Longitudinal Predictions Using Regression-Corrected Grouping to Reduce Regression to the Mean

S. Stephens and M. Marder
Department of Physics
The University of Texas at Austin

Coarse-graining procedures make it possible to model the movement of large numbers of objects, such as particles in a fluid. Longitudinal student performance data can be modeled similarly by sorting students into score bins and following the flow of scores through time: trajectories depict the average scores over time for initial score bins and streamlines provide an approximate way to calculate the flow of student scores over many years based on only two consecutive years of data. However, due to the partially stochastic nature of observed scores, the coarse-graining procedure that sorts students into score bins amplifies a statistical phenomenon known as regression to the mean (RTM). As a result, streamlines do not provide an accurate prediction for the future performance of students. Here we discuss a new coarse-graining procedure, regression-corrected (RC) grouping, which reduces RTM in the streamlines. We apply RC streamlines to the Texas State Longitudinal Data System, which contains standardized testing data for students throughout primary and secondary school since 2003. We show that the RC streamlines accurately predict trajectories, using two or three years of data. Therefore RC streamlines can be used to identify the effects of academic interventions on a time scale comparable to that of policy changes. We illustrate this assertion by examining a particular policy intervention, Texas' Student Success Initiative.

I. INTRODUCTION

The No Child Left Behind Act of 2001 (NCLB) mandated widespread, high-stakes standardized testing for primary and secondary students throughout the United States [1]. Struggling students may be required to receive additional instruction to pass these exams and failure can result in grade retention [2]. Teachers evaluations and placements are often determined by the performance gains of their students on standardized exams [3–5]. Schools can be closed or restructured due to persistent low scores, whether overall or for disaggregated subgroups [6]. These accountability measures for students, teachers, and schools create a strong incentive for policy makers to find interventions that raise student performance. However, evaluating the consequences of legislative actions is difficult as the policies affect large numbers of students and may last only as long as an election cycle, whereas the impacts of the policies on valued metrics, such as high school graduation rates and college readiness, require up to 13 years to reach fruition.

A common way to analyze longitudinal datasets is to use hierarchical linear modeling or structural equation modeling to model outcomes with respect to observed variables or unobserved latent variables [7]. Marder and Bansal (2009) [8] and Bendinelli and Marder (2012) [9] developed an alternative approach that makes use of standard ideas from statistical mechanics. Trajectories can represent the position of any extended object over time. An example is the parabolic trajectory of a projectile. This is a useful idea especially when the object is composed of many correlated sub-units. It can be extended to fluid systems by coarse-graining the fluid at some point in time and following the motion of packets of fluid as the flow evolves. This provides a model for student score trajectories in an education context, showing average observed scores over time. A coarse-grained fluid is also frequently described by a velocity vector field that depicts the average velocity of particles in a fluid in each coarse-grained cell. Streamlines can be interpolated from the velocity vector field to show the anticipated movement of a particle throughout the space, given some initial position. Similarly, in the context of student scores, streamlines can be interpolated from a vector field that represents the change in score for each grade transition. Thus, streamlines depict the anticipated flow of scores over time. For fluids, streamlines are good approximations to trajectories when the flow is steady. In the education context, the testing environment must be stable for streamlines to provide a good estimate of trajectories, but this condition is only necessary and not sufficient.

Instead of analyzing individual student outcomes or the average outcomes for the aggregation of all students, outcomes are evaluated after coarse-graining into score bins by sorting students into groups by their percent score (90-100%, 80-90%, etc.) in some initial grade. When dealing with student scores, we will use the terms 'coarse graining' and 'binning' interchangeably.

A consequence of the partially stochastic nature of observed scores and the distribution of observed scores is the regression of the scores toward the mean score. This is exacerbated when students are binned

according to their observed scores. Regression to the mean (RTM) is present in both trajectories and streamlines; however, the multiple binning processes in streamlines exaggerate RTM, resulting in inaccurate depictions of student scores over time.

Here we present a new binning technique, called regression-corrected (RC) grouping, which reduces RTM in the streamlines so that they accurately represent the flow of student scores over time. RC streamlines can follow a single cohort of students over time, or they can be constructed using an accelerated longitudinal design to predict longitudinal outcomes with only two or three years of data. In this way, RC streamlines can predict longitudinal outcomes within a time suitable for informing policy; in addition, the comparison of streamlines from different periods can identify the effects of intermediate interventions.

