REL Technical Brief

REL 2008 – No. 001

A review of *Avoidable losses: high stakes accountability and the dropout crisis*

Prepared by

C. Wilkins
Edvance Research

February 2008

REL Technical Briefs is a new report series from Fast Response Projects that helps educators obtain evidence-based answers to their specific requests for information on pressing education issues. Technical Briefs offer highly targeted responses across a variety of subjects, from reviews of particular studies or groups of studies on No Child Left Behind Act implementation issues, to compilations or quick summaries of state or local education agency data, appraisals of particular instruments or tools, and very short updates of Issues & Answers reports. All REL Technical Briefs meet IES standards for scientifically valid research.

February 2008

REL Southwest received a request to review the report *Avoidable Losses: High Stakes Accountability and the Dropout Crisis* to assess the soundness of the study methodology and the appropriateness of the conclusions drawn in the report.

This REL Technical Brief was prepared for IES under Contract ED-06-CO-0017 by Regional Educational Laboratory Southwest administered by Edvance Research. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Wilkins, C. (2008). *A Review of "Avoidable Losses: High Stakes Accountability and the Dropout Crisis"* (REL Technical Brief, REL 2008–No. 001). Washington, DC: U. S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs

This REL Technical Brief is available on the regional educational laboratory website at http://ies.ed.gov/ncee/edlabs.

Abstract from the study reviewed

In the state of Texas, whose standardized, high-stakes, test-based accountability system became the model for the nation's most comprehensive federal education policy, more than 135,000 youth are lost from the state's high schools every year. Dropout rates are highest for African American and Latino youth, more than 60% for the students we followed. Findings from this study, which included analysis of the accountability policy in operation in high-poverty high schools in a major urban district, analysis of student-level data for more than 271,000 students in that district over a seven-year period under this policy, and extensive ethnographic analysis of life in schools under the policy, show that the state's high-stakes accountability system has a direct impact on the severity of the dropout problem. The study carries great significance for national education policy because its findings show that disaggregation of student scores by race does not lead to greater equity, but in fact puts our most vulnerable youth, the poor, the English language learners, and African American and Latino children, at risk of being pushed out of their schools so the school ratings can show "measurable improvement." High-stakes, test-based accountability leads not to equitable educational possibilities for youth, but to avoidable losses of these students from our schools.

Citation for the study reviewed: McNeil, L. M., Coppola, E., Radigan, J., & Vasquez Heilig, J. (2008). Avoidable losses: High-stakes accountability and the dropout crisis. *Education Policy Analysis Archives, 16*(3). Retrieved February 20, 2008, from http://epaa.asu.edu/epaa/v16n3/

Summary of the review

The authors have made strong causal conclusions about the effect of Texas's test-based accountability system on the high school dropout rate: that the accountability system directly increases dropout rates throughout the state. Given the nature of the data collected and analyzed in this study, such conclusions cannot be scientifically validated.

REL Southwest received a request to review the report *Avoidable Losses: High Stakes Accountability and the Dropout Crisis* to assess the soundness of the study methodology and the appropriateness of the conclusions drawn in the report.

The review found that conclusions drawn in this study cannot be generalized and are greatly overstated.

- This study was conducted in one school district in one city in the state of Texas. The in-depth ethnography that makes up a large portion of the study was conducted in a single high school in that district.
- Generalizing from that one district to all 1,090 school districts in Texas is invalid.
- The authors assume that because Texas's accountability system was a model for the No Child Left Behind Act of 2001 their conclusions have "great significance for national education policy." They claim to have shown that "disaggregation of student scores by race . . . puts [minority students] at risk of being pushed out of their schools." Even if the study demonstrated this finding for one district, such broad generalization is unwarranted.

The causal claims by the authors do not meet the standards for scientific rigor, even in the single district in which the study was conducted.

- Many of the data in this study were collected through ethnographic methods, essentially unstructured surveys designed to help discover causal hypotheses that could be scientifically tested in later experimental studies. Claiming that such data have "explanatory power" (p. 11) is not justified scientifically.
- An "in-depth" ethnography of a single high school may yield detailed insights. But no matter how in-depth the observations, a single case study simply cannot be used to draw causal conclusions and "verify" that a "positive answer" exists to the question whether waiver policy caused minority students to drop out of school (p. 19).
- A statistical analysis of data was conducted as part of this study. But simply doing a statistical analysis does not lead to causal conclusions unless the study was designed to collect data that can be used in such a manner (as in a randomized controlled trial). That was not the case in this study. The longitudinal data analyzed can be used only to make correlational inferences, not causal ones, no matter what statistical techniques are used.

