Visualization of Longitudinal Student Data

Anthony J. Bendinelli and M. Marder*
Department of Physics, The University of Texas at Austin, 78712 USA

We use visualization to find patterns in educational data. We represent student scores from high-stakes exams as flow vectors in fluids, define two types of streamlines and trajectories, and show that differences between streamlines and trajectories are due to regression to the mean. This issue is significant because it determines how quickly changes in long-term educational patterns can be deduced from score changes in consecutive years. To illustrate our methods, we examine a policy change in Texas that put increased pressure on public school students to pass several exams, and gave them resources to accomplish it. The response to this policy is evident from the changes in trajectories, although previous evaluation had concluded the program was ineffective. We pose the question of whether increased expenditure on education should be expected to correspond to improved student scores, or whether it should correspond to an increased rate of improvement in student scores.

I. INTRODUCTION

Students have always taken tests in school, but since 2002 the United States has stored unprecendented numbers of test results and made them available for evaluation of schools and for research. One result is a huge quantity of data describing students over time. Our aim in this article is to present new ways to analyze these data.

Our methods draw heavily upon traditions of analysis from statistical and fluid mechanics. This is the reason that we have submitted the work to a journal concerned with Physics Education Research. Our topic does not directly concern the improvement of the teaching of physics, but it does involve the improvement of education on a broad scale, and the technical details may be most accessible to an audience with a background in physics.

We have four motivations to develop new methods of analysis for longitudinal student data.

Patterns: A general advantage of visualization in science is that it enables researchers to detect patterns that might otherwise go unobserved. Visualization can serve the same function in education research as well.

Accessibility: Because of our emphasis upon visual representations, the primary results of our analysis will be more transparent and accessible to the broad public than methods whose end product is a set of coefficients in a linear model.

Long Observation Times: Our approach makes it natural to ask and answer questions about educational progress over long periods of time, rather than mainly focusing on changes in the course of a single year.

Causality: We are able to inquire into the causal effects of educational interventions in new ways.

These claims are unlikely to be persuasive in the abstract. For this reason we will present an example using data from Texas where our methods enabled us to detect the influence of a large-scale educational initiative that previously had been thought to have failed. Similar methods could be used to follow students within colleges, or to track them between secondary schools and colleges, although this has not yet been done.

The structure of this article is as follows: In Section II we provide a brief overview of high-stakes testing and some reasons that traditional methods for establishing causality in education should be regarded with caution. In Section III we provide conceptual definitions of snapshot and cohort velocity plots, snapshot and cohort streamlines, and trajectories. These are supplemented by mathematically formal definitions of the same plots in Appendix A. In Appendix B we present a statistical model that allows us to compute the difference between cohort streamlines and trajectories, and we show explicitly how the difference is related to the phenomenon of regression to the mean. In Section IV we describe the data set we have employed. In Section V we present visual evidence for a large and abrupt change in the flow properties of Texas students in mathematics. In Section VI we make the case that a particular statewide initiative was responsible for the change in the student flow pattern. In Section VII we pose final questions and conclude.

II. HIGH-STAKES TESTS AND CAUSALITY

School reform in the United States is a quantitative subject. The results of high-stakes tests are used to judge not just the students themselves, but also teachers, schools, and districts. Schools have to reach numerical benchmarks each year. Steadily increasing fractions of disaggregated student populations must reach benchmark scores, or else students must make Adequate Yearly Progress towards the benchmarks [1]. Schools and districts obtain labels such as Acceptable and Unacceptable in connection with these targets. The labels are significant both because they communicate to the public how well schools are doing, and also because schools that fail to meet standards five years in a row can be reorganized, meaning that the personnel can be replaced. Usher [2] estimates that 48% of public schools failed to meet Adequate Yearly Progress standards in 2011; thus half the nation's public schools had started down the path to dismissing their teachers and administrators.

These pressures coincide with a flood of reform efforts: new technologies, new forms of school organization, new routes for teacher certification, new evaluation and compensation policies for teachers, and new policies at the state and national level [3, 4]. It

*Electronic address: marder@mail.utexas.edu: To whom correspondence should be addressed.

is natural to ask which if any of these changes have had positive effects. These inquiries lead to further questions about how one can establish causal effects in education.

Several reports have laid out guidelines for establishing causation. The National Research Council produced reports in 2002 [5] and 2005 [6], and the American Educational Research Association produced a White Paper in 2007 [7] specifically designed to provide guidance to the National Science Foundation on educational research. All these reports take the position that causality is best established through carefully controlled experiments involving random assignment or, upon failing that ideal, through designs such as regression discontinuity. They also discuss practices for analyzing large educational data sets, almost exclusively using linear modeling.

The visualization techniques in this article were partly stimulated by worries that conventional quantitative methods in education research are more limited than many proponents recognize, that they are prone to certain sorts of errors that are rarely acknowledged, and that understanding educational data should accommodate the creation of specific alternatives with complementary strengths and defects.

We briefly relate some of our concerns:

1. The Institute for Education Sciences has been promoting random controlled trials for a decade, and the What Works Clearinghouse has carried out exhaustive searches for articles satisfying its methodological requirements. Searching the What Works Clearinghouse for Mathematics Achievement (9-12) in April 2012, one finds seven curricula [8]. In five cases the Extent of Evidence is Small [9–13]. In two cases the Extent of Evidence is Medium-to-Large but the Improvement Index is 0 or slightly negative [14, 15]. Thus adhering to the greatest methodological rigor for studying mathematics achievement in high school only restricts attention to a small number of suggested curricula, and only a few of those are slightly preferable to others.

2. Researchers do recognize that the study of one population does not trivially generalize to another [7, p 29]. A random controlled trial placing low-income students from Manhattan in charter schools does not necessarily provide guidance on how charter schools will serve low-income students in rural Iowa. Nevertheless, studies involving random assignment to treatment are often uncritically claimed to establish causality. For example, see the discussion of Case I, pp 59-69 in [7] that investigates whether there are "teacher effects on student achievement."

3. Analyses of large data sets tend to focus on very small numbers of outcome variables analyzed with linear methods such as Hierarchical Linear Modeling. The results of the modeling are expressed as tables of coefficients, and have to be interpreted by experts. This makes it possible for technical problems to creep in. For example, models sometimes compensate for large effects with linear terms, even if a simple scatter plot shows them to be highly nonlinear. Figure 1 provides an example showing that score changes of students in Texas depend on prior year score, but in a large and highly nonlinear fashion. Yet controlling for prior year score with a linear term appears to be common; for example, see the formula for New York City Value Added [16].

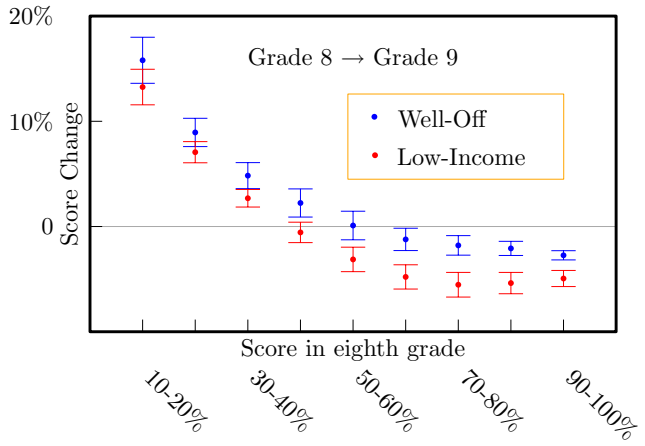

Figure 1: Average score changes of eighth-grade Texas students on Texas' high-stakes mathematics exam, TAKS, as a function of score in eighth grade, and as a function of student economic status measured by eligibility for free or reduced lunch. Average score changes across the whole state are treated as independent random events and used to construct standard error bars. That is, using data from 2003 to 2010, there are 7 independent measurements of student score changes. The point to take away from this figure is that controlling for prior year score with a single straight line would be a technical error, yet some statistical models of student performance do just that.

Because of these concerns, we will explore alternatives.

Here we develop visual representations of student scores using a method inspired by fluid mechanics [17]. In fluid mechanics one must keep track of large numbers of variables in space and time. Fluid particles undergo complicated motions that involve a mixture of deterministic and diffusive effects. The motion of students through time, grades, and scores lends itself to a similar representation. The role of visualization in science is to make it possible to examine large quantities of data and find patterns that may or may not be expected. The role we assign to visualization in education is similar.

III. DEFINITIONS

Conceptual descriptions

The public supports education so that US citizens can graduate high school with a certain level of skills and knowledge. Fifth-grade students no longer leave school for the workplace. The success of educational interventions eventually should be tested through their effect on the final outcome.

The country gathers educational data at time intervals on the order of a year. In some cases (e.g. the Texas mathematics exams we will discuss later in detail), the scores form a time series. In other cases (e.g. the end-of-course exam in a particular science), a single score is all one has to indicate student knowledge. Nevertheless, exam scores always reflect to some extent the growth in student knowledge and skill over time.

For the cases where exam scores constitute a detailed time series we have developed a representation in *flow plots*. A flow plot is like a weather map for student scores. It provides an immediate overview of how student scores develop from grade to grade for a variety of starting points.