Section II summarizes some techniques in education research that are currently used to analyze student performance over time. While RC grouping and streamlines could be used in many contexts, Section III discusses the setting used for this paper: the Texas State Longitudinal System and standardized testing in Texas. Section IV lays the foundation for the RC streamlines by summarizing the trajectory and streamline techniques described in Bendinelli and Marder (2012) [9]. Section V presents a simple theory for understanding the consequences of RTM, specifically in regards to trajectories and streamlines. Section VI describes the RC grouping technique and uses the same theory to demonstrate how RTM is reduced. Section VII applies RC streamlines to identify the effects of a state-wide intervention called the Student Success Initiative, while Section VIII explores additional applications including disaggregation by demographic factors, coarse-graining by alternate exams, and future predictions of student performance. In Section IX, we summarize our conclusions.

II. BACKGROUND

Longitudinal data analysis methods have been developed in many fields to study both within-person changes and between-person differences in outcomes over time [10]. Within-person changes refer to outcomes for a single individual and are particularly important for the analysis of educational policy impacts because they can establish a connection between an intervention and the resulting changes for the affected individuals. Between-person differences are also important to study because they can show how interventions affect students on a larger scale, and they can shed light on the differential impacts of an intervention on different groups of students, which can lead to more efficiently targeted or more equitable interventions.

Most longitudinal studies utilize a statistical technique known as hierarchical linear modeling [11, 12], or a similar technique called structural equation modeling [13–15], to analytically describe both the within-person changes and the between-person differences in the longitudinal data. The analytical model may use observed variables or unobserved latent variables combined with fitted parameters to best represent the growth of the outcome variable. Individual growth curves represent the within-person component of the model, and person-specific parameters (often within a normal distribution of values) that adjust the intercepts or slopes account for the between-person variation. The model often takes the form of a linear, polynomial, or piece-wise function, although it can have any non-parametric form.

This idea can be extended to grouped individuals through techniques such as group-based trajectory modeling [16] and latent class growth modeling [17]. These techniques assume that the population is mixed, containing several distinct groups that are categorized by a latent variable. Using the longitudinal data, individuals are first sorted into groups based on similar initial growth patterns and then their average outcomes over time are modeled with a trajectory. Intervention effects can be compared for treated groups and control groups within the same pre-treatment growth trajectory. This technique requires several years of data both before and after an intervention to establish comparison groups and then to observe the intervention effects.

The regression techniques in hierarchical linear modeling or structural equation modeling are familiar to statisticians. However, the language, equations, and coefficients in these frameworks can be difficult to understand for policy makers and educators. These techniques often involve computational packages such as HLM [18] and Mplus [19], which can act as a black box, potentially leading to misuse or misinterpretation. The techniques in this paper are designed to be more intuitive to policy makers and educators by using visualizations rather than relying on parametric solutions with tables of computed coefficients. Nonetheless, the standard regression methods provide a foundation and a source of comparison for the new methods described in this paper.

A. Age-Period-Cohort Effects

We now develop some of the concepts needed in order to discuss educational data gathered over time. Between-person differences and within-person changes in the data can be observed as any of three time-related variations: age effects, period effects, and cohort effects. Age effects represent changes related to aging although in the context of education, age effects could be substituted by grade effects, changes that occur between each grade. Period effects represent changes occurring during a specific time period, affecting people of all ages similarly. Cohort effects represent formative experiences, changes that are unique to people who experience the same events at the same time, often because they were born in the same time period. In an education context, cohort effects may relate to changes that affect a cohort of students progressing through school together, completing one grade each year and graduating in the same year.

It can be difficult to separate age, period, and cohort effects from each other. This is due to the inherent relationship between age, time, and cohort. For studies using birth cohorts—people born in the same time period—the relationship between age, year, and birth cohort is $Year - Age = Birth\ Cohort$. For students in primary or secondary school, the relationship between the year, grade, and graduating cohort is $Year - Grade + 12 = High\ School\ Graduation\ Cohort$. These relationships make it difficult to design a study to isolate grade, period, or cohort effects because it is difficult to control for more than one of these effects. In addition, there may be several concurrent influences, causing a combination of grade, period, and cohort effects. Educational policy changes are usually either period effects or cohort effects, depending on the process of implementation and the intended recipients.

