

## Comments on Papers by Elaine Allensworth and Rob Warren

### Workshop on Improved Measurement of High School Dropout and Completion Rates: Expert Guidance on Next Steps for Research and Policy

Aaron M. Pallas  
Teachers College, Columbia University

October, 2008

It is an odd sensation to come back to a literature that you haven't stayed up with very closely. More than two decades ago, when I first began worrying about the calculation of high school dropout and graduation statistics, we had nagging suspicions about the reasons why various statistics differed from one another, and a vague understanding of the limited utility of different measures. Now, thanks to the work of many of the people in this room, including Rob and Elaine, we have a much more sophisticated understanding of why the available statistics don't tell us what we want to know. The most positive spin on the relationship between research and policy is Carol Weiss' enlightenment model, in which research gradually sensitizes policymakers about how to reframe or think about policy problems and policy alternatives. I suppose that the literature on high school dropout and completion statistics might be characterized as conforming to a bewilderment model of the relationship between research and policy. That's because we think we know what we want in the way of measures of high school dropout and completion, but research has shown us that we don't know how to get there.

And nothing I say this afternoon is going to alter that disappointing fact. The calculation of high school dropout and graduation statistics is at once a political, technical and social process. In my view, the political questions come first: what do we want to know about dropout and graduation statistics, and why? The technical questions come second: Conditional on what we want to know, what are the technical requirements for a data system that can generate the desired statistics? The social questions come third: Given the technical requirements, what social forces can actually populate the data system with data that are accurate and timely? These questions recognize that the production of dropout and completion data is a social process, a point made explicit in Elaine's account of what clerical personnel might actually have to do to produce useful data.

I'm going to highlight just a few themes. The first is that the two major uses of dropout and completion statistics, as indicators and accountability mechanisms, have very different data demands. The second is that it may be unwise to privilege administrative records over self-reports. And the third is that I think it's high time that we start designing information systems that assume that students are mobile, rather than constantly being surprised by student mobility.

Indicators and Accountability. As we've heard, dropout and completion statistics are used both as education indicators, and as the drivers in complex performance accountability systems, including No Child Left Behind. Education indicators are like a social thermometer, providing broad, but non-specific information about the status and direction of various features of the education system. When a thermometer shows an elevated temperature, we know something is wrong, but we're not immediately sure of what that is; rather, running a temperature suggests

that something is going on that warrants further investigation. Education and other social indicators operate much the same way.

I'm a lot more comfortable using indicators to assess the direction of system performance than to compare performance across jurisdictions such as school districts and states. Perhaps I've seen too many instances where variations in the technology of measurement across sites create more confusion than clarity regarding the health of a system. The medical analogue is instructive. There's a standard set of panels that are done to analyze blood, and when a doctor sends your blood to the lab for analysis, she or he receives a report that typically includes both the value that the lab found, and a reference range, often conditioned on sex and age, for that value. It was a bit of a shock to me the first time I saw that two different labs had different reference ranges for the same test. A given red blood cell count could be judged as outside the reference range, and thus an indicator of a potential problem, or inside the reference range, depending on the lab. Similarly, doctors prefer that their female patients have mammograms done on the same equipment with the same radiologist from one year to the next, so that changes from one year to the next can be interpreted more easily as evidence of a change in the underlying condition, rather than changes in the measurement technology. In the medical world, there's more comfort in assessing changes over time within a person than in making comparisons across people.

I feel the same way about high school dropout and completion statistics as education indicators. They're fine as a source of evidence about trends within a jurisdiction such as a district or a state, but because of variation in the measurement technology across jurisdictions, I'm hesitant to use such statistics to compare the relative performance or health of one state versus another. If my scale at home is consistently showing that my weight is five pounds too low, it won't be very useful in comparing my weight to my brother's weight, based on his scale in his California house. (Not that it's a competition.) But even if my scale is consistently underestimating my weight, I can use what the scale says over time to judge whether my weight is going in the desired direction, which would be down.