Snapshot flow plots

A *snapshot* flow plot is constructed from data obtained over two consecutive years. As an example of how it works, take a group of students and organize them into bins or cells that have two indices. One index describes the student's **grade level** in a given year. The other index describes the student's **score level** in that same year. We typically index score levels by the fraction of maximum possible score that the student obtains in that year; we will discuss later the implications of focusing on raw scores in such a simple fashion. Next for each cell, compute the average score change for all students who also took the mathematics exam the next year. Plot an arrow in the cell whose area is proportional to the number of students and whose direction points to the new average score. In this particular representation, every cell describes a different collection of students from every other cell.

Figure 2 displays examples of snapshot flow plots. All the White students from Texas have now been placed in one of the 72 bins of Figure 2(A) according to their grade level and mathematics score level in the spring of 2009; for example, one of the bins contains all White fifth graders who obtained a score of between 70% and 79% on a mathematics exam. Figure 2(B) provides another example of a snapshot flow plot, but for Hispanic students in the state of Texas.

We provide a few observations about Figure 2 that apply generally. We draw two lines in each plot that correspond to Panel Recommendation and Commended scores respectively. We will discuss the definitions of these two terms in further detail in Section IV, but essentially, 90%-100% is a Commended score, and the Panel Recommendation line indicates that 70%+ is a passing score in elementary school and 60%+ is a passing score in middle and high school. We typically present subsets of students in such plots; in this case we examine White and Hispanic students. Figure 2 shows the relative performance of two groups of students on a common assessment across all grades and levels of performance [18]. One feature that stands out is the large downward motion of Hispanic students when moving from eighth to ninth grade. One sees that White students similarly moved downward, but there were proportionally far fewer of them. In particular, we single out the cell corresponding to eighth graders scoring between 70% and 79%. The magnitude of the score drop in this cell for Hispanic students is 12.5%, while for White students it is 9.0%; the difference is statistically significant with approximately 22,600 students ($p < 10^{-345}$). Thus the plot suggests that Hispanic students are particularly impacted by the transition from middle to high school, and those with scores between 60% and 90% are particularly at risk. We use this example to illustrate a natural way to interpret flow plots. When a feature stands out to the eye, one can turn to other sources of information and other tools to investigate it further. This is the normal function of scientific visualization. Its goal is not to test hypotheses, but to suggest them.

Cohort flow plots

A *cohort* flow plot is constructed similarly, but with one difference. The first column contains all third graders grouped according to their score in a given year. The second column contains all fourth graders grouped according to their score in the next year. The third column contains all fifth graders grouped according to the score in the year after that, and so on. If all students advanced by one grade each year and students never entered or left the testing system after third grade, then each column would contain the same cohort of students as every other column. However, the number of students in each class is not the same from column to column, due to students entering or leaving the school system, students who are not promoted to the next grade, or students who are not tested for some other reason. Thus, unlike most physical systems, particle number in this flow is not conserved. The space of grades and scores has been divided into cells and the plot computes the rate of score change in each cell (this description is like an Eulerian description of a fluid [17, p. 3]). Figure 3 is an example of this kind of plot.

In these examples, the lowest scoring bin has been omitted. Our analysis has shown that very few students score less than 10%, but there is an overwhelming number of tests that are marked zero. These zeros do not represent a lack of knowledge; a student may have been absent and missed the test, or the test may have been exempt from scoring [19]. Therefore, we have suppressed these bins in our plots.

Streamline plots

A *streamline* plot can be constructed from either a snapshot flow plot or a cohort flow plot. From a given starting grade, we draw a line that follows the direction of the flow plot arrows. We vary the thickness of the line to represent the number of students present at any given time. This is almost exactly like the definition of a streamline in fluid mechanics, but it is important to remember that the arrows of a snapshot flow plot correspond to several different cohorts of students. Consequently they are only an estimate of the likely path that students take through our score-grade continuum. Figure 4 shows an example of streamlines.

Trajectory plots

Our final flow plot is a *trajectory* plot. In this plot, students are divided into score bins in a starting grade such as grade three. These collections of students are fixed and do not change; we display the average mathematics score of each group for all subsequent years up through grade 11 (this representation is like a Lagrangean description of a fluid [17, p. 5]). Figure 7 shows an example of trajectory plots.

Thinking about the aims of education, trajectory plots are the ones of fundamental interest. They provide the record of how students starting at any particular level of performance in some starting grade turn out by the end of schooling. The reason to develop other sorts of plots is that the trajectories take so long to obtain.

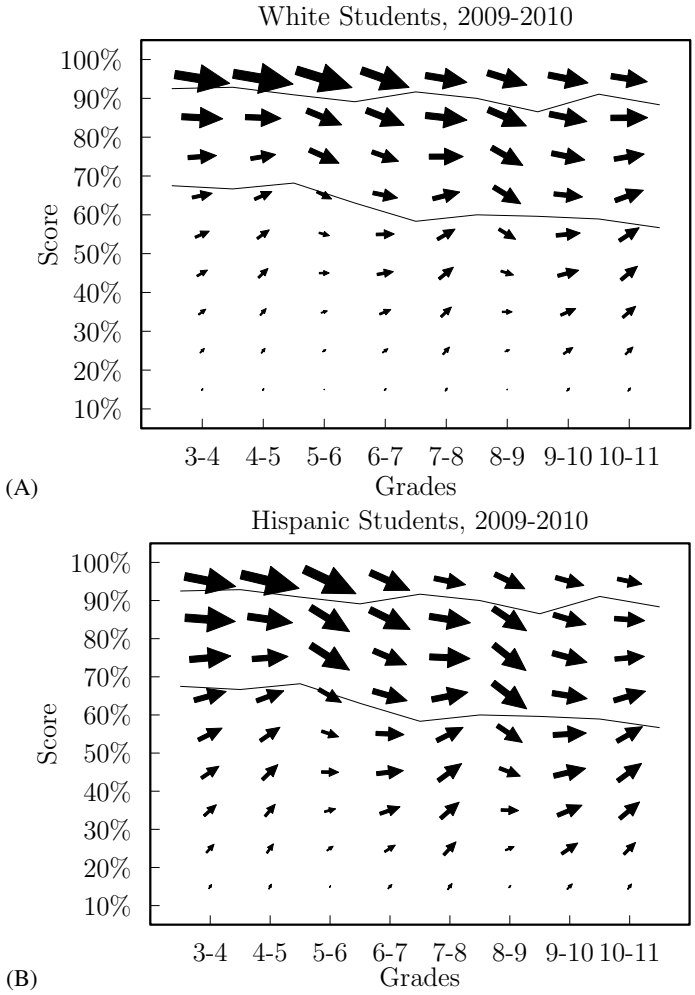

Figure 2: (A) Example of a snapshot flow plot computed from statewide mathematics scores on the Texas Assessment of Knowledge and Skills (TAKS, see Section IV), showing average score changes of White students from Spring 2009 to Spring 2010. The area of each arrow is proportional to the number of students. (B) Same plot but for Hispanic students. Lines indicate Commended and Panel Recommendation score cutoffs.

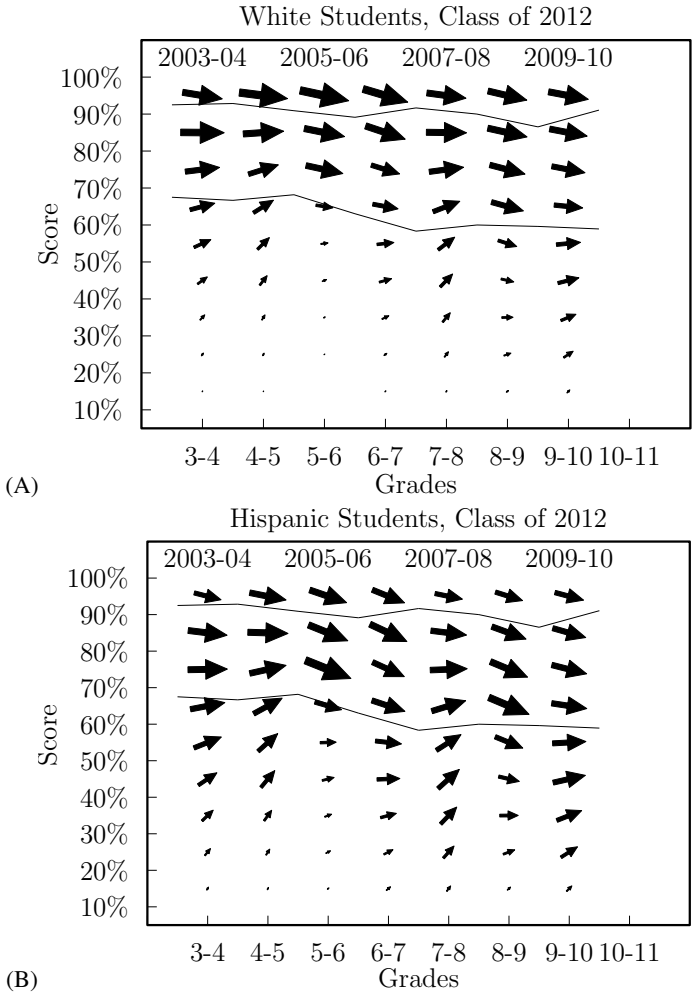

Figure 3: (A) Example of a cohort flow plot, again from TAKS mathematics scores, showing average annual score changes of White students in the Class of 2012. Lines indicate Commended and Panel Recommendation score cutoffs. (B) Same plot, but for Hispanic students.