To minimize or isolate the potential sources of time-related change, studies may focus on a single time period or a single cohort. Studies analyzing data from one time period are known as cross-sectional models. By using data from only one time, period effects are removed from the analysis and the time required for data collection is minimized, an attractive feature for costly studies. Cross-sectional models show the range of outcomes at a moment in time and can be helpful for identifying between-person variation. The students within a cross-section can be aggregated into a *synthetic cohort*, which contains students at every grade throughout school in a single year. If there were no cohort effects, the outcomes from a synthetic cohort would accurately identify grade effects. However, without other time periods for comparison, grade effects and cohort effects are completely confounded. Extending the study longitudinally can help to separate the grade effects from the cohort effects, although this can then introduce period effects.

By contrast, single-cohort studies use the longitudinal data from a single graduating cohort of students to study within-person change [20]. Cohort studies are helpful for studying the evolution of individuals, especially considering that a person may have many experiences that build upon one another to cause a particular growth pattern. Single-cohort studies require several data points for each individual, which can take years of data collection and can suffer from attrition. In addition, while cohort studies remove cohort effects from the analysis, grade effects and period effects are confounded. By comparing outcomes from multiple cohorts, grade effects and period effects can be separated, while possibly reintroducing cohort effects.

Accelerated longitudinal design (ALD) studies are a compromise between cross-sectional studies and longitudinal single-cohort studies [21]. ALDs use data from multiple overlapping cohorts beginning at different ages to span a large age range while using only a few years of data. For example, Miyazaki and Raudenbush used the National Youth Survey that contained data for 7 adjacent cohorts over 5 years to study the development of antisocial attitudes from ages 11 to 21 [22]. ALDs study growth over a large age/grade range without needing to wait the full time period as in longitudinal studies. This reduces the cost of the study as well as the attrition due to missing data, while producing results in a time period that allows for more political influence. The techniques discussed in this paper can be categorized as single cohort designs or ALDs.

III. SETTING

Although we expect the methods of this paper to have general utility, the possibility of testing them is due to a unique resource provided by the Texas Education Research Center (ERC), which holds extensive longitudinal information about Texas students, teachers, and schools [23]. The Texas Education Agency (TEA) provides the ERC with much of their student-level data, including demographics and standardized testing information. Starting in 2003, the state-wide exam was the Texas Assessment of Knowledge and Skills (TAKS), which consisted of annual mathematics and reading exams between 3rd and 11th grade, as well as periodic exams in other subjects [24]. In 2012, the exams started transitioning to the State of Texas Assessments of

Academic Readiness (STAAR), which had mostly the same schedule of exams before high school but replaced grade-specific high school exams with end-of-course subject exams [25, 26]. To use a consistent metric, we mostly limit the focus of this paper to the TAKS mathematics exams (2003-2012), except for some future predictions that we make with the STAAR data.

In our research, we have decided to use the raw percent scores instead of the scaled scores provided by the TEA. While, psychometrically, scaled scores are preferable to raw scores when comparing the results from multiple exams, we feel justified in using raw scores for several reasons. First, the exams were compiled from banks of items that were designed and tested to be similar in content and difficulty [27]. Each year the exams were designed to be very similar, resulting in a consistent raw score metric. Second, the raw score conversion tables [28] were examined and the raw scores corresponding with a passing score within each grade for the mathematics TAKS exams was fairly consistent; the score usually varied by only one question, occasionally by two questions, and rarely by three (out of approximately 40-50 questions). The maximum possible raw score within a grade and subject was consistent over time. Third, the TEA used a one-parameter logistic function as the item-response theory [27], which resulted in a one-to-one mapping of the raw scores to scaled scores; in this model, students with the same raw score have the same "ability". Fourth, the iterative process of estimating the abilities using the one-parameter logistic model requires proprietary software designed for item response theory calculations and student-level data for specific test items that are often unavailable for researchers. Fifth, the TEA converted raw to scaled scores using scaling constants with a proprietary method based not only on item response theory, but also subject to the constraints that a scaled score of 2100 be a panel-recommended passing score, and 2400 a panel-recommended commended score. Sixth, the TEA's scaling method switched from a horizontal scaling method (comparable between years within grade and subject) to a vertical scaling method (comparable between years and grades within subject) in 2008 due to changes in the Texas Education Code [6]; this resulted in a large shift in scaled score ranges, without the ability to track longitudinal progress before 2008. Therefore, on the basis of simplicity, transparency, and consistency, we have chosen to use raw score percentages as our metric.