Thus, I'm most concerned about those sources of measurement error in high school completion and dropout statistics that affect the quality of inferences about changes over time in the status of completion or dropping out within a jurisdiction. If the measurement error is primarily owing to variations across states in measurement technology, I'm inclined to suggest simply not using a particular statistic to make comparisons across jurisdictions, and rather to focus on changes over time within a jurisdiction. Elaine does show that even within a jurisdiction such as Chicago there can be discontinuities, such as a promotion or retention policy that produces a sharp change in the rate of grade retention, but generally a data series is disrupted for just one data point, and thereafter the data series will be measured consistently.

I suppose I need to say something critical about these papers, so let me take issue with Rob's contention that high school completion rates can be useful in describing the level of human capital in some meaningful jurisdiction—a school district, or a state, for example. I think a U.S. high school diploma is a really crummy measure of human capital, because the knowledge, skills and orientations that a high school diploma represents vary so widely from one state to the next. We all know that a high school diploma does not signify that a young person knows anything in particular. Moreover, if State A and State B both require a student to obtain a passing score on a high-stakes exit exam in order to get a high school diploma, and the exams differ in the academic

proficiency needed to pass the exam, then the possession of a high school diploma in State A could represent a very different level of human capital than the possession of a high school diploma in State B. And we already know that some high school completion credentials, such as a GED credential, signify lower quantities of human capital than a “regular” high school diploma.

I was hoping to show this with data from U.S. states, but couldn’t pull together the data that I think are out there. So I’ll have to make do with data from the Adult Literacy and Life Skills study, an international assessment of adult literacy in six countries: Canada, Switzerland, Italy, Norway, Bermuda, and the U.S. (I’m sure the rationale for these countries is obvious.) Among adults with a terminal secondary school credential, the average total literacy—prose literacy, document literacy, and numeracy combined—differed substantially from one country to the next. Adult literacy among terminal secondary school graduates was about half a standard deviation higher in Norway than in Bermuda and the U.S., and the gap was nearly .8 standard deviations between total literacy in Norway and in Italy. Not the ideal comparison for the point I want to make about the ambiguous meaning of a high school credential in human capital terms, but I hope you get the idea.

The use of high school completion and dropout rates for accountability purposes differs from the use of such rates as education indicators in important ways. The most important is that accountability typically involves well-defined rewards and sanctions, whereas indicators do not. Jurisdictions have an inherent incentive to manipulate accountability data to obtain rewards and avoid sanctions, whereas they do not have such incentives to manipulate data used as indicators of the status or direction of the system. There is both anecdotal and systematic evidence of the distortion of completion and dropout rates, just as there is parallel evidence regarding standardized test scores which are components of accountability systems.

One strategy to minimize the likelihood of distortion is to have the relevant data processed by individuals and organizations who have little at stake in the accountability outcomes. School and district personnel may feel substantial pressure to produce particular outcomes, whereas state personnel may not. Thus, shifting the locus of data collection, processing and reporting from the local level to the state level may improve the quality of the data. Moreover, a state clearinghouse or data warehouse may be able to compensate for the limited capacity for data processing that may exist at the local level. There are tradeoffs that cannot be ignored here, however; tracing students may be more successful at the local level than from a centralized state agency.

The future of accountability measures, in my opinion, lies in model-based measures of organizational performance. For reasons that are still not entirely clear, accountability measures in education have evolved differently than accountability measures in medicine, which moved much earlier to model-based measures of the performance of hospitals and specialty physicians. The critical move was to adjust outcomes for “risk”—the mix of patient inputs that was seen as predetermined relative to a stay in a hospital or treatment by a particular physician. The reasoning was that hospitals and physicians should not be penalized in their performance measures for treating patients whose overall health status placed them at high risk of an adverse outcome, regardless of the quality of treatment the hospital and/or physician might administer. The use of unadjusted measures, it was feared, would create disincentives for doctors and hospitals to take on cases with a higher risk of mortality.