Following students all the way from third to eleventh grades takes 9 years. Seeing whether whole trajectories change will take even longer. This is a very long time scale to wait to understand the effectiveness of educational interventions.

Snapshot plots require only two years of data. They provide a vector field one can sum up across grades to obtain an estimate of trajectories; that is, they provide an estimate of where students will end up by the end of schooling. This estimate arrives quickly enough to provide timely information for policy decisions. But is the estimate reliable? What are some of the technical problems that can undermine the correspondence between snapshot streamlines and trajectories? We begin to address these questions in Appendix A and Appendix B.

IV. STUDY SETTING

Features of specific example

Our hope is that our methods of analysis will prove useful to the education community, and we therefore devote the remainder of this article to a specific example. We analyze an educational reform with the following characteristics:

1. The reform was a linked set of policies and prescriptions in Texas involving almost all students. It would have been possible at its inception to assign students randomly to treatment groups, but this was not done.

2. The study population is all public school students in Texas over the course of eight years. Texas has nearly 10% of the population of the United States and has large populations in urban, rural, and suburban communities. Ethnic and racial "minorities" make up a majority of the school-age population. Thus one can plausibly interpolate from the results to a

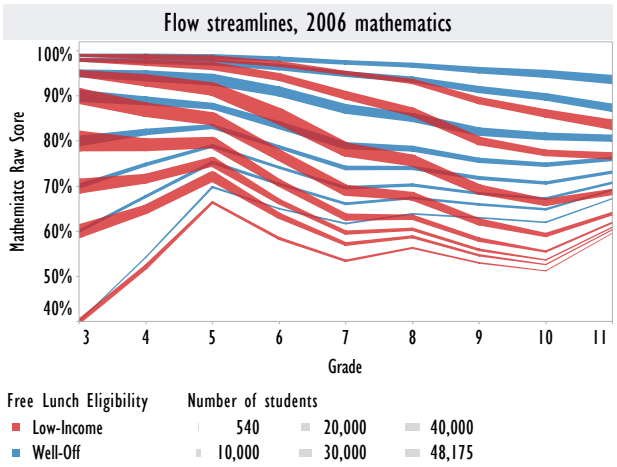

Figure 4: These streamlines were obtained from snapshot flow plots similar to those in Figure 2 using TAKS mathematics scores in 2005 and 2006 for low-income and well-off children, as determined by eligibility for free and reduced lunch. We disaggregate by poverty here rather than by race to illustrate the variety of comparisons that are possible. The comparisons seem rather extreme. By eleventh grade, low-income students who started in the 90th score percentile at third grade are below well-off students who started at the 60th score percentile at third grade. We will show later in this article that this type of plot exaggerates differences between groups and that real trajectories do not diverge as much.

wide range of communities and student groups in the United States.

3. Positive effects of the reform were not mentioned in evaluation reports, and it was canceled.

We were initially unaware of the existence of this educational reform and learned of it while trying to understand the change in flow visible in Figure 5. Using our visualization methods, we are now able to ask many worthwhile questions about what the intervention did and did not accomplish. However, before coming to the intervention, we need to discuss the exams we were studying and the way they were constructed.

TAKS examinations

The Elementary and Secondary Education Act of 2002 (No Child Left Behind, or NCLB) requires that each state adopt challenging academic standards and hold schools accountable to meeting these standards [1]. NCLB requires schools to show Adequate Yearly Progress for all public elementary schools and secondary schools, with separate annual objectives for economically disadvantaged students, students from major ethnic groups, students with disabilities, and students with limited English proficiency. NCLB also requires that within 12 years of the 2001-2002 school year, 95% of all students in each of these groups should meet or exceed each state's definition of proficiency in the form of a standardized test. In other words, by the 2013-2014 school year, all

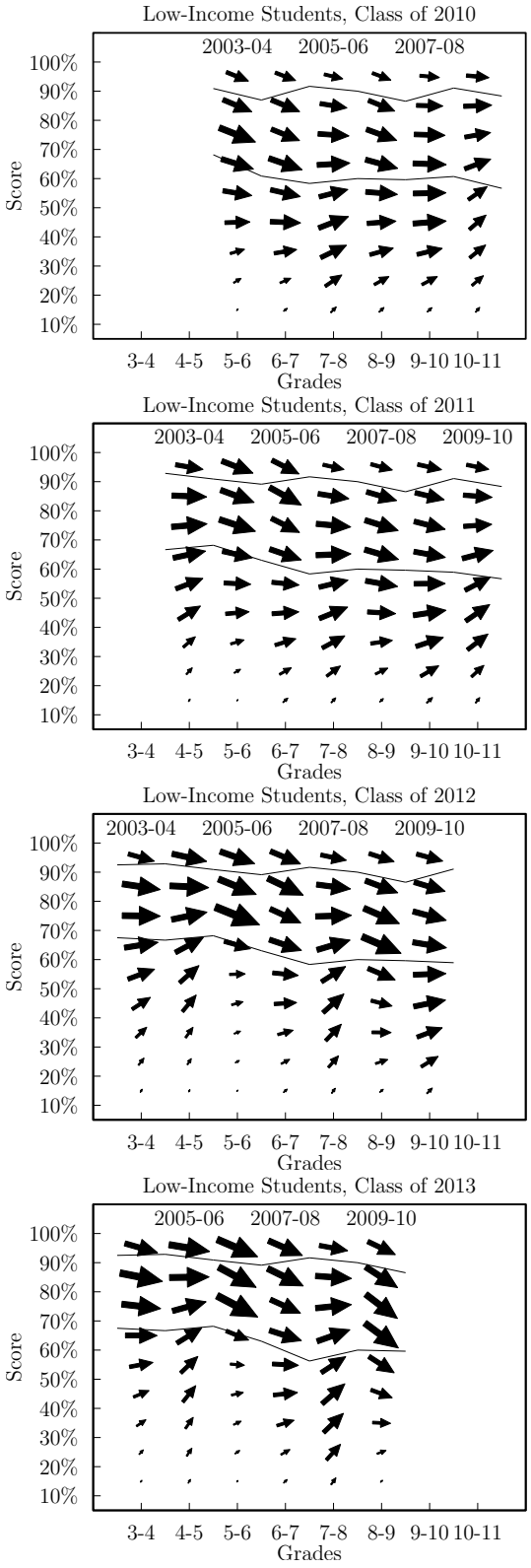

Figure 5: The flow pattern of Texas students changed with the Class of 2012. We display a series of cohort flow plots. Each arrow indicates the average one-year change in mathematics scores of students with the same starting score. The size of each arrow is proportional to the number of students. Upper and lower lines indicate Commended and Panel Recommendation scores respectively. Students eligible for free and reduced lunch (Low-Income) depicted only; well-off students showed similar improvement.

students (within 5%) must be able to pass each state's version of its standardized test.

In Texas, the standardized assessment used since 2003 is the Texas Assessment of Knowledge and Skills (TAKS). TAKS evaluates students between 3rd and 11th grade in mathematics, reading, writing, science, and social studies, although not every subject is tested in each year. Only mathematics and reading are tested at each grade level, and reading is combined with writing in 10th and 11th grade. We use the mathematics portion of TAKS to construct a longitudinal record of the progress of students over the course of several years. Using the methods described above in Section III, we can then use this longitudinal profile to calculate score distribution and snapshot flows, cohort flows, streamlines, and trajectories (careful definitions of the quantities appearing in the plots are contained in Appendix A).

We applied for access to the TAKS data set from the Texas Educational Research Center and used the data to construct multirank tensors containing information about Texas students from which all plots could be constructed. The dimensions of the tensors disaggregate the students into various ethnic and socio-economic groups, as well as grades, years, and score bins. The complete TAKS data set through 2010 consists of over 29 million individual examinations in its raw form. Each student is identified in the data set with a unique, anonymous number. Our Python scripts condense these files into a more manageable form where each row of the data set corresponds to a single student's progress over the time period of 2003-2010. In the case of re-tests, we look for the student's highest score for each subject. There are defects that our scripts attempt to reconcile, such as when math scores for a given student are in one row but the reading scores are located in another row. After reconciling as many defects as possible and combining re-tests with the main tests, we end up with approximately 23 million individual examinations corresponding to 6 million unique students.

Some of the data set's defects cannot be remedied. For the first year of the test in 2003, scores of third-graders were very incomplete and therefore we cannot use them. There are over 27,000 students with invalid records that have the same unique identifying number. Fortunately the number of entries with invalid student identifiers is very small in comparison with the 23 million valid entries. Another problem with the data arises from the termination of the State-Developed Alternative Assessment II (SDAA II) in 2008. Prior to 2008, students taking the SDAA II were exempt from having their tests scored. When the SDAA II was discontinued, approximately 60,000 students had their scores counted for the first time. This "source" of students is again relatively minor compared to the millions of students present in the data set, but it does provide a visible source of students appearing seemingly from nowhere between 2007 and 2008.

Raw scores vs. scaled scores

We now address a number of questions that have to do with the legitimacy of making comparisons between scores in consecutive years, and our use of raw rather than scaled scores.

