

MEMORANDUM

To: Wendy Kopp, President, Teach For America

From: Paul Freedman, Assistant Professor, Department of Politics, University of Virginia

Date: September 17, 2002

Re: Laczko-Kerr-Berliner study

Introduction

This memo contains my reactions to the methods employed by Ildiko Laczko-Kerr and David Berliner, in their article, “The Effectiveness of Teach for America and Other Under-certified Teachers on Student Academic Achievement: A Case of Harmful Public Policy” (2002). I am an Assistant Professor in the Department of Politics at the University of Virginia, where I have taught graduate-level research methods since 1997. As you know, I have not been involved in any way in the debate over teacher certification. Indeed, I confess that if anything, my prior, naïve assumption is that requiring teachers to demonstrate some level of competence and to receive state certification is probably a good thing. None of the findings in the Laczko-Kerr-Berliner study, however, support this belief.

The authors claim that, compared with Teach For America (TFA) and other “under-certified” teachers, certified teachers lead to increases in student performance that correspond to two months of class time. I am unconvinced by this claim, for three reasons:

- problems of selection and inadequate matching fundamentally undermine the validity of the study;
- the authors overstate the substantive importance of their estimates;
- the statistical approach employed is not well suited to the research question.

The first concern is paramount; it trumps any other criticism that can be raised about the study, including the other two. I raise the additional points only to illustrate that even if the fundamental flaw were absent, there would still be problems with the study. Ultimately, Laczko-Kerr and Berliner’s conclusion that, “the TFA program appears to be a failure” (p.33) is unwarranted, as is their more general claim about the impact of certified versus under-certified teachers.

1. Comparing Apples to Apples? Sampling and Matching Procedures

In an ideal world, researchers could randomly assign teachers to classrooms without taking account of background, experience, skills, or demographic factors. With enough teachers, randomization would ensure that there was no correlation between a given factor – certification status, for example – and the attributes of a district, school, or classroom of students. Thus, one could make unbiased inferences about the impact of any given teacher characteristic (e.g., teacher certification) on any given outcome measure (e.g., SAT-9 scores). In the real world, of course, teachers are not randomly assigned to students. Instead, teachers are assigned in ways that are correlated with student factors, that in turn are correlated with

achievement scores. As Laczko-Kerr and Berliner themselves explain, “teachers in schools that serve the poor are often under-certified, inexperienced, and may be teaching out-of-field. Teachers who serve wealthier students overwhelmingly hold regular certification, have accumulated considerably more teaching experience, and are less often required to teach out-of-field” (p.31).

As a result, any observed differences between the achievement scores of students taught by certified teachers and those taught by “under-certified” teachers are very likely due to pre-existing differences between the two groups. Valid estimates of the impact of any given teacher attribute (including certification status) would require careful steps to ensure that comparisons are being made between teachers in comparable classroom contexts (i.e., teaching students who are as close to identical as possible). It is for this reason that the authors attempt a matching procedure, which ideally would enable them to compare comparable teachers. *The entire Laczko-Kerr and Berliner study rests on the success of this matching procedure, which is ultimately inadequate.*

There are at least three reasons why the authors’ matching procedure and selection process fail.

- First, the authors make unrealistic assumptions about the distribution of teachers within districts and schools. “It is assumed,” the authors write, “that teachers in the same school teach similar students, an imperfect but reasonable assumption” (p.19). This assumption is clearly imperfect, but it is not clearly reasonable. “Under-certified” teachers in general, and TFA teachers in particular, may be more likely than their certified counterparts to end up in tougher classrooms within a given school. The authors make a similar assumption about the comparability of schools within a given district, “since Arizona school district boundaries are based on relatively homogenous geographic areas” (p.19). “The assignment of teachers to schools, and to classrooms within schools,” the authors conclude, “appears to have been unbiased. Similarly, we have no reason to believe that class size or student ability was different in any way for the certified or under-certified teachers in our sample” (p.19).