Review of the study

Avoidable Losses: High-Stakes Accountability and the Dropout Crisis, by McNeil, et al., is an examination of various educational policies and dropout rates in Texas. Problems were identified with the methodology, data analysis and interpretation, and conclusions in this study. This review is limited to an analysis of the study's methodology and conclusions.

Many of the conclusions drawn by the authors cannot be justified scientifically given the type of data collected and analyzed. The primary conclusions the authors draw is that Texas's "high-stakes accountability system has a direct impact on the severity of the dropout problem." They further claim that this study has "great significance for national educational policy" and that in general, "high-stakes, test-based accountability leads to . . . avoidable losses of . . . students from our schools."

We evaluated the appropriateness of the study's conclusions for both external validity and internal validity. External validity, or generalizability, refers to whether any valid causal relations demonstrated in the study can be inferred to hold across a broader spectrum of conditions not specifically examined in the study. Internal validity refers to whether scientifically valid conclusions can be drawn about the causal relationship between two variables for the specific conditions directly examined in a study. We look at each of these separately.

External validity (generalizability)

The authors make some strong claims about the generalizability of the study's results. The study was conducted entirely in one school district in one city in Texas. Portions of the study were conducted in a single high school in that district. Yet the authors make broad generalizations about the entire state. Even if all the conclusions the authors make were valid for this district, it is not scientifically valid to generalize to the other 1,089 school districts in Texas based on the results from one district.[1] Such generalizability would require data from a sufficiently large number of randomly sampled, or demonstrably representative, districts from the population to be generalized to. This district cannot be considered as representative of all the districts in Texas. Nor was this district selected at random. And even if it were, it is essentially, as the authors admit, a single case study (in other words, a sample of districts of size n=1).

While an in-depth case study can provide much information about the district in question, it does not change the fact that it is a single district, and the results from a single district do not achieve any additional generalizability from being "in-depth."

The authors also contend that the Texas accountability system they examine became the model for the No Child Left Behind Act of 2001 (NCLB) and attempt to use the results of this study to draw conclusions about NCLB. But NCLB encompasses 50 states with 50 different accountability systems (50 state standards, 50 state tests, and so on). To draw causal conclusions about all of NCLB across 50 states based on results from one district in Texas does not become scientifically valid simply because the Texas accountability system was part of the inspiration for NCLB.

We evaluated the appropriateness of the study's conclusions for both external validity and internal validity

Internal validity

Some conclusions drawn in this study are of questionable validity even in the single district. The rest of this review examines which conclusions are valid given the nature of the data collected and analyzed.

This study collected and examined both qualitative and quantitative data. The types of causal conclusions that can potentially be

1. Texas Education Agency, October 2007, reports 1,090 independent school districts and active charter school districts with assigned school district numbers.

drawn for these two types of data are very different, so we examine them one at a time.

Qualitative data

This study was conducted primarily by collecting and analyzing qualitative data, which the authors describe as falling into three categories:

1. An ethnographic study of urban high schools in one school district.
2. School-site interviews and observations in urban high schools in the district.
3. An in-depth ethnography of one high-poverty, mostly Latino high school within the district.

The study was conducted primarily by collecting and analyzing qualitative data

The district examined in the study is not identified for reasons of privacy, but it is identified as a large urban district in Texas.

Most of the qualitative data were collected through ethnographic methods, so we first describe the ethnographic methodology and appropriate uses of the resulting data before describing the specific data in the study.

An ethnography is a qualitative technique, defined in Shadish, Cook, and Campbell (2002) as an "unstructured exploratory investigation, usually of a small number of cases, of the meaning and functions of human action, reported primarily in narrative form." They note that "such explorations of explanation help identify hypotheses about possible causal contingencies to be tested."

Essentially, an ethnography allows the ethnographer to examine a situation in great detail and come up with insights about possible causal effects that might not emerge without such an intensive examination. However, no causal conclusions can be made from an ethnography, which is qualitative and exploratory. It allows one to form hypotheses about causal relationships, but those hypotheses must be tested and conclusions drawn using other methodologies, namely an experimental study.

1. ***Ethnographic study of urban high schools in the district.*** The study is not entirely clear on this category of data, but it appears that the researchers refer to "our ethnographic investigation leading up to the present study" rather than a part of the current study. This is summarized briefly as having "suggested that retention of large numbers of 9th graders . . . had its basis in the accountability system." These earlier results became the motivation for the primary question addressed by the study: "Is there a connection between high-stakes accountability and high numbers of dropouts?"

   The use of previous ethnographic results to form the key question to examine is entirely appropriate and exactly what an ethnography is designed to do.