We assume that within a single subject and at a single grade level, TAKS is equatable over time. That is, exams in different years contain problems of equivalent nature and difficulty. This assumption can be challenged. In particular, some assert that the methods from item response theory used to select exam problems (see Appendix B) preserve statistical features of student responses over time rather than the inherent difficulty and nature of the problems [20]. A primary use to which we put flow plots is to compare subgroups of students. These comparisons remain valid even if the exam is not equatable from one year to the next. However, in what follows we will examine changes in flow patterns over time, and these changes are most persuasive if the exam is equatable.

Two separate issues must be considered in the comparison of exams. One is whether the exams at any given grade are equatable over time. The second is whether they are vertically scaled. At least prior to 2009, TAKS was not vertically scaled. For example, if a student got a lower score on 6th grade mathematics than on 5th grade mathematics, it did not automatically mean that the student did not make progress over the year. Exams from one grade level to another were not directly comparable.

The Texas Education Agency also advises against comparing raw score on exams within one grade level from one year to another. The Agency provides a scaled score for this purpose. Nevertheless, we present raw score percentages in our plots. The reason for this decision is that the scaled score leads to a great loss of transparency, and it is not even available for every year.

To address concern over this issue, we summarize the procedures used to equate the exam from one year to another. More details can be found in Chapter 18 of Ref. [21]. The scale for the mathematics exam at each grade level was set during the construction of the original exam when large numbers of items were field-tested, and the difficulty of each item was ascertained. Items were also examined for content validity and grade appropriateness by a variety of experts, and passing scores for each grade were determined by a panel. Each subsequent examination contained a mixture of questions; some determined the student's score, while others were new and were undergoing field testing. From the field tests, the difficulty of each new item was ascertained. This made it possible to construct an exam of comparable difficulty each year. That is, exam items were chosen so as to equate the exams from year to year. Each TAKS mathematics exam was subject to an additional equating procedure after it was given that resulted in the final scaled score. The precise algorithm used in this process is proprietary. The net result of the post-equating procedure is a correspondence between raw and scaled scores for each examination. If the pre-equating process is successful, then the conversion between raw and scaled scores should be stable over time.

We examined every conversion table from raw to scaled score for every grade available from 2003 to 2010, and observed that the raw score corresponding to Panel Recommendation varies by at most two raw score points from year to year. In the few cases where the Panel Recommendation score differs from the most typical value by more than 5%, the tests are alternate versions of the TAKS test (e.g. online tests, re-tests administered in months other than April). We found one instance, ninth grade mathematics in 2010, where the Panel Recommendation score was three points lower than in prior administrations. This is the most problematic

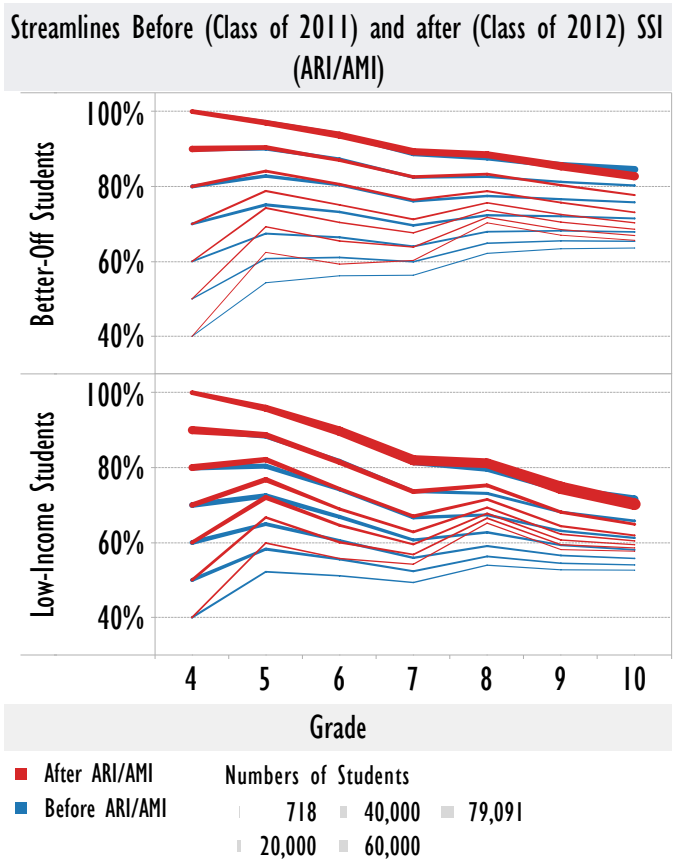

Figure 6: Streamlines computed by averaging over the data from the Classes of 2005 through 2011, and from the Classes of 2012 through 2016 in the cohort flow plots of Fig. 5. The thickness of the streamlines is proportional to the number of students.

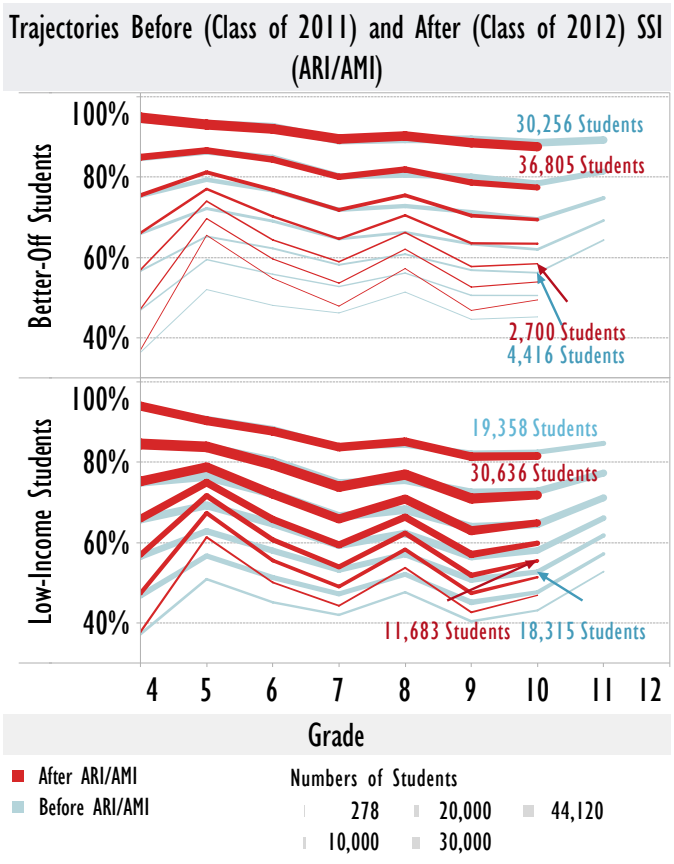

Figure 7: Trajectories computed by tracing explicitly the average scores over time of cohorts of fourth graders grouped by their fourth grade mathematics scores, before and after the introduction of SSI (ARI/AMI). Comparison with Fig. 6 shows that following cohorts of students in time explicitly rather than integrating up year-to-year changes produces trajectories that compress less towards the mean.

case for our use of raw scores, and plays a role in the score drops of high-performing ninth graders in Figure 9.

V. CHANGE IN FLOW

While examining flow plots for individual cohorts of students (Figure 5), we were surprised to see that the patterns for the Texas graduating Classes of 2012 and 2013 differed dramatically from the flow pattern for the Class of 2011 and before. Starting in the 2004-2005 school year, mathematics scores of fifth-graders leaped up to such an extent that the percentage of low-income students failing the mathematics exam dropped from 36% the year before to 28%. A second leap is visible when these students arrive in eighth grade. The graduating Class of 2013 shows an essentially identical pattern of improvement relative to the Class of 2011.

We averaged together the available average score changes for students from the Classes of 2005 through 2011 and compared it to the average score changes for students from the Classes of 2012 through 2016. From these averages we integrate over time to compute streamlines, as shown in Figure 6. A strong difference is apparent after the Class of 2012 passes through school. However, we were skeptical of potential inaccuracies resulting from the process of integration, and found no completely satisfactory way to allocate numbers of students to particular streamlines as time progressed.