Risk-adjusted models of organizational performance require longitudinal data, as the key analytic question is what outcomes are observed conditioned on the characteristics of a client prior to exposure to treatment. Hospitals have been able to provide the data such models need primarily because of their billing and claims practices.

Ultimately, for school accountability measures, I think we need to move towards a system in which the basic unit is a student's spell of enrollment in a particular school, at the end of which a student may either graduate or be discharged, with an array of discharge codes, including transfer and dropout. At the beginning of every spell, a student is at risk of—or eligible for, if you prefer—high school completion. The duration of a spell is particularly important. A student who enrolls in a school, and after two months transfers to another school or drops out, might be weighted differently than one who transfers or drops out after three years of exposure to the school.

The data demands of such models are extensive. We need precise information on the onset and end of an enrollment spell, for each spell, and there need to be clear and consistent rules about what counts as enrollment.

One of the challenges of model-based measures of school performance, whether we consider test scores or high school completion as the outcome of interest, is that they're so darned complicated. It's extremely difficult to explain to policymakers and the public the innards of the models and how they work.

Administrative data and self-reports I have a confession to make. Twenty-four years ago this month, I completed my doctoral dissertation, which used the Base Year and First Follow-Up of the High School and Beyond study to look at the determinants of dropping out of high school. (I say "completed" rather than "defended" because at the time Johns Hopkins had an unusual system in which doctoral students could either formally defend their dissertation proposal or the dissertation itself, and I opted for the preliminary defense, reasoning that I could actually make substantive changes at that point.) Ever since, I've been listing the date I received my degree as 1984. A few years ago, I learned that the Department and University had recorded me as a 1985 graduate. So there was a discrepancy between my self-report and the administrative records.

How did this arise? Part of the explanation is that I was moving into a postdoctoral research position at Johns Hopkins, and my contract specified that my salary would increase by about 20% upon completion of the degree. That was a powerful motivation to finish, and as of November, 1984, my title and salary were in fact adjusted. So in that sense the University had formally acknowledged that I had completed the degree in 1984. There also was some social desirability at play. I had entered graduate school in 1979, and it was more palatable to have taken five years to finish than six years. Moreover, as I went on the job market that fall, it was clearly preferable to declare that I had my Ph.D. in hand than to say that it would be awarded the following spring.

So what's the point here—besides exposing my slipperiness? I suppose one point is that it's to demonstrate that gaps between self-reports and administrative records are ubiquitous and inevitable. If we can't trust Ph.D.'s to accurately report the timing of their degree completion, can we count on high school graduates? A second point is that the notion of "accuracy" is a bit

problematic. There is a tendency to privilege administrative records over self-reports as the arbiter of what is “accurate” about the reporting of the receipt of an educational credential. Social desirability aside, in my case my institution did recognize the receipt of the degree as having occurred in 1984, despite the formal recording of the degree as occurring in 1985. For this reason, I felt justified in describing myself as a 1984 Ph.D. graduate of the University. Was this “inaccurate?” Experiences that convey meaning about the timing and/or receipt of a credential may legitimately shape self-reports in unpredictable ways. For this reason, I’m skeptical that audit studies that compare self-reports to administrative records will ever be definitive in characterizing the magnitude and direction of the differences between them, and whether those differences ought to be characterized as “bias.” And finally, the gap between administrative records and self-reports is likely to be most evident around the timing of events such as high school graduation rather than whether or not a credential was awarded. Here too there is some ambiguity about the notion of “accuracy,” but a general design principle might be to seek to construct measures that minimize dependence on the self-reported timing of events. One-year dropout rates are particularly vulnerable to errors in timing; completion rates that extend beyond the “on-time” time horizon are probably the least vulnerable.

Designing information systems that assume that students are mobile Imagine an information system for a public hospital serving people without health insurance. On October 1<sup>st</sup> of each year, the hospital reports the number of patients in beds, and maybe their diagnoses (or maybe not). Based on this headcount, the state pays the hospital a sum that is intended to cover its operating costs. Sounds ridiculous, doesn’t it? In a typical hospital, the stays are relatively short, and no one would be expected to remain a patient for an entire year. Moreover, as patients turn over, the mix of their symptoms, and hence the treatment plans, vary over time, and some treatments are more expensive than others. A snapshot on October 1<sup>st</sup> could yield a distorted picture of both the hospital’s utilization rate and the mix of patients.