2. ***School-site interviews and observations of urban high schools in the district.*** Seven high schools in the district were selected, and interviews were conducted with teachers, administrators, and students. It is not clear in the paper whether these high schools are the same as in the earlier ethnographic study. There is no clear statement about selection, but it does not appear that the schools were randomly selected from high schools in the district. The interviews and focus groups were "approximately one hour long, taped and transcribed, and then coded for consistent themes," an appropriate methodology for collecting such data. The results of these interviews and observations are reported in a purely qualitative fashion, with resulting conclusions. In keeping with the ethnographic approach, interviews were unstructured (for example, "during a chance conversation with another principal in 1997 . . ."). Some outside quantitative data are incorporated in the description of the interviews. There are no qualitative data from or about the interviews themselves.[2]

2. These data are not quantitative summaries of interviews, but school-level statistics from another study.

3. ***In-depth ethnography of one high school.*** One high school was selected for an in-depth ethnography. The high school was not selected at random, but was purposely selected because it "fit the pattern of rising ratings and rising dropouts" of interest to the researchers. A large part of the paper is devoted to the detailed description of this ethnography. The study is infused with other information (outside data, interviews from other schools, and so on), and conclusions are drawn throughout. This approach is in keeping with the stated intent of the authors to "triangulate results" and to look at "interactions"[3] between the different phases of the study. A mixture of both qualitative and quantitative data is presented at this point, but the quantitative data are only descriptive, not analyzed statistically.

How the results of the in-depth ethnography are used appears to be inappropriate. The authors do not seem to use the in-depth ethnography to further develop or refine hypotheses. Instead, they attempt to draw causal conclusions from the results.

For example, the authors talk about the "explanatory power" of an ethnography and claim that the in-depth ethnographic single case study can "verify" that a "positive answer" exists to the question of whether waiver policy caused minority students to drop out of school. Later they claim that the results "demonstrate that the accountability system aggravates the high dropout rate." There are many other instances where causal conclusions are drawn or implied from ethnographic data.

Simply put, such causal conclusions from this type of data are not scientifically justified, and they represent an inappropriate use of such data.

Quantitative data

This study uses quantitative data in two ways. First, various graphical representations of descriptive data about aspects of the district are incorporated in the discussion of the ethnographic results. While this use is perfectly legitimate, no statistical analyses were conducted on the data, and (as discussed below) the data were not collected in a manner that would allow any causal conclusions to be validly drawn. As mentioned, the authors attempt to draw causal conclusions from the ethnographic data. The inclusion of the descriptive quantitative data does not make such conclusions legitimate.

The second way quantitative data are used is in a statistical analysis of what the authors describe as "student-level" data in the district over a seven-year period.

The data were not collected in a manner that would allow any causal conclusions to be validly drawn

This part of the study examines the relationship between changes in Texas Education Agency (TEA) school ratings and various changes in "student progression, demographics, and teacher capacity." It began with a longitudinal dataset of 271,000 students, but that dataset was aggregated into school-level data before any analyses were done. It is therefore questionable whether this analysis should be considered an analysis of "student-level" data,[4] as opposed to school-level data, including student characteristics.

The data are analyzed using a multinomial logistic regression. One point of confusion arises because the study is inconsistent in describing the reference group used for the analysis. At one point it says that "a decrease in TEA school rating [is] used as the reference group" and later that "the reference category . . . denotes schools that remained at the same accountability rating." It appears in table 3 of the study that the latter is done in the analysis. As noted by the authors, the dependent variable is ordinal, but it does not appear that

3. The term is used in a nonstatistical sense.

4. If the data really were student level, it would be important to use hierarchical linear modeling to account for the clustering of students within classrooms and schools, which was not done. But because the data are actually school level, and all the schools are part of one district, those analyses are not necessary.

an ordered logistic regression was used. That approach would be technically preferable to an unordered multinomial logistic regression.

Aspects of the analysis and the corresponding conclusions are not clear. The interpretation of the odds ratio speaks of the "odds *of* a one-percent increase in 9th grade retention" (emphasis added). It appears that the authors probably mean the odds *due* to a one-percent increase. However, while the results are not as clearly explicated as they could be, it does appear from examining table 3 that an increase in 9th grade retention rate is associated with a significantly higher likelihood that a school will increase its school rating rather than maintain the current rating. But interpreting this result is complicated because many schools are already rated at the highest level and cannot move up. Given the nature of the data collected, it is also not possible to conclude that changes in 9th grade retention rates caused changes in TEA school ratings.

The point here is that there is an important distinction between valid inferential conclusions that can be drawn and invalid causal conclusions that cannot be drawn in this case

The point here is that there is an important distinction between valid inferential conclusions that can be drawn and invalid causal conclusions that cannot be drawn in this case.