IV. FOUNDATION

The methods detailed in Bendinelli and Marder (2012) [9] establish a foundation for the RC streamlines developed in this paper. We briefly describe them again here to provide a background for the modifications that follow.

A. Trajectories

In physics, trajectories are used to represent the position of a physical object over time. In an education context, trajectories can be used to represent scores over time. Using a coarse-grained binning procedure, students are sorted into score bins determined by their TAKS mathematics score in 3rd grade, the earliest exam in primary school. Specifically, the score bin limits are determined by the raw percent scores (90-100%, 80-90%, etc.). Grouping students into decile score bins provides a compromise between an overwhelming number of individual student trajectories and a single aggregate trajectory, which would fail to convey most of the information available in the data. Once the groups have been established, the average scores for each group are plotted for each grade, up to the exit mathematics exams in 11th grade. Therefore, trajectories are single-cohort studies, depicting the observed average scores for the score bins over time. Decile score bins were chosen in part because they are familiar and correspond to intuition about 'A', 'B', 'C', 'D', and 'F' students. For a discussion of grouping from a more general point of view, see Guthrie (2018).[29] Only by grouping students according to prior academic performance and other characteristics are we able to understand effects of educational policy changes, for these work differently on students of different backgrounds. However, as we will see, the process of grouping creates a set of technical problems that we have set ourselves in this paper to resolve.

Figure 1 shows the trajectories for the cohort of students who graduated high school in 2012. The thickness of each line is proportional to the number of students in that score bin. The average scores in each grade are connected by linear segments, although the connection does not need to be linear. The performance peaks in 5th and 8th grade will be discussed in Section VII; otherwise, the trajectories are fairly flat, without intersections. This is not surprising when averaging thousands of students, as prior performance is the strongest indicator of future performance. Students could be further aggregated by demographic variables or course taking, for example, which would likely result in more movement

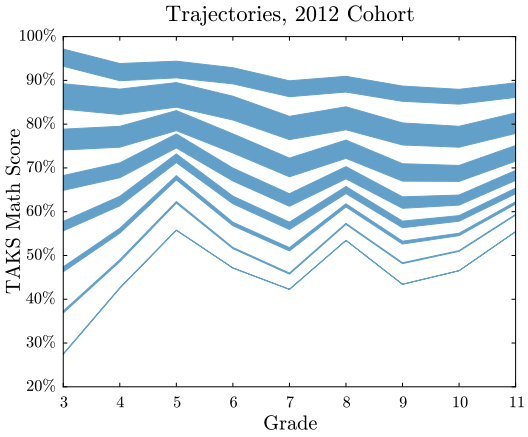

FIG. 1. Trajectories for the cohort of students who graduated in 2012. The students are sorted into percent score deciles by their maximum 3rd grade mathematics TAKS score in 2003. The average scores for these groups of students are then plotted for the subsequent grades in the following years, and these average scores are connected in linear segments. The thickness of the trajectory is proportional to the number of students in that group. There are approximately 250,000 students included in this analysis.

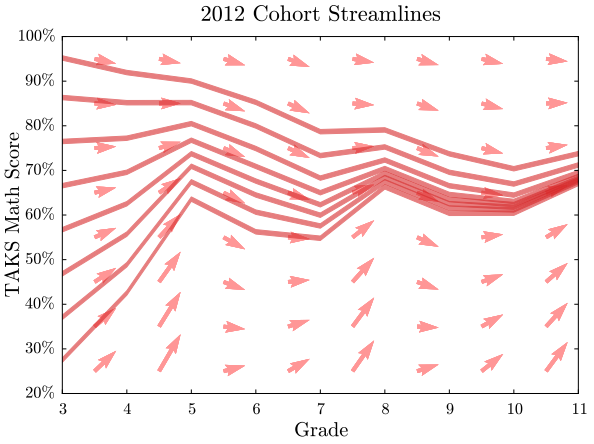

FIG. 2. Vector field of score changes and corresponding streamlines for the cohort of 2012. The students are sorted into new bins each grade, and the average change in score for each group is calculated from that grade to the next. The slope of an arrow is equal to the change in score for that grade and bin. Streamlines are constructed, beginning with the average score in 3rd grade for each bin and then interpolating changes in score from the vector field to determine the slope of each segment of the streamline. These streamlines should capture the flow of student scores throughout the grades, however, the evident convergence of the streamlines due to RTM prevents this method from producing more meaningful results.

in the trajectories (see Section VIII).