Now, think about how we construct information systems for public schools. We often rely on October 1<sup>st</sup> headcounts of enrollments in particular grades as the key components of rates of school dropout and completion. The use of a single snapshot assumes some measure of stability in enrollments during the school year. In some schools, this may make sense, whereas in others, freezing a single frame of a moving picture could yield a very distorted impression. The system works best for schools that are stable and subject to few transient shocks. An interesting thought experiment is to imagine what kind of reporting system would be more appropriate for a school that has frequent movement into and out of the school. How frequently would we want to take a snapshot of enrollments? Twice a year? More frequently? A general design principle in social research is that the measurement of various phenomena should be aligned with a theory of how quickly the phenomenon changes. Something that doesn’t change very often does not need to be measured as frequently as something that changes more often. In some schools—perhaps the ones for which we are most interested in accurate measures—enrollments change rapidly.

Ultimately, the issue here is one of organizational routines, both at the local level and perhaps at the state level. If there’s a mismatch between the desired statistics and the organizational routines that would produce them, there are essentially two courses of action: either learn to live with the statistics that are produced by the existing organizational routines, or change those organizational routines so that they can produce the desired statistics.

Here's an example of a state-level organizational routine that could be changed—and why it might be very difficult to do so. One of the reasons that medical researchers are able to calculate mortality rates for individuals who have been treated by particular physicians or hospitals is that there is a definitive record of who has died, and when. Local and state departments of vital statistics maintain such records. What kinds of data are deemed to be “vital”? Births, deaths, marriages and divorces are part of the National Center for Health Statistics' National Vital Statistics System. Not, of course, high school completion. We don't have any official repository of lists of individuals who have completed high school, regardless of the source of the high school credential. The fact that these lists do exist for these other vital statistics suggests that the reason we do not have high school graduation in a vital statistics system is not technical; rather, it's political. It's not like marriage and divorce records are matters of public health; but we have deemed the state to have a legitimate interest in keeping track of family formation and dissolution, and there seems to be more skittishness about government reaching as deeply into the education system.

It's interesting to consider the implications of the interest in weighted student funding plans for the accounting of individual students. In theory, weighted student funding has the potential to make schools look more like hospitals. In weighted student funding plans, a school's operating budget is driven by the dollars that individual students bring with them to the school. Typically, there's a base allocation that comes with each student, with additional funds allocated when students have particular characteristics that are weighted more heavily—such as being free-lunch eligible, or an English language learner, or in a part-time or full-time special education setting. As I understand it, New York City's weighted student funding plan makes a preliminary funding allocation based on a projection of the mix of students who will be attending a given school in the coming year, and then adjusts that allocation based on the actual distribution of students enrolled on a particular date such as October 1<sup>st</sup>. But no further adjustments are made, even if students subsequently leave or are newly enrolled. If the money truly followed the child, then keeping track of enrollment spells would be essential, as it is in hospitals.

Of course, I can hear principals cursing me upon hearing of this plan. (I often imagine people cursing me.) Principals in particular just love mid-year budget cuts.

Let me conclude by returning to a point that both Rob and Elaine make: the most important issue in high school dropout and completion statistics is the use to which the data will be put. What do we want to know, and why? The known flaws in the existing statistics and data systems might not matter so much if the things we want to know are not influenced by them. I've argued today that data on high school completion is more useful than data on high school dropout, and that high school completion data to be used as indicators of the health and direction of the education system in a particular jurisdiction are most valuable for looking at changes over time, rather than variability across jurisdictions. I guess I still need to be convinced that improvements in the quality of the data are that critical to inferences to be made for this purpose. Conversely, the data demands for using high school completion data for accountability purposes are very high, and further investment in the information systems needed to generate those data is a high priority.