There are two issues—the nature of the data being analyzed and the type of analysis being conducted. We next review the relationship between these issues and the subsequent conclusions that can be validly made.

To make any valid inferential statistical statements about a sample, that sample must be representative of the population of interest. This goal is often accomplished through some sort of random sampling. In this study the sample was not randomly selected, and there is a fairly brief exposition of the selection of schools. It seems reasonable that they would be relatively representative of the high schools in the district, though they may be representative of only a particular subpopulation (for example, only urban schools).[5]

While a representative sample allows statistical inferences about the population, the nature of the data can limit those statistical inferences. In the relations between two or more variables, inferences can be either correlational or causal, depending on the nature of the data.

Correlational inferences can almost always be made from a representative sample.[6] This type of inference shows that two variables are related, or co-vary. So, for example, it might be possible to show that higher levels of teacher satisfaction tend to go with higher test scores. But there would be no way of knowing whether higher teacher satisfaction caused higher test scores, whether higher test scores caused higher teacher satisfaction, or whether some other factor caused both higher teacher satisfaction and higher test scores.

Showing a causal relationship (for example, smaller class sizes cause both higher teacher satisfaction and higher test scores) is much more difficult. The details are beyond the scope of this review (see, for example, Morgan & Winship, 2007). The primary method is to conduct an experiment in which subjects are randomly assigned to treatment or control conditions. Such an experiment in education is known as a randomized controlled trial. For example, researchers might randomly give half the students special help and then see if those students get better test scores than those who did not get the help. If done properly, that research could support valid conclusions such as "the special help led to higher test scores."

Consider again the primary research question of the study: "Is there a connection between high-stakes accountability and high numbers of dropouts?" As stated, this question is correlational. The conclusions drawn and implied in the study, however, are causal, asserting that high-stakes accountability causes higher

5. For example, there is no way to know whether the "urban" district selected for this study is 100% urban or only "mostly" urban given the privacy concerns.

6. Correlational inferences might not be possible if the data on one variable were collected at a different time under very different circumstances than those on another variable.

numbers of dropouts. This question would, admittedly, be difficult to test experimentally. Under some conditions carefully designed quasi experiments, instead of true experiments, might be needed to address causality. But the data collected and examined in this study, no matter how they are looked at, cannot be used to draw causal conclusions, though they could certainly provide important causal hypotheses that could be addressed in another study.

Finally, the statistical analyses themselves must be considered. The conclusions that can be drawn depend largely on the statistical analyses conducted. But no statistical analyses can lead to valid causal conclusions without appropriate data to test such hypotheses. With a different type of data collection, the analyses conducted in the study could have been appropriate to show causal relationships. Given the data available, however, no statistical analyses could yield scientifically legitimate causal claims.

Unsupported assumptions

We discuss one further issue related to the interpretation of the results from this study: avoidable losses.

The title of the study alludes to the conclusion that the losses (high school dropouts) would be avoidable if the accountability system did not exist. Such a conclusion is based on assumptions that are not supported by scientific rigor.

It is possible to examine the type of data available in this study and to draw such conclusions as "9th grade retention rates are positively correlated with dropout rates." But to draw that conclusion, the authors would have to make three assumptions:

- 9th grade retention rates cause higher dropout rates.
- The accountability system causes higher 9th grade retention rates.
- If the accountability system were not in place, students currently retained in 9th grade would not be retained and would not dropout. Hence, they are avoidable losses.

Some of these assumptions may be true, and they can be argued in a political realm. But given the lack of scientific rigor, none of them can be proven in the current study. So, the conclusion—that these dropouts would not occur in the absence of the 9th grade waiver and are therefore avoidable losses through some unspecified alternative mechanism to the accountability system—is scientifically unwarranted.

The conclusion that high school dropouts would be avoidable if the accountability system did not exist is based on assumptions that are not supported by scientific rigor

References

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge: Cambridge University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

* * *

REL Southwest at Edvance Research is one of 10 educational laboratories in the Regional Educational Laboratory Network (REL Network), under the Institute of Education Sciences (IES) of the U.S. Department of Education. REL Southwest serves the states of Arkansas, Louisiana, New Mexico, Oklahoma, and Texas and works for the benefit of over 7 million students and 500,000 teachers in approximately 14,500 schools in pre-kindergarten through college in this five-state region. The REL Network encompasses 10 geographical regions that span the nation, with the primary mission to serve the educational needs of designated regions. The REL Network uses applied research, development, dissemination, training, and technical assistance to bring the latest and best research and proven practices into school improvement efforts. For more information, see http://edlabs.ed.gov/RELSouthwest.