Trajectory plots are of fundamental interest because they track the average performance of a cohort of students at all grades and performance levels exactly. Therefore, trajectories provide an accurate depiction of longitudinal student performance on a large scale and they can be used to analyze the outcomes of policies in the long term. However, policies typically have a shorter duration than the nine years of data that it takes to construct a full trajectory, so educational policy has likely already changed by the time the results can be analyzed with this method. In addition, attrition can introduce bias to the sample and this becomes more of an issue with longer studies. It is therefore necessary to identify other techniques that permit more timely analysis.

B. Streamlines

Streamlines are used in fluid mechanics to represent the motion of particles in a fluid. Velocity vector fields are constructed from the average velocity of particles at various positions in a fluid. Streamlines begin at an initial position and then follow the flow created by the velocity vector field, conveying the anticipated motion of a particle in the fluid from the starting position. In an education context, vector fields can be constructed from the change in score over time to represent the flow of scores through the grades. Streamlines then represent the anticipated scores over time for a student with an observed initial score.

Streamlines can utilize a single cohort design or an ALD. *Cohort streamlines*, as in Figure 2, use data from a single cohort of students as they progress through school. The students are sorted into score bins in each grade, and the change in score is calculated for those groups from that grade to the next. This differs from a trajectory where the students are sorted in only the initial grade, as the students in streamlines are re-grouped into score bins in each grade. The changes in score for each score bin and grade transition comprise a vector field from which streamlines can be interpolated. The streamlines begin at the average scores for each score bin in 3rd grade and then the slope of each segment is determined by a linear interpolation function for that grade transition, which is derived from the observed score changes (it does not need to be linear).

Snapshot streamlines, as in Figure 3, use an ALD with data from eight sequential cohorts, following each for two consecutive years. Snapshot streamlines are constructed in the same way as the cohort streamlines, except each grade transition is represented by a different cohort of students. In the first

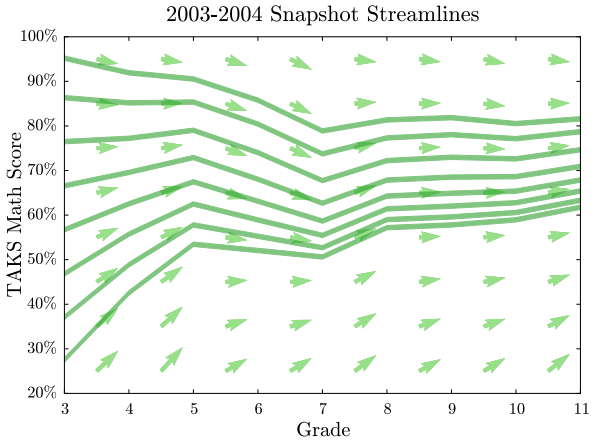

FIG. 3. Vector field and corresponding streamlines for each grade transition between 2003 and 2004. The students in the synthetic cohort are sorted into groups according to their score in 2003, and the average change in score for each group is calculated from 2003 to the subsequent grade in 2004. The slope of an arrow is equal to the average change in score between the two neighboring grades for that bin. The streamlines begin with the average score for each bin in 3rd grade and then follow the flow designated by the vector field. The convergence is a result of RTM.

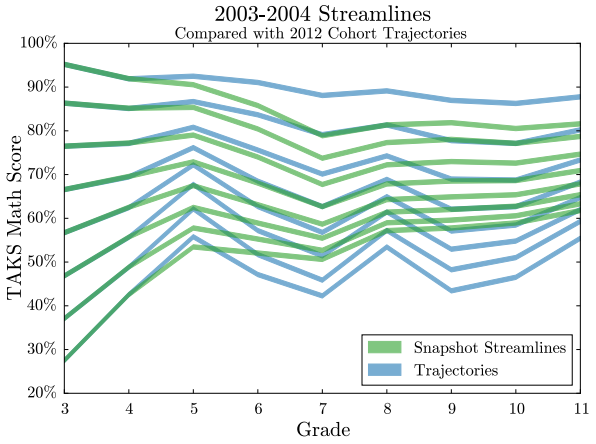

FIG. 4. Snapshot streamlines for 2003-2004 and trajectories for the cohort of 2012. The snapshot streamlines represent the synthetic cohort of 2003, whereas the trajectories represent the cohort of 2012. In the absence of cohort effects, the snapshot streamlines could be used to predict the longitudinal performance of the students, which is represented by the trajectories. The convergence of the streamlines due to RTM makes the predictions from the snapshot streamlines inaccurate.

