School of Education, University of Colorado at Boulder
Boulder, CO 80309-0249
Telephone: 802-383-0058

NEPC@colorado.edu
http://nepc.colorado.edu

Due Diligence And The Evaluation Of Teachers

Fact Sheet Concerning L.A. Times Article of February 7, 2011

On February 7, 2011, the Los Angeles Times published an article titled, "Separate study confirms many Los Angeles Times findings on teacher effectiveness," and with the subtitle, "A University of Colorado review of Los Angeles Unified teacher effectiveness also raises some questions about the precision of ratings as reported in The Times."[1]

The *Times* story was written by Jason Felch, the reporter who also wrote the August 2010 *Times* story that relied on the problematic research. As is explained in the new study (http://nepc.colorado.edu/publication/due-diligence), the August publication by *Times* of teacher effectiveness ratings were based on unreliable and invalid research.

Mr. Felch's article on Monday, February 7th, is therefore incorrect in characterizing the re-analysis as "confirm[ing] the broad conclusions of a Times' analysis." In fact, the new study, by Derek Briggs and Ben Domingue and titled *Due Diligence and the Evaluation of Teachers*, confirms very few of the Times' conclusions.

We asked the study's lead author, Professor Derek Briggs, to comment on the *Times'* story. For clarity of reading, his comments are inserted below in green italics. The non-italicized text (in blue) is from the *Times* story of February 7th.

[1] Although this article starts by identifying the study as to be released on Monday, the "embargoed" copy of the report obtained by the Times states in bold red print, "NOT FOR RELEASE BEFORE 12:01 AM EST, FEBRUARY 8, 2011" – that is, it was embargoed until Tuesday morning.

A study to be released Monday confirms the broad conclusions of a Times' analysis of teacher effectiveness in the Los Angeles Unified School District while raising concerns about the precision of the ratings.

> *I don't see how one can claim as a lead that our study "confirmed the broad conclusions"—the only thing we confirmed is that when you use a value-added model to estimate teacher effects, there is significant variability in these effects. That's the one point of agreement. But where we raised major concerns was with both the validity ("accuracy") and reliability ("precision"), and our bigger focus was on the former rather than the latter. The research underlying the Times' reporting was not sufficiently accurate to allow for the ratings.*

Two education researchers at the University of Colorado at Boulder obtained the same seven years of data that The Times used in its analysis of teacher effectiveness, the basis for a series of stories and a database released in August giving rankings of about 6,000 elementary teachers, identified by name. The Times classified teachers into five equal groups, ranging from "least effective" to "most effective."

After re-analyzing the data using a somewhat different method, the Colorado researchers reached a similar general conclusion: Elementary school teachers vary widely in their ability to raise student scores on standardized tests, and that variation can be reliably estimated.

> *It is not clear here if Mr. Felch is referring to our replication or our sensitivity analysis. We didn't say anything to the effect that "variation can be reliably estimated". I don't even know what this means.*

But they also said they found evidence of imprecision in the Times analysis that could lead to the misclassification of some teachers, especially among those whose performance was about average for the district.

> *I think this statement is true, However, it is a strange point to highlight because he skipped over the heart of our study, which contrasted the LA Times model to the one we specified with additional control variables. He never mentions this.*

The authors largely confirmed The Times' findings for the teachers classified as most and least effective.

> *No, we did not, quite to the contrary. Mr. Felch seems to be again focused only on the precision issue and not on the accuracy problems that we primarily focus on in our report.*

But the authors also said that slightly more than half of all English teachers they examined could not be reliably distinguished from average. The general approach used by The Times and the Colorado researchers, known as "value added," yields estimates, not precise measures.

The Colorado analysis was based on a somewhat different pool of students and teachers from the Times analysis, a difference that might have affected some of the conclusions. The Colorado researchers began with the same dataset released to The Times, but their ultimate

analysis was based on 93,000 fewer student results and 600 fewer teachers than the analysis conducted for The Times by economist Richard Buddin.