Therefore we turned to trajectories, as shown in Figure 7. We follow particular cohorts over time, starting in fourth grade, and can therefore speak unambiguously about their mean scores and the numbers of students present on the trajectory at any given time. Comparison of Figures 6 and 7 shows that the two computations are similar in overall features but rather different in detail. The trajectories are clearly superior to the streamlines, since they literally report the mean score over time of a particular cohort of students. Streamlines computed from flow snapshots have the advantage that they can be computed from just two years' worth of

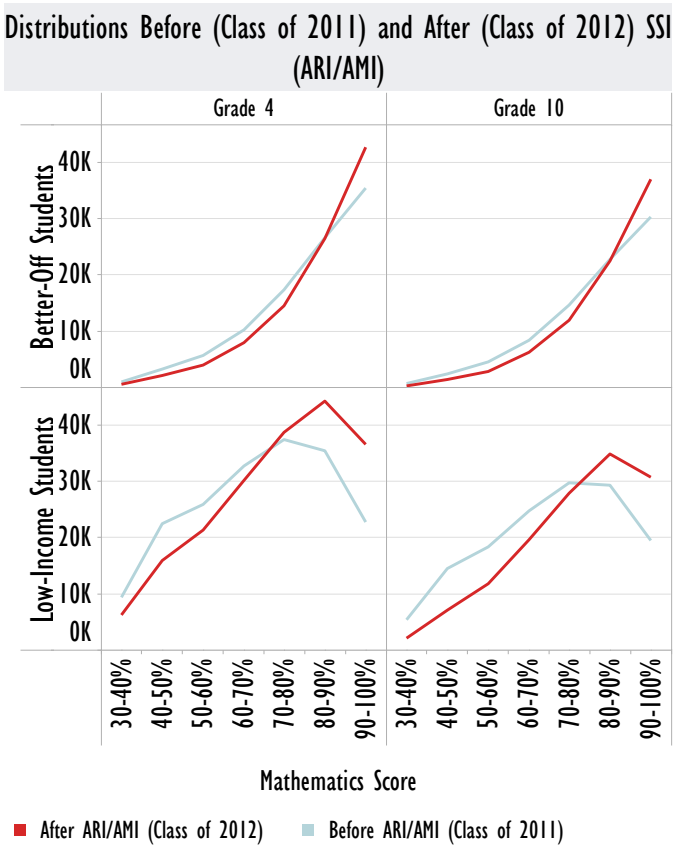

Figure 8: Distribution of student scores for the two cohorts before and after ARI/AMI. The distributions are shown for fourth and tenth grades. Note that the number of low-scoring students drops when moving from the first cohort to the second, while the number of high-scoring students increases. This trend appears for both better-off and low-income students, and it persists both for fourth grade and six years later in tenth grade.

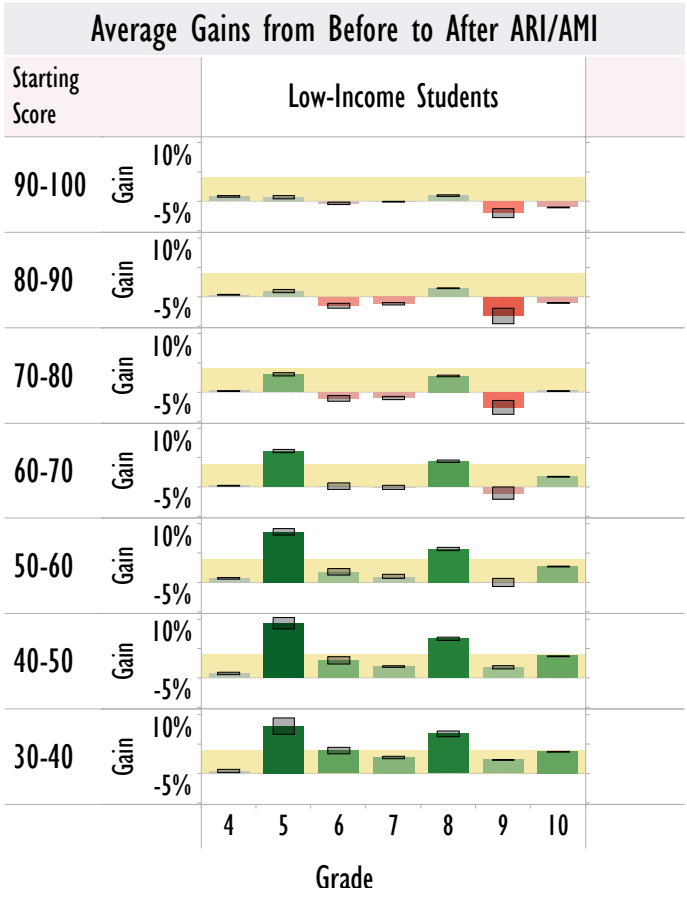

Figure 9: Low-performing low-income students saw gains in all grades. Standard error bars follow from averaging data over several cohorts. For *Before ARI/AMI* all data are employed. For *After ARI/AMI*, all cohorts graduating in 2012 and after are used. However, for students in eighth grade and below, 2010 data are excluded since funding for the program dropped by a factor of 3 in 2009-2010. Data for students above eighth grade are retained on the grounds that they received the boost ARI/AMI provided, and it is reasonable to see if they retain momentum or not. The data in ninth grade, which include a contribution from 2010 when funding dropped, show a strong decrease.

data, but unfortunately they perform only a qualitative job of providing trajectories. The streamlines are compressed and regress towards the mean, as explained by Eq. (B11).

We refer to students scoring 60% or less in fourth grade as low-performing. One sees that after their scores rose in fifth grade, low-performing students held on to score gains in sixth and seventh grades, rising again in eighth. This does not mean that their scores remained constant; the mean scores dropped in sixth and seventh grade for all groups of students shown in the plot. Rather, it means that in sixth and seventh grades, low-performing students from the Class of 2012 onward achieved score gains relative to the Class of 2011. The gains were initially around 10%, then dropped, but not to zero, and stabilized at 2–4%. To display the magnitudes of the gains sustained by students, we display them explicitly in Figure 9, which incorporates data from all available cohorts.

Even more striking is a dramatic shift in the numbers of students that populate different trajectories. This shift had already occurred by the time students reached fourth grade, presumably due to interventions in earlier grades for which we do not have data. We focus on the top-most trajectory in Figure 7. This trajectory can be identified as the one for which fourth-graders had a raw score between 90% and 100%. The students do not continue to score over 90% as time goes forward; their mean score rises and falls over the years, eventually ending up near 80% in tenth grade for the low-income students. However, what is noteworthy is that in moving from the Class of 2011 to the Class of 2012, the number of students in this upper trajectory increased by over 50% without the shape of the trajectory changing. There are both red and blue lines running along the top of the figure, but they track so precisely that the red obscures the blue. In particular, for the Class of 2011, 22,662 low-income students in fourth grade were in the

highest trajectory, and of those 19,300 reached tenth grade. For the Class of 2012, 36,403 students scored between 90% and 100% in the fourth grade, and of those 30,600 students reached tenth grade. This increase from one cohort to the next is not explained by population growth, which was at the level of around 2% per year for Texas [22]. In Figure 8 we show the distributions of low-income and better-off students at fourth and tenth grades explicitly.

VI. STUDENT SUCCESS INITIATIVE

What produced these results? At any given time Texas is funding many education initiatives. However, the score gains we observe have the fingerprints of a particular policy on them: the Accelerated Reading Instruction/Accelerated Math Instruction (ARI/AMI) component of the Texas Student Success Initiative (SSI). In 1999-2000 ARI was implemented for struggling reading students, and in 2003-04 AMI was implemented for students struggling with mathematics. ARI/AMI allowed all students who failed either the TAKS reading or the TAKS mathematics exams to receive "accelerated instruction," which is to say intensive tutoring that continued into the summer for some students [23]. When the Class of 2012 reached third, fifth, and eighth grades, they had to pass the mathematics and reading exams or they would not be allowed to advance to the next grade level. However, students who failed these TAKS exams in these years could take them a second and a third time.

For the Class of 2012 the TAKS mathematics exam at third, fifth, and eighth grades not only could label a school as low-performing; it also had high stakes for students. Students were given multiple chances to learn the material, and extra instruction to do so. It appears to have worked. The evidence lies not so much in the fact that the number of students failing at fifth and eighth grade dropped, but that score gains were retained, the number of low-income students passing through school in the highest performance trajectory nearly doubled in the space of a year, and these results were sustained without any drop in scores. The only signs of negative results come from high-scoring students at ninth grade (maybe the result of a difficult exam in 2010), but these losses did not persist to tenth grade and can be balanced against the large increase in the numbers of students in the higher trajectories.

How large were the gains? Score gains are frequently discussed in terms of standard deviations; the standard deviations of scores on Texas mathematics exams range from 15% to 20% of the raw score. Annual student gains that could supposedly be obtained by replacing teachers in the 25th percentile of the quality distribution by teachers in the 75th percentile are .2 standard deviations [24], or around a 3–4% score increase on the exams. Our data show that low-performing students across Texas in the Class of 2012 on average made gains of this order in fifth grade and afterwards.

In searching for alternative explanations, we wondered if low-scoring students were being pushed out of the data set, perhaps moving to take the State-Developed Alternative Assessment II, which is allowed for special education students. We did not find evidence that low-scoring students vanished from the TAKS data set more frequently for the Class of 2012 and after than for the prior cohorts. We also could find no reason to believe that Texas' mathematics exams became systematically easier precisely so as to correspond to the shifting flow plots of the Class of 2012 and beyond.

Although the ARI/AMI was evaluated every two years, the observations we present here appear to be new. The Texas Education Agency evaluated the initiative by looking for score gains between successive cohorts of students in the same grade after the policy was in effect [25]. The policy had produced a new steady state, but steady states do not change over time, and the policy was judged a failure [26]. Funding was moved elsewhere in 2009-2010 [26] and the changes in funding over time are shown in Table I. It should be noted that despite the reduction in funding, schools are still required to identify struggling students and provide them with additional instruction.