year, the students from each cohort are sorted into score bins in their grade and the score change is calculated for those groups from that year to the next, when the students have progressed to the next grade. Therefore snapshot streamlines represent the score flow throughout the grades while using only two years of data. Assuming that different cohorts perform similarly, snapshot streamlines can be used to predict longitudinal results from a single cohort.

Figures 4 and 5 show the snapshot streamlines for 2003-2004 and 2012 cohort streamlines, each compared to the trajectories for the cohort of 2012. It is evident in both cases that the streamlines are converging, differing from the relatively flat trajectories. The trajectories directly plot the average observed scores, so the discrepancies between the trajectories and streamlines points to an issue with the streamlines, as they do not accurately represent the flow of scores through the grades. This inaccuracy is due to RTM. The next section presents a simple theory to explain why the iterative score binning process is causing strong RTM in the streamlines, and then we present a simple solution to create more accurate streamlines.

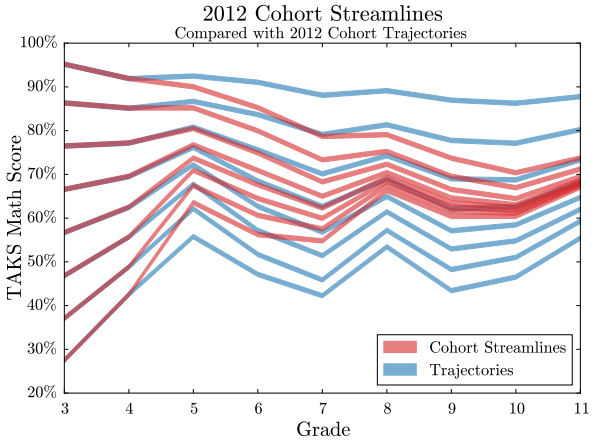

FIG. 5. Cohort streamlines and trajectories for the cohort of 2012. Both plotting schemes are representing the same cohort of students, however the streamlines are more inclusive of students who join the cohort after 3rd grade. Due to RTM, the streamlines converge considerably compared to the trajectories. The streamlines therefore do not accurately represent the flow of student scores.

V. REGRESSION TO THE MEAN

Regression to the mean (RTM) is a bulk statistical phenomenon, and is a consequence of the stochastic component of the observed scores. RTM depends on the magnitude of these random fluctuations and

on the score distributions. RTM is exacerbated by selection or classification processes. In particular, by sorting students into score bins by their observed scores on a single exam, the random component of their score may force the student into a different bin than expected. For a score distribution with more students performing near the mean than at the extremes, most of the "wrongly binned" students typically perform closer to the mean of the distribution. On a second exam, the average for the score bin will regress toward the mean as the expected scores are closer to the distribution's average.

We use a simple theory to demonstrate the influences of RTM on the trajectories and streamlines with the understanding that observed data are never so simple, but general behavior is best elucidated in the simple case. Using conditional expectation values for the exam scores [30], we can better understand the significance of RTM in the coarse-graining procedures. We invoke classical test theory [31, 32] to establish the relationship between the observed score, the true score, and the random component. If x_i are the raw scores for exam i, then we can say that $x_i = t_i + e_i$ where t_i is the true score and e_i is the random component, or error score. The true score, $E(x_i) = t_i$, is unobserved and is defined to be the expected score if a student were to be tested repeatedly to average over short-term fluctuations due to temporary circumstances such as a bad night's sleep. The error scores for the repeated measurements come from a normal distribution with a mean of zero, and they are uncorrelated with each other and with the true score. We can also define z-scores, z_i, such that $z_i = (x_i - \mu)/\sigma_{x_i}$, where σ_{x_i} is the standard deviation of x_i and μ is the mean raw score or expected score, $E(x_i) = \mu$. For z-scores, $E(z_i) = 0$ and $\sigma_{z_i} = 1$ for all i.