These are all critical assumptions. If “under-certified” teachers tend to teach in under-performing districts, in worse schools within a given district, or worse students within a given school, then we have a serious problem of endogeneity: pre-existing differences among classrooms, schools, and districts contribute to the assignment of teachers to students. Thus, *to the extent that TFA and other “under-certified” teachers are assigned to lower-performing schools within districts and lower-performing classrooms within schools, the matching procedure has failed and the validity of the study is called into question.* If TFA teachers end up in tougher schools and tougher classrooms than certified teachers, we have reason to question this study.

- Second, the authors’ own data, properly interpreted, call into question the “sameness” of the schools and districts on which they have attempted to match teachers. They report a series of ANOVA analyses, designed to demonstrate that there are no

significant differences in SAT-9 scores by school and district. These analyses are intended to address the concern that there are underlying differences among schools and districts that could serve to bias subsequent analyses. Such concerns, however, appear well-placed: Table 2 reports findings for “school sameness.” The authors claim that there are significant differences by school only for the 1999 reading scores. This is true only if one adheres to a strict $p < .05$ criterion for significance. If one relaxes the criterion to $p < .10$, significant differences emerge for all three sets of 1999 scores.¹ With respect to “district sameness” (Table 3) the authors themselves admit that, “the procedures we used to match teachers across districts were not faultless” (p.24). ***These results indicate that the matching procedures employed by the authors were inadequate.***

- Finally, it is important to note the process by which districts were selected into the study. Thirty-six districts were invited by the authors to participate (24 in 1998-99, 12 in 1999-2000). Of these, only five agreed to participate. Such a low cooperation rate (13.9 percent) is not necessarily a cause for concern, unless one has reason to believe that the districts self-selecting into the sample are not representative of all districts statewide. This is most likely the case (participating districts were, for example, urban with large minority populations), undermining the generalizability of the study to the entire state. Of greater concern, however, is the possibility that these districts are not representative of the population of districts *with similar populations*. Because the sample of districts was essentially self-selected, it is possible that these districts had particular difficulties with under-certified teachers, making them more likely to participate in a study about teacher certification.² ***To the extent that this is the case, the five districts in the sample are not representative even of similar urban, largely minority districts, and the findings from the study should not be generalized.***

The result of their sampling and matching strategies is an initial sample of 293 teachers that is reduced (by throwing out “unmatched” cases) to 218 (109 certified, 109 under-certified). Included in the under-certified group are 29 TFA teachers.³ Given the problems enumerated above, these 218 teachers provide no insight into the relative effects of certified versus under-certified teachers on student performance.

¹ Doing so is equivalent to accepting a 90 percent rather than a 95 percent confidence level, akin to betting on a horse that had a 90 rather than a 95 percent chance of winning.

² In fairness, this concern does not apply if the purpose of the study was hidden from district personnel responsible for deciding whether or not to participate.

³ These 29 cases are the basis for all subsequent analysis and conclusions about the impact of the TFA program *per se*. Where the authors do consider TFA teachers (e.g., Table 10), the findings are subject to the concerns raised in points 1 and 2 of this memo.

2. How Large is Large? The Substantive Significance of the Estimated Effects

Putting aside, for a moment, the serious concerns about the sampling and matching methodology, one can question the substantive claims that the authors make on their own terms. The authors make a strong assertion: “Students of under-certified teachers,” they write, “make about 20% less academic growth per year than do students of teachers with regular certification” (p.2). But the data do not support such a strong claim: First, the authors provide six sets of differences between SAT-9 NCE scores (reading, math, and language for 1998-99 and for 1999-2000) for students of certified and under-certified teachers. These differences are reported in terms of *effect size* (i.e., the difference in mean NCE score divided by the NCE standard deviation of 21.06). The author’s base their “20% less academic growth” claim on an effect size of .20. However, the average effect size for these differences for the full sample of teachers is .18, and only two of the six estimated effects sizes are greater than .20.⁴ ***Indeed, it is only by incorporating additional analyses that exclude 7th and 8th grade teachers (thereby reducing the sample size) that the authors can reach the .20 effect size, the basis for the “two-month’s growth” claim.***