VII. CONCLUSIONS

In closing, we emphasize two of our main points:

- Streamlines, even cohort streamlines that require many years of data to construct, are quite different from trajectories. Streamlines (Fig. 6) strongly overstate the divergence in performance of different groups, and give a false impression that students at many different starting points will regress to the mean. For example, the top streamline for better-off kids drops nearly 20 points from fourth to tenth grade, while the top trajectory (Fig. 7, which due to details of construction starts in a slightly different place) drops only around 5 points over the same span of grades. The difference between the top streamline for better-off and low-income kids by tenth grade is around 12 points; the corresponding difference for trajectories is 6 points. Putting it another way, trying to deduce long-term results (in the simple way we first tried) from year-to-year changes produces errors on the order of a factor of two.
- Test score increases in early grades can persist over long periods of time. The strongest evidence for this claim comes from the upper trajectories of Figure 7. The number of low-income fourth graders in the Class of 2012 with a fourth grade score above 90% in mathematics increased over 50% in comparison with the previous year, and the students' scores then tracked those from the cohort before almost perfectly, despite the increase in number. This is not a trivial achievement. Furthermore, this growth came at the expense of student population of lower trajectories, while mean scores of the lower trajectories went up. We note that the intervention of ARI/AMI began in early grades, particularly in third grade, for which we do not have usable test data. However we find it plausible that the large rise in high-scoring fourth graders is due to this cause.

An interesting question raised by these observations is how the cost-effectiveness of educational interventions should be judged. An implicit assumption of No Child Left Behind is that by spending essentially fixed amounts per year, student performance can increase steadily until all students in the United States reach proficiency in mathematics and reading in 2014 [1]. The evaluation reports for the Texas Student Success Initiative appear to expect effects of this type: steady increases in performance in each

Year	Funding (M$)	Year	Funding (M$)
99-00	65.2[a]	05-06	149.5
00-01	57.5[a]	06-07	144.2
01-02	106.4[a]	07-08	124.9[e]
02-03	75.1[a,b]	08-09	123.3
03-04	80.9[c]	09-10	44.2[f]
04-05	144.1[d]	10-11	44.4[f]

Table I: Funding history of initiative we associate with score gains. (a) Accelerated Reading Initiative (ARI) funding only (b) First year grade 3 had to pass (c) Accelerated Mathematics Initiative (AMI) funding begins (d) First year grade 5 had to pass (e) First year grade 8 had to pass (f) ARI/AMI defunded; Student Success Initiative only [23]

grade level, with the rate of increase proportional to expenditure of funds. This sort of change is very easy to check. "This year 73% of third graders passed mathematics, while last year 70% of them passed." What is easy to overlook is that different cohorts of students are being compared, and the changes can correspond to many things, including small shifts in the difficulty of the exams due to technical challenges involved in year-to-year equatability. Public reports on the comparisons can either focus on the absolute numbers being achieved or the changes, and accountability laws use both.

Our analysis looks at changes of a different sort. We look at the long-term trajectories of cohorts of students. What we found was that the Student Success Initiative produced a change in trajectories and that the change itself responded to expenditure, rather than the rate of change. Table I shows the history of funding for the ARI/AMI component of the Student Success Initiative. The changes in student scores correspond to the increased funding in the 2004-2005 school year, but once trajectories shifted upwards, they did not keep shifting year after year, but simply stayed at a new elevated location. One might argue that this shows the initiative was in fact defective as it produced a static increase in scores and nothing more. Initial examinations of the TAKS data suggest that in fact scores began to decrease again when the program was defunded in the 2009-2010 school year, but further data will be required to accurately determine the impact. Our tentative prediction is that cuts in public funding in Texas that started in 2009-2010 will result in actual declines in student performance that will display themselves as cohorts move upwards through the school system. The cuts in funding happen to coincide with a change in the Texas high-stakes accountability system from TAKS to end-of-course exams in high schools. Thus if this tentative prediction can be tested, it will have to be through state-level results on the National Assessment of Educational Progress, as TAKS will no longer be administered.

We arrive in the end at a set of simple questions:

1. When funding for education increases or decreases, does it affect the *level* of student scores in a given grade, or does it affect the *rate of increase* of student scores in a given grade? Do some interventions increase levels while others increase rates?

2. When a cohort of students experiences an educational intervention, how does it play out over time? Are there some interventions that have positive effects on students in one grade, but negative effects as the students proceed? Are there some interventions that are inherently durable? Are there some that last longer than others?

3. What is the minimum number of years of longitudinal data that is necessary to entertain predictions about changes in student outcomes all the way out to the end of schooling, and therefore to determine the long-term effects of interventions? Results we presented here suggest that a minimum of three years of student data are necessary, but we do not yet know whether three years are sufficient.

Visualization methods provide a powerful technique for evaluating the progress of students over time. They make no *a priori* assumptions about linearity, and instead allow the data to describe the system. They suggest new forms of mathematical models, which we are refining in order to improve predictions about long-term consequences of perturbations such as the Student Success Initiative. But prediction is always likely to have limits. We should have the patience to watch for the consequences of policy changes, and should be willing to give credit to hard work by teachers and schools in cases where it is deserved.

Acknowledgements

This work was partially supported by the U.S. National Science Foundation Materials Theory program, DMR1002428. Access to the Texas longitudinal data set was made possible through the UT Dallas Educational Research Center. Any opinions expressed in this article are not necessarily shared by either the National Science Foundation or the Texas Education Agency.

Appendix A: Formal definitions of snapshot flow, cohort flow, and trajectory plots

Table II records the notation used to describe scores and grade levels in this paper. We first describe the *velocity* or mean score change of students over time. We select a collection of students in a single year t who are in the same grade g and whose score falls in bin k (bins correspond to 90% and above, 80%-90%, etc). Our choice of the bins deliberately intends to exploit the frequent understanding that 90%-100% is an "A", 80%-90% is a "B", etc., along with the levels of competence these gradations imply. We further select the subset of these students who, in the next year $t+1$, advance to the next grade $g+1$ and have a nonzero score; we call this subset of students $A_{t,g,k.}$ In addition to specifying grade, year, and score, the set might also include restriction to a particular ethnic or economic group (e.g. White, or eligible for Free and Reduced Lunch), but we have chosen not to put an additional index on A or the other symbols to denote subgroups. Then the mean score change of students in set A from year t to year $t+1$ is:

$$v_{t,g,k} \equiv \frac{\sum_{\alpha \in A_{t,g,k}} \left(s^{\alpha}_{t+1} - s^{\alpha}_{t}\right)}{N_{t,g,k}}, \tag{A1}$$

Symbol	*Meaning*
t	An integer denoting the year in which a test is taken. When a test is taken in an academic year such as 2009-2010, we use $t = 2010$.
s_t	A test score in year t, in units of percentage of maximum score.
s_t^α	The test score of student α (an integer) in year t in units of percentage of maximum score. When students take multiple administrations of the exam during the year, we choose the maximum.
g_t^α	The grade level of student α in year t.
$S(k)$	The k'th boundary of bins used to make scores discrete: $S(k) = \frac{k}{10}100\%$, $k \in [0, 1, \ldots 10]$. A score s_t is in bin k when $S(k) < s_t \leq S(k+1)$.
$A_{t,g,k}$	A set of students who in year t are in grade g, whose test score is in bin $k \neq 0$, who advance to grade $g+1$ the following year, and who have nonzero score the following year.
$N_{t,g,k}$	The cardinality of the set $A_{t,g,k}$ (i.e. the number of students in year t, grade g, and bin k).
$v_{t,g,k}$	The average score change of students in year t, grade g, and bin k (in set $A_{t,g,k}$).
$\bar{s}_{k_0,g_0,t_0\to t}$	The average score in year t of students who in year t_0 had score given by k_0 and were in grade g_0.
$\mathsf{S}_{t'}^{s,t}$	The score in year t' of a trajectory passing through score s in year t.

Table II: Notation and conventions used to define flow plots in this paper.

where $N_{t,g,k}$ is the number of students in $A_{t,g,k}$. In [27] we showed that this definition arises formally when one analyzes the change of student test scores using procedures from statistical mechanics to derive a Fokker-Planck equation [28]. However, they are the most obvious definitions one could adopt, quite independent of any formalism.

A *snapshot* flow plot is a visual representation of $v_{t,g,k}$ and $N_{t,g,k}$. All the data in the plot come from data in two consecutive years, t and $t+1$. The horizontal axis gives grade level g; we put grades g and $g+1$ into the tick labels to clarify the starting and ending points for each arrow. The vertical axis gives score level k. Every cell indexed by k and g in the plot has an arrow. Arrows point at an angle so that if their horizontal length is 1, their vertical height is $v_{t,g,k}$; that is, the arrows point towards the mean score of students the following year. The area of each arrow is proportional to the number of students involved in $N_{t,g,k}$. Figure 2 provides examples of snapshot flow plots.

A *cohort* flow plot makes use of the same ingredients, but instead of plotting $v_{t,g,k}$ for a single year t, each successive column of the plot advances the year by 1. It is a plot of $v_{t,g+t-t_0,k}$ and the precise cohort under investigation can be tuned by selecting the offset t_0. Thus the plot follows a cohort of students advancing through school together, subject to corrections due to students who leave and enter school, or students who are retained a grade. Figure 3 gives examples of these plots. This kind of plot provides a more accurate representation of progress through school than a snapshot, but it requires many years of data to produce, while the snapshot can be produced with two years of data. The two versions of plots are the same when schools are in steady state, and differ when schools change substantially over time. Figures 2 and 3 are quite similar, but Figure 2 needs two years of data while Figure 3 requires eight.