To make concrete computations, we assume that student scores on exams in successive years are linearly correlated. This assumption holds for a bivariate normal distribution, and it can also hold for non-normal distributions. Figure 6 shows the score distributions and bivariate distributions for the 3rd, 4th, and 5th grade scores of the 2012 cohort, as examples of typical distributions in the ERC dataset. All of the score distributions are skewed toward higher scores, with a ceiling effect from the maximum score. The 5th and 8th grade distributions, in particular, had floor effects due to a program called Student Success Initiative (see Section VII). Despite these floor and ceiling effects, the assumption of linear conditional expectation values is not unreasonable.

Assuming linear conditional expectation values, the expected score on exam y given a score a on exam x is:

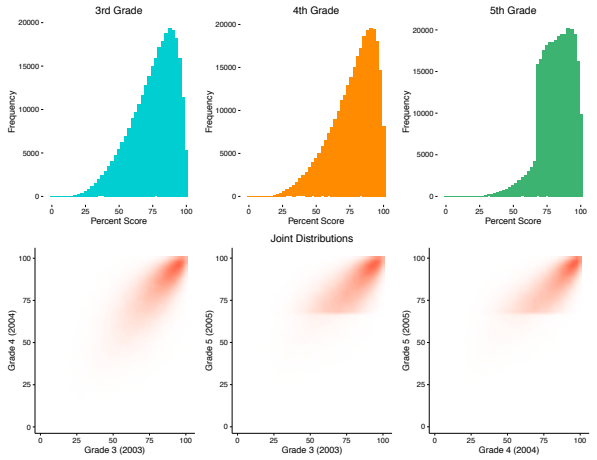

FIG. 6. Score distributions for TAKS mathematics exams in 3rd, 4th, and 5th grade, and correlations between students' scores for each grade pair. The score distributions for most of the exams are skewed as a result of the ceiling effects from the maximum score. The 5th grade and 8th grade exams have an additional floor effect due to the Student Success Initiative, discussed in Section VII. The relationship between the scores for most exams is roughly linear.

$$E(y|x = a) = E(y) + \rho_{x,y}\frac{\sigma_y}{\sigma_x}(a - E(x)), \quad \text{(V.1)}$$

where $\rho_{x,y}$ is the Pearson correlation coefficient between the exams and σ is the standard deviation. For z-scores, this expression simplifies to:

$$E(z_y|z_x = a) = \rho_{z_x,z_y}a. \quad \text{(V.2)}$$

Following from the Cauchy-Schwarz inequality, $|\rho_{z_i,z_j}| \leq 1$. Therefore, this expression explains RTM; given the score on an exam, the expected score on another exam is closer to zero, the mean for z-scores. The less correlated the exams, the more RTM.

This theory can be expanded to multiple years in several ways. Trajectories are modeled by conditioning every exam by the initial exam, similar to how the students are sorted into bins by their initial scores. The necessary conditional expectation values are:

$$E(z_2|z_1 = a) = \rho_{z_1,z_2}a$$
$$E(z_3|z_1 = a) = \rho_{z_1,z_3}a$$
$$E(z_4|z_1 = a) = \rho_{z_1,z_4}a$$
$$\ldots$$

The sequence of expected scores is shown in Table I and depends on the correlation coefficients between

Trajectories		Streamlines	
Expression	Equivalent Correlation Coefficients	Expression	Equivalent Correlation Coefficients
$\varsigma_3 = a$	a	$\varsigma_3 = a$	a
$\varsigma_4 = \rho_{3,4}a$	ρa	$\varsigma_4 = \rho_{3,4}a$	ρa
$\varsigma_5 = \rho_{3,5}a$	ρa	$\varsigma_5 = \rho_{4,5}\rho_{3,4}a$	$\rho^2 a$
$\varsigma_6 = \rho_{3,6}a$	ρa	$\varsigma_6 = \rho_{5,6}\rho_{4,5}\rho_{3,4}a$	$\rho^3 a$
$\varsigma_7 = \rho_{3,7}a$	ρa	$\varsigma_7 = \rho_{6,7}\rho_{5,6}\rho_{4,5}\rho_{3,4}a$	$\rho^4 a$

TABLE I. Anticipated sequences of scores for trajectories and streamlines with respect to the initial score. The full expression and the expression for equivalent correlation coefficients are given for several exams.

each exam and the first exam. If it is assumed that all of the exam pairs have the same correlation coefficient $\rho_{z_i,z_j} = \rho$, this sequence would become $a, \rho a, \rho a, \rho a$, etc. Therefore the RTM takes place during the first grade transition but not thereafter. The assumption of equivalent correlation coefficients is not completely borne out in the data, as exams tend to be less correlated with time, but it is illustrative of the basic patterns of RTM.