There is, moreover, a second, more important reason to regard the “two months’ growth” claim with caution. One must first be clear about what the findings indicate. The *largest* estimated difference found is the six NCE points for reading during the 1998-99 period. Because these are NCE scores, the impact of certified teachers in this case is six points on a scale that ranges from 1 to 99 and has a standard deviation of 21.06. It is tempting to consider an effect of this size as relatively small in an absolute sense. Indeed, in his classic text on effect size, Cohen regards effect sizes in this range (.20 and below) as representing a “small difference between means” (1988, p.25). ***How one interprets the alleged gains from certification reported here, then, will depend on how substantively large one considers a six-point (or smaller) rise in NCE scores to be, and there may be good reasons to consider it modest.***⁵

Where, then, do the authors get their two-month estimate? They appear to base this on the work of Glass (2002), who, in a footnote, asserts, “It is an empirical fact that the standard deviation of most achievement tests is 1.0 years in *grade equivalent units*” (p.8.27, emphasis in original). If this assertion (which is unsupported by citation or evidence) does in fact apply to the specific relationship between NCE scores and grade equivalent scores, the authors’ claim has some basis (putting aside the fundamental flaws identified in the first section of this memo). But if we suspect that one NCE standard deviation does not neatly correspond to one academic year, we would arrive at a different conclusion. ***Thus the two-month claim hinges upon a particular assumption about the relationship between the standard deviation of NCE scores and grade equivalent units.***

⁴ As Table 8 (p.28) reveals (to the authors’ credit), the estimated effect sizes range significantly from .09 for language scores in 1999-2000 to .28 for reading in 1998-99.

⁵ The Northwest Regional Educational Laboratory, for example, suggests that as a rule of thumb, a difference of *seven* NCEs is needed for an effect “to be considered to have practical importance” (Yap, et. al., 2000, p.76).

3. Choice of Analytic Method

Finally, the research question for the present study concerns the impact of a particular teacher characteristic (certification status) on student performance. Because the dependent variable is an individual-level outcome, the authors could have undertaken a different (and more useful) strategy of modeling individual student SAT-9 scores as a function of a range of explanatory variables. Such variables could include student-, school-, and district-level characteristics, along with teacher attributes such as certification status. As the authors themselves note, there is a range of multivariate modeling techniques (including Hierarchical Linear Models) appropriate to the task. Undertaking even slightly more sophisticated techniques could have provided greater analytic purchase. Additionally, the present research would be improved by the collection and analysis of individual-level longitudinal data.

Conclusion

The authors make a series of strong claims about the impact of under-certified teachers in general and TFA teachers in particular. ***Given the limitations of the research design, most notably the inadequacy of the matching procedure, these claims are not supported by the findings of the present study.***

References

- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences* (Second edition). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Glass, Gene V. 2002. "Teacher Characteristics." In A. Molnar (ed.) *School Reform Proposals: the Research Evidence* (chapter 8). Retrieved 9/13/02 from: www.asu.edu/educ/eps1/EPRU/documents/EPRU%202002-101/Chapter%2008-Glass-Final.pdf
- Laczko-Kerr, Ildiko and David.C Berliner. 2002. "The Effectiveness of Teach for America and Other Under-certified Teachers on Student Academic Achievement: A Case of Harmful Public Policy" *Education Policy Analysis Archives*, 10(37). Retrieved 9/12/02 from: <http://epaa.asu.edu/epaa/v10n37/>
- Yap, Kim, Aldersebaes, Inge, Railsback, Jennifer, Shaughnessy, Joan and Timothy Speth. 2000. *Evaluating Whole-School Reform Efforts: A Guide for District and School Staff* (Second edition). The Northwest Regional Educational Laboratory.