From either snapshort or cohort flow plots one can derive *streamlines*. These are obtained in exactly the same way that particle streamlines in a fluid can be obtained from a velocity vector field. For any grade level, we use linear interpolation across the $v_{t,g,k}$ values to construct a series of continuous functions $v_{t,g}(x)$, where the score x varies continuously from 0 to 1. This allows us to estimate the average score change for a student regardless of what he or she scored in year t. The value of $v_{t,g}(x)$ for $x = 1$ is determined by linear extrapolation, and if it ever turns out to be positive, it is set to zero, since students getting a perfect score on the mathematics exams cannot improve any further. A similar correction is made if $v_{t,g}(x)$ for $x = 0$ is negative. To get streamlines, we pick a starting score in third grade and use $v_{t,g,k}$ to calculate the average score of those students in fourth grade. From there, we use the continuous $v_{t,g}(x)$ to estimate the scores of those students in fifth grade and all grades after that. In snapshot and cohort flow plots, we set the area of the arrows to be proportional to $N_{t,g,k}$; similarly, we set the width of the streamlines to be proportional to $N_{t,g}(x)$. Figure 4 provides an example of streamlines computed from a snapshot flow plot for well-off and low-income children, determined by whether the student is eligible for free/reduced lunches or not. As we will show in Appendix B, the striking difference between the streamlines of the two groups exaggerates their actual differences, and is partly an artifact of regression to the mean.

Finally, we can derive *trajectories* from the data set. We start by selecting the subset of students A_{t_0,g_0,k_0} that are initially in score bin k_0, grade g_0, and year t_0. Instead of using interpolated velocities to estimate their scores in future years, we follow this cohort of students explicitly and record their actual scores in all years $t > t_0$. By plotting the average scores $\bar{s}_{k_0,g_0,t_0\to t}$ of this subset of students over several years, we track their path through our score-time continuum.

Appendix B: Difference between streamlines and trajectories

We have raised the question of the conditions under which it is possible to make predictions about nine years of progress through school by measuring two years of data. In order to address this question, we establish a formal structure that resembles the structure of classical testing theory. However, it is conceptually somewhat different.

Classical testing theory posits that every student has an underlying knowledge state T_i [29]. When the student takes a test, he or she gets a score s_i that differs from her underlying knowledge by a random error term ξ_i:

$$s_i = T_i + \xi_i. \tag{B1}$$

What is the underlying knowledge state T_i? Assuming that ξ_i has mean zero, it can be obtained by posing sufficient numbers of tests. There can be many different opinions on what really constitutes underlying knowledge. Perhaps the random error results only from students randomly bubbling in questions when they do not know the answers. In this case the error could be reduced to any desired level by making the test long enough. Perhaps it results from

fluctuations in student mood from day to day. In this case testing would need to be spread over several days. Perhaps the underlying knowledge means the student's true knowledge of mathematics, and the random error includes a contribution from biased test construction, which should be compensated by having completely independent groups prepare tests. All of these conceptual constructs are consistent with Eq. (B1).

The TAKS mathematics exam is based upon a more complicated statistical framework, one-parameter Item-Response Theory [21]. In this framework, every item i on an exam has a difficulty δ_i and every student n has a proficiency θ_n; the probability P_{ni} of student n correctly solving item i is

$$P_{ni} = \frac{\exp(\theta_n - \delta_i)}{1+\exp(\theta_n - \delta_i)}. \tag{B2}$$

The computations involved in deducing student proficiencies over time and testing hypotheses about whether they have changed can become extremely elaborate, particularly since both proficiencies and problem difficulties must be estimated from the data.

There is another community of practice in statistics that we should mention, the study of longitudinal data. In contrast to physics, where the introduction of time has been fundamental to the field since inception, in statistics the practice is newer. As recently as 1970, Cronbach and Furby said that researchers trying to study change should "frame their questions in other ways" [30]. The subject has been developed since, but we have not yet found results that resemble those we present here. The closest resemblance is from hazard models [31, Ch. 11]. These involve the probability of observing binary events over time, and have a considerably different flavor.

When we compute cohort streamlines and trajectories explicitly, we find in Figures 6 and 7 that they are quite different. Snapshot streamlines can be expected to differ from trajectories simply because of changes in the flow over time. Explaining a difference between cohort streamlines and trajectories is more difficult because time dependence in the flow seems at first to be present in both of them in the same way. Here we carry out a formal analysis of a stochastic system designed to analyze this difference.

We emphasize that although we are now introducing a formal framework with a random component, it is different from testing theory. We take test scores at face value and simply aim to supply a compact mathematical description with a deterministic component and a random component. The random component does not describe a description of the difference between a student's score and the student's underlying knowledge. Instead, it describes the random difference between a student's actual score in year $t+1$ and an attempt to predict the score based upon past performance.

Our main finding at a qualitative level is that streamlines regress to the mean [32] more quickly than trajectories. The rate at which this happens depends how strongly score changes depend upon history. To demonstrate this result, we adopt a Langevin equation framework [33, ch. 15]. This means that we take the score of each student to be a deterministic function of past scores plus a random component. In particular we suppose that the score of a student α in year $t+1$ is related to that student's score in year t by:

$$s_{t+1}^{\alpha} - s_t^{\alpha} = V\left(s_t^{\alpha}, s_{t-1}^{\alpha}; t\right) + \xi_t. \tag{B3}$$

Here $V(s_t^{\alpha}, s_{t-1}^{\alpha}; t)$ is a deterministic function that predicts score changes based upon *two* prior scores and ξ_t is a normally distributed random variable with the following properties:

$$\langle \xi_t \rangle = 0; \quad \langle \xi_t \xi_{t'} \rangle = \delta_{tt'} D. \tag{B4}$$

where D is the variance of the distribution. It is the dependence on two previous times rather than one in V that will lead streamlines and trajectories to diverge.

In the limit where the noise amplitude D vanishes, the problem is completely deterministic. We use $\mathsf{S}_t^{s_0,t_0}$ to denote the score in year t along a trajectory that has initial score s_0 in the year t_0. Suppressing superscripts that describe a common initial condition, these deterministic trajectories obey:

$$\mathsf{S}_{t+1} - \mathsf{S}_t = V\left(\mathsf{S}_t, \mathsf{S}_{t-1}; t\right). \tag{B5}$$

There are families of trajectories, and individual trajectories are selected by specifying the value s_0 in some particular year t_0. If the probability distribution of ξ_t is normal, then the deterministic trajectories obeying Eq. (B5) are the most likely paths for students to take in the presence of noise.

In particular, the difference between a velocity $v_{t,g,k}$ computed for a student getting score s_t in year t and the trajectory $V(s_t, \mathsf{S}_{t-1}^{s_t,t})$ passing through s_t is the product of two terms. The first term describes the degree to which the deterministic trajectory $V(s_t, s_{t-1})$ depends upon the score in year $t-1$. The second term is a quantitative measure of how much scores are regressing to the mean.

The probability of having a value of the noise ξ_t variable in Eq. (B3) is given by the normal distribution:

$$N(\xi_t) = \sqrt{\frac{1}{2\pi D}} \exp\left\{-\xi_t^2/2D\right\}. \tag{B6}$$

Using Eqs. (B3) and Eq. (B6) we can derive the joint probability distribution P for obtaining a sequence of scores $\sigma_0, \sigma_1 \ldots \sigma_T$. We are using variables σ rather than s because we need to compute expectation values, and we will adopt a convention in which we integrate over score variables σ. That is, the expectation value of some quantity $\langle Q \rangle$ is given by multiplying Q by the probability $P(\sigma_0 \ldots \sigma_N)$ and integrating over $\sigma_0 \ldots \sigma_T$.

Assume that $V(\sigma_0, \sigma_{-1}; 0) = V(\sigma_0; 0)$ does not depend upon s_{-1}. That is, score changes depend upon two prior years, except for the lowest grades in which students take tests. Let P_0 be some probability distribution for scores in the lowest grade where they are recorded. Using Eq. (B3), we can derive:

$$\begin{aligned} P(\sigma_0 \ldots \sigma_T) &= \int d\xi_{T-1} \left[N(\xi_{T-1})\right. \\ &\quad \times \delta\left(\sigma_T - \sigma_{T-1} - V(\sigma_{T-1}, \sigma_{T-2}) - \xi_{T-1}\right) \\ &\quad \left. \times P(\sigma_0 \ldots \sigma_{T-1})\right]. \end{aligned} \tag{B7}$$

That is, the probability of getting σ_T given $\sigma_0 \ldots \sigma_{T-1}$ is given by the probability of having the value of ξ_{T-1}needed according to Eq. (B3). Performing the integral, Eq. (B7) becomes

$$\begin{aligned} P(\sigma_0 \ldots \sigma_T) &= N\left(\sigma_T - \sigma_{T-1} - V(\sigma_{T-1}, \sigma_{T-2})\right) \\ &\quad \times P(\sigma_0 \ldots \sigma_{T-1}) \\ &= \exp\left\{\frac{-\left(\sigma_T - \sigma_{T-1} - V(\sigma_{T-1}, \sigma_{T-2})\right)}{2D}\right\} \\ &\quad \times P(\sigma_0 \ldots \sigma_{T-1}). \end{aligned} \tag{B8}$$

Applying Eq. (B8) recursively makes it possible to write the probability distribution explicitly as

$$P(\sigma_0 \dots \sigma_T) = \left(\sqrt{\frac{1}{2\pi D}}\right)^T P_0(\sigma_0) \times \exp\left\{\sum_{t'=0}^{T-1} \frac{-(\sigma_{t'+1} - \sigma_{t'} - V(\sigma_{t'}, \sigma_{t'-1}; t'))^2}{2D}\right\}. \quad \text{(B9)}$$

In order to find the difference between trajectories and streamlines, we need to find the average score change from year t to year $t+1$ when nothing is specified about the score in year $t-1$. We will compare with results that come by following the deterministic trajectory S of Eq. (B5). These deterministic trajectories cause the argument of the exponential in Eq. (B9) to vanish, and maximize the probability distribution.