> *He doesn't note that we used all the same decisions for what students and teachers to include (having emailed Dr. Buddin, the Times' researcher, to get this information directly) and still had this difference. That's part of the serious concerns that we raise—we were trying to replicate what Dr. Buddin did and couldn't do it. These paragraphs come across as an attempt to mischaracterize our due diligence and our conclusion that there may be an additional problem with the underlying study – an attempt to instead make us look sloppy.*

In addition, to improve the reliability of the results it reported, The Times excluded from its ratings teachers who taught 60 students or fewer over the study period. The Colorado study excluded only those teachers who taught 30 students or fewer.

> *As I pointed out to Mr. Felch in my email (full email here), this doesn't change the point we made regarding precision.*

After a Times reporter inquired about that difference, Derek Briggs, the lead researcher on the Colorado study, said in an e-mail that he had recalculated his figures using only those teachers who had taught more than 60 students. Doing so reduced the number of discrepancies, he said; but still, up to 9% of math teachers and 12% of English teachers might have ended up in different categories using Colorado's method than they did in The Times' analysis.

> *The numbers here are just wrong, and it is surprising that my email to Mr. Felch could be so blatantly misunderstood. Here is what I wrote in the email:*
>
> For classifications based on reading outcomes, we find 516 false negative and 612 false positives—10.2% and 12.1% of the total sample of teachers that would be included in ratings released by the L.A. Times. For classifications based on math outcomes, we find 257 false negatives and 454 false positives—5.0% and 8.9% of the total sample of teachers that would be included in ratings released by the L.A. Times.
>
> *This is about false negatives and false positives, and the correct numbers based on the N>60 distinction are 10.2+12.1= 22.3% of English teachers, and 5+8.9 = 13.9% of math teachers. Using the screen applied by the Times, we found a likely misclassification of 22% (reading) and 14% (math). Again, this is the precision issue, not the accuracy issue; even a precise misclassification can be based on a flawed model.*

The authors also found that the way school administrators assign students to teachers — giving especially challenging students to a certain teacher and not to others, for example — could have skewed the value-added results. But recent research by a Harvard professor using Los Angeles school data did not find that such assignments created a bias in value-added scores.

> *This is, in my view, the most misleading part of the article – even worse than the incorrect numbers mentioned above. Mr. Felch is referring here to a study by Kane & Staiger (2008). In that study, they used a model that controlled for FAR more variables than was done in Dr.*

Buddin's approach. Readers are strongly encouraged to take a look at Table 1 on p. 6 of our report. Mr. Felch seems to think that all value-added models are the same—they are not.

Buddin said that although most conclusions of the two studies were similar, the differences in data analyzed made it difficult to directly compare his results with those of the Colorado study.

This is just a red herring—trying to distract readers by implying we used fundamentally different data. This is not true. We obtained the exact same database as did Dr. Buddin, and we attempted to use that database to replicate his work. We reached very different conclusions regarding validity and reliability.

The Colorado study comes as education officials in Los Angeles and across the country are moving to incorporate more objective measures of performance into teacher evaluations. In the process, they are confronting the technical challenges involved in value-added analysis, which attempts to estimate a teacher's effect on student learning by measuring each student's year-to-year progress**.**

Developing value-added scores requires numerous judgment calls about what variables to use and how to obtain the most reliable results. Each school district that has used value-added follows slightly different methods, and supporters of the approach say it should not be used as the sole measure of a teacher's ability.

Briggs said his goal was to raise awareness about those issues. "You have an obligation to have an open conversation about the strengths and weaknesses" of the methodology, he said.

These paragraphs are fine.

Briggs' study was partly funded by the Great Lakes Center for Education Research and Practice, which is run by the heads of several Midwestern teachers unions and supported by the National Education Assn., the largest teachers union in the country.

Note from Kevin Welner, director of NEPC: Neither the Great Lakes Center nor any other funder of NEPC work has any editorial voice in our work. In addition, neither I nor any other NEPC leader had any editorial voice in the work of the researchers who authored this report, Briggs and Domingue.