption that controlling for a number of school contextual variables will remove all factors unrelated to teacher performance.

Third, and the most important point, is about the use of statistical inference *en masse*. This is a little harder to explain, but I will try. Let's assume that the incidence rate of SIDS is 1 in 1000 births[8]. For two SIDS deaths to happen to one family, the probability is 1 in a million (1000 X 1000), assuming that SIDS happens independently. "1 in a million" sounds like a rare event, but if we look at 20 million families[9] in the world, we would expect to find about 20 families with two SIDS deaths happening just by chance. Surely we will not charge these families for murdering their children, since it is with almost certainty that we will find families with two SIDS deaths happening just by chance, if we look at a large number of families. So chances are that Sally Clark is just one of these random cases. Sally came to the attention of the authorities because of the two SIDS deaths, not because of other evidence.

Statistical inference is not meant to be applied to a large number of (unsuspecting) cases[10]. The right way to use statistics is to provide supporting evidence in the presence of other evidence. Suppose someone comes to the attention of the authority based on evidence of child abuse and other evidence, then the statistics of having two SIDS deaths can be used to add further evidence that two SIDS deaths are not likely. In this case, we only use the statistical inference once about one person, and we only use the statistical information because there is already other evidence.

The key message is that statistical inference alone should never be used to convict anyone in the absence of other evidence. That is, we should not look for suspicious cases by looking at a whole population of cases, because there is almost certainty to find an unusual case that happened by chance.

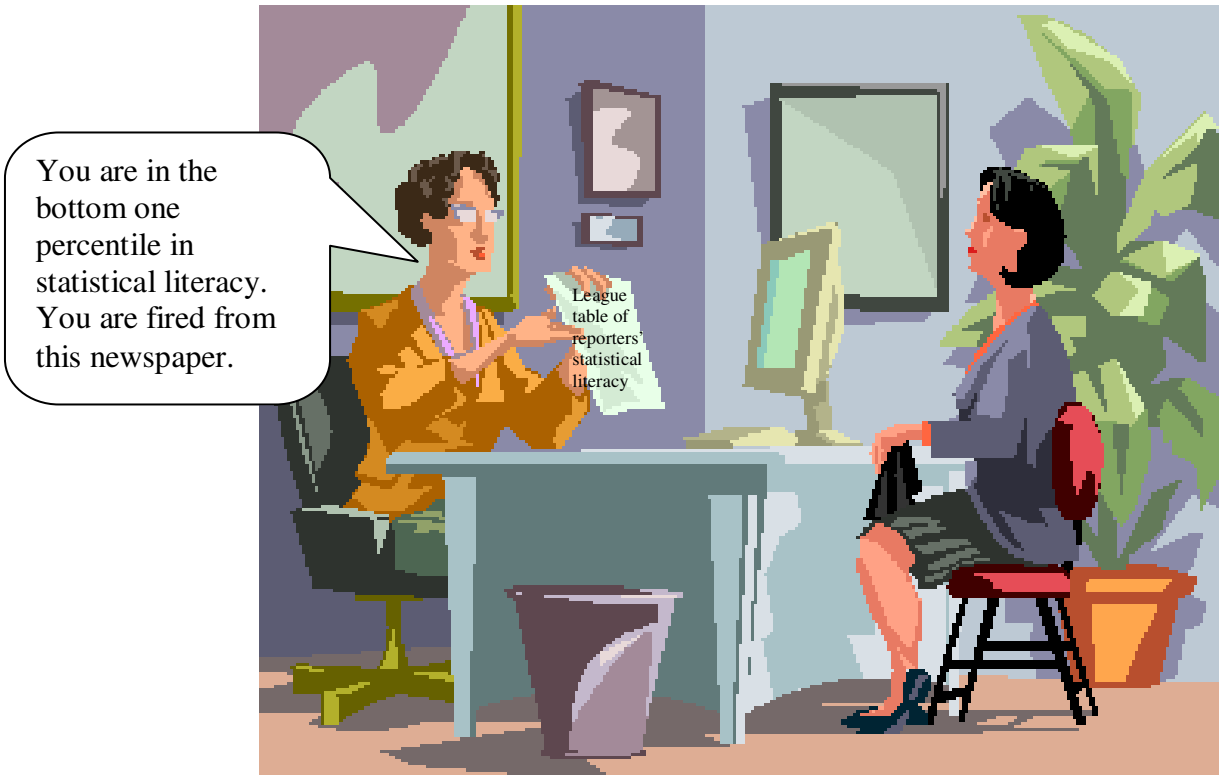
Similarly, if we look at 12,000 teachers' student test scores, we are bound to find teachers in the bottom percentile by pure chance. (Technically, to identify teachers in the bottom percentiles, we have conducted 12,000 statistical significance tests, not just one or two.). Consider this example. If there is in fact no difference between teachers' performance, we will still find teachers' class test scores to spread over a range by chance. We will always be able to find 5% of low scoring teachers, but there is no evidence why these are not just due to chance. Surely we can't name and shame these teachers. This example should be sufficient to demonstrate that statistics alone should not be used for high-stakes purposes.

Statistical testing should be used when a hypothesis is tested in a *confirmatory* way. For example, if it comes to the attention of the authorities through community feedback that a school is not functioning well, then the school academic results can add evidence to support that. In contrast, if we use an exploratory approach by looking at school academic results first with no prior hypothesis of which schools may be dis-functioning, we cannot use low academic results as indicators of low school performance[11]. If you insist on doing exploratory analysis like ranking teachers' class test scores and identifying the bottom 5% of teachers, you can only put these cases up for "possible further investigation", because, chances are, there are good reasons for why the schools are in the bottom 5%. You should never apply sanctions or rewards based on just the rankings.

Finally…

Students' test gain scores (value-added model) simply cannot be used for measuring individual teacher's performance, especially for high-stakes decision making such as linking the measures to teachers' pay and contracts. Those behind these schemes and those who make these measures public must be aware of the consequences of their actions. I have now put it plainly so there should be no excuse for not knowing about the unreliability and invalidity of using students' test gain scores to judge teachers.

If the ranking of teachers does not stop, I will have no choice but to compile a league table of people by their proficiency level in statistical literacy. I can already identify some people at the bottom of the table. Further, I have far more confidence in identifying people with low statistical literacy than the newspaper has in identifying low performing teachers. As a final note, I hope teachers will file a class action to put abusers of statistics behind bars and the world will be a better place.



[1] http://www.nypost.com/p/news/local/ratings_of_public_school_teachers_agH4xQDbuen0uBUpEoSuDM
[2] For example, an excellent report by Darling-Hammond et al can be accessed on the internet: http://aera.net/uploadedFiles/Gov_Relations/GettingTeacherEvaluationRightBackgroundPaper%281%29.pdf

[3] See, for example, Wu, M.L. (2010). Measurement, sampling and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29(4), 15–27.*

[4] Schochet, P., & Chiang, H.S. (2010). *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains,* July 2010, NCEE. Retrieved 2010/10/26, http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf
[5] http://aera.net/uploadedFiles/Gov_Relations/GettingTeacherEvaluationRightBackgroundPaper%281%29.pdf
[6] Amrein-Beardsley & Collins (forthcoming).
[7] See, for example, http://en.wikipedia.org/wiki/Sally_Clark
[8] It's actually a little lower than this, but I will use 1 in 1000 to simplify the computation.
[9] Families with several births.
[10] If there is a large number of statistical tests being conducted, we need to use a multiple-comparison correction, such as the bonferroni correction. After applying this correction, we will not find many statistical tests significant.
[11] Technically, this is because we carry out many statistical tests in an exploratory approach.