More specifically to carry out the comparison, we want to find the expectation value of $(\sigma_{t+1} - \sigma_t)\delta(s_t - \sigma_t)/P(s_t)$. Here $P(s_t) = \langle \delta(s_t - \sigma_t) \rangle$ is the probability that the score at time t has value s_t and we need it in the denominator to keep the expectation value properly normalized. We sketch the ensuing computation. All the integrals of $P(\sigma_0 \dots \sigma_T)$ over variables $\sigma_{t'}$ where $t' > t+1$ can be performed immediately and give unity, reducing $P(\sigma_0 \dots \sigma_T)$ to $P(\sigma_0 \dots \sigma_{t+1})$. The integral of $(\sigma_{t+1} - \sigma_t)P(\sigma_0 \dots \sigma_{t+1})$ over σ_{t+1} gives $V(\sigma_t, \sigma_{t-1}; t)P(\sigma_0 \dots \sigma_t)$ after using Eq. (B3). Integrating $P(\sigma_0 \dots \sigma_t)$ with respect to all $\sigma_{t'}$ with $0 \le t' < t-1$ produces by definition $P(\sigma_{t-1}, \sigma_t)$. Let score s_t be in bin k for student in grade g and in year t. Then we can write the score velocity $v_{t,g,k}$ in this framework as

$$\begin{aligned} v_{t,g,k} &= \langle (\sigma_{t+1} - \sigma_t) \frac{\delta(\sigma_t - s_t)}{P(s_t)} \rangle \\ &= \int d\sigma_{t-1} d\sigma_t [\frac{P(\sigma_{t-1}, s_t)}{P(s_t)} \\ &\quad \times V(\sigma_t, \sigma_{t-1}; t)\delta(\sigma_t - s_t)] \\ &= \int d\sigma_{t-1} \frac{P(\sigma_{t-1}, s_t)}{P(s_t)} V(s_t, \sigma_{t-1}; t). \end{aligned}$$

The probability functions can be interpreted as the conditional probability of score σ_{t-1} given score s_t.

We are finally ready to compute the difference between trajectories and cohort streamlines. Suppressing the final argument t of the function V it is given by:

$$v_{t,g,k} - V(s_t, \mathsf{S}_{t-1}^{s,t}) = \int d\sigma_{t-1} \frac{P(\sigma_{t-1}, s_t)}{P(s_t)} [V(s_t, \sigma_{t-1}) - V(s_t, \mathsf{S}_{t-1}^{s,t})]. \quad \text{(B10)}$$

Expanding V in Eq. (B10) to first order in $\sigma_{t-1} - \mathsf{S}_{t-1}^{s,t}$ for the second argument gives

$$\begin{aligned} v_{t,g,k} &- V(s, \mathsf{S}_{t-1}^{s,t}) \quad \text{(B11)} \\ &\approx \frac{\partial}{\partial s'} V(s, s'; t)\Big|_{\mathsf{S}_{t-1}^{s,t}} \times \left(\bar{s}_{t-1}(s_t) - \mathsf{S}_{t-1}^{s,t}\right) \end{aligned}$$

and

$$\bar{s}_{t-1}(s_t) \equiv \int d\sigma_{t-1} \frac{P(\sigma_{t-1}, s_t)}{P(s_t)} \sigma_{t-1}.$$

The interpretation of $\bar{s}_{t-1}(s_t)$ is this: find students who got score s in year t, then find their mean score the year before. Now suppose that s_t is above the mean score. The students who got this score are a mixture of those who reproducibly get this score year after year and those who benefited from a positive random fluctuation. Therefore by the logic of regression to the mean, the score of this group the previous year is lower than one would expect from deterministic reasoning. That is, for s_t above the mean, $\bar{s}_{t-1}(s_t)$ is less than $\mathsf{S}_{t-1}^{s,t}$ and the term in parentheses is negative. Similarly, for s_t below the mean, the term in parentheses is positive. The first multiplicative factor on the right hand side of Eq. (B11) should be positive. The reason is that of two students with the same score this year, one should expect on average that the one with better scores the prior year will do better in future score changes. The bottom line is that when student scores next year depend in fact on the last two years, but one throws away information on the year before last (as in cohort flow plots) the result is that students regress to the mean more rapidly than they would be seen to do if one kept more information about them over time. This is why the streamlines in Figure 6 do not correspond well to the accurate trajectories in Figure 7.

[1] *No Child Left Behind Act of 2001*, Pub. L. no. 107-110, 115 Stat 1425 (2002).

[2] A. Usher (2011), URL `http://www.cep-dc.org/cfcontent_file.cfm?Attachment=Usher_Report_AYP2010-2011_121511.pdf`.

[3] D. Ravitch, *The Death and Life of the Great American School System* (Basic Books, 2010).

[4] S. Brill, *Class Warfare* (Simon and Schuster, 2011).

[5] National Research Council, *Scientific research in education.* (National Academies Press, Washington, 2002).

[6] National Research Council, *Advancing scientific research in education.* (National Academies Press, Washington, 2005).

[7] B. Schneider, M. Carnoy, J. Kilpatrick, W. Schmidt, and R. Shavelson, *Estimating causal effects using experimental and observational designs* (American Educational Research Association, Washington, DC, 2007).

[8] What Works Clearinghouse, *Find What Works* (2012, Retrieved April 2012 from), URL `http://ies.ed.gov/ncee/wwc/findwhatworks.aspx`.

[9] D. R. Thompson, S. L. Senk, D. Witonsky, Z. Usiskin, and G. Kaeley (2006).

[10] S. Ritter, J. Kulikowich, P. Lei, C. McGuire, and P. Morgan (2007).

[11] H. L. Schoen and C. R. Hirsch, *The Core-Plus Mathematics project: Perspectives and student achievement* (Lawrence Eribaum Associates, Hillsdale, NJ, 2002).

[12] L. Barrow, L. Markman, and C. E. Rouse, American Economic Journal: Economic Policy **1**, 52 (2009).

[13] B. J. Abrams, Ph.D. thesis, University of Colorado, Boulder (1989).

[14] J. J. Baker, Dissertation Abstracts International **58**, 2573A (1997).
[15] J. V. Cabalo, A. Jaciw, and M.-T. Vu, *Comparative effectiveness of Carnegie Learning's Cognitive Tutor Algebra I curriculum: A report of a randomized experiment in the Maui School District* (Empirical Education, Inc, Palo Alto, CA, 2007).
[16] Value-Added Research Center, Wisconsin Center for Education Research and NYC Department of Education, NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model (2010), URL http://bit.ly/J4FrfW.
[17] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics* (Butterworth and Heinemann, 1987), 2nd ed.
[18] Note that students in Texas have the option of taking a Spanish-language exam until seventh grade.
[19] Texas Education Agency (2011), URL http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147497081&libID=2147497078.
[20] V. H. Pham, Ph.D. thesis, The University of Texas at Austin (2009).
[21] Texas Education Agency, *Texas education agency – technical digest 2007-2008* (2008), URL http://bit.ly/PUkI0e.
[22] *Table 1. Intercensal Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2000 to July 1, 2010 (ST-EST00INT-01)* (U.S. Census Bureau, Population Division, September 2011).
[23] Texas Education Agency, University of Texas at Dallas Education Research Center, Gibson Consulting, and Learning Points Associates an affiliate of American Institutes for Research (2010), URL http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147495699&libID=2147495696.
[24] E. A. Hanushek and S. G. Rivkin, American Economic Review **100**, 267 (2010).
[25] Texas Education Agency (2009), URL http://ritter.tea.state.tx.us/opge/progeval/ReadingMathScience/SSI_ARI_AMI_Evaluation_2009.pdf.
[26] Texas Education Agency (2009), URL http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147484895&libID=2147484894.
[27] M. Marder and D. Bansal, PNAS **106**, 17267 (2009).
[28] E. M. Lifshitz and L. P. Pitaevskii, *Physical Kinetics* (Butterworth Heinemann, 1981).
[29] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Wadsworth Group/Thomson Learning, 1986).
[30] L. J. Cronbach and L. Furby, Psychological Bulletin **74**, 68 (1970).
[31] J. D. Singer and J. B. Willett, *Applied Data Longitudinal Analysis* (Oxford University Press, 2003).
[32] S. M. Stigler, *Statistics on the Table* (Harvard University Press, 1999), pp. 157-188.
[33] F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw Hill, 1965).