Streamlines are constructed using a first-order Markov process, with each exam being conditioned only by the previous exam. This process is described by the following conditional expectation values:

$$\begin{aligned} E(z_2|z_1 = a) &= \rho_{z_1,z_2}a \\ E(z_3|z_2 = b) &= \rho_{z_2,z_3}b \\ E(z_4|z_3 = c) &= \rho_{z_3,z_4}c \\ &... \end{aligned}$$

To create a continuous streamline, the score for one exam is used as the initial condition for the next, so that the anticipated values for the scores are

$$\begin{aligned} &\varsigma_1 = a \\ &\varsigma_2 = \rho_{z_1,z_2}a \\ &\varsigma_3 = \rho_{z_2,z_3}\varsigma_2 = \rho_{z_2,z_3}\rho_{z_1,z_2}a \\ &\varsigma_4 = \rho_{z_3,z_4}\varsigma_3 = \rho_{z_3,z_4}\rho_{z_2,z_3}\rho_{z_1,z_2}a \\ &... \end{aligned}$$

For streamlines, the scores are proportional to the product of previous correlation coefficients. For equivalent correlation coefficients ρ, the sequence becomes $a, \rho a, \rho^2 a, \rho^3 a$, etc. It is obvious in this case that the scores continually regress towards the mean of zero. The sequences for trajectories and streamlines can be compared in Table I. While this theory is simple and requires assumptions that may not hold with observed scores, Figure 5 exhibits RTM mainly in the first segment of the trajectories but throughout the streamlines, aligning with the observations from the theory.

VI. REGRESSION-CORRECTED GROUPING

The sequence of anticipated scores in the trajectory framework demonstrates the magnitude of RTM each year after a binning process. The majority of the RTM occurs between the binning exam and the next, in the first grade transition. Streamlines are essentially several short trajectories strung together; each segment is the first grade transition in a new trajectory. Since this is exactly the portion of the trajectory that contains the majority of the RTM, it is no surprise that the streamlines have such extreme RTM. However, the RTM in the trajectories is mostly resolved by the second grade transition. Therefore, if the streamlines instead utilized the second grade transition for each short trajectory, the RTM would have already been mostly resolved. This is the premise of the new binning procedure, called regression-corrected (RC) grouping.

For the uncorrected streamlines described above, students are sorted by their scores in grade i for $i \in [3, 10]$ and the average change in score is calculated for those groups between grades i to $i+1$. The scores from the binning exams are included in the change-in-score calculations. With RC grouping, we delay the change-in-score calculation by one year. In RC streamlines, students are sorted by their scores in grade i for $i \in [3, 9]$ and the changes in score are calculated for those groups between grades $i + 1$ to $i + 2$. Excluding the binning exam scores from the change-in-score calculations reduces the magnitude of RTM. After the vector field is established from the change-in-score calculations, continuous streamlines are constructed. In essence, the students for each grade transition are sorted into score bins by a separate exam from the exams used to calculate the change in score. We have chosen to use the previous exam of the same subject (mathematics) but it could be any other exam that is reasonably well correlated with the exams used to calculate the change in score. An example of using the reading score for sorting is discussed in Section VIII and it demon-

strates that RC streamlines can be constructed with only two years of data.

We can analyze RC streamlines using the conditional expectation values, as above. For each binning process, we calculate a pair of expectation values conditioned on the binned scores. The necessary pairs of expectation values are:

$$\begin{cases} E(z_2|z_1 = a) = \rho_{z_1,z_2} a \\ E(z_3|z_1 = a) = \rho_{z_1,z_3} a \end{cases}$$

$$\begin{cases} E(z_3|z_2 = b) = \rho_{z_2,z_3} b \\ E(z_4|z_2 = b) = \rho_{z_2,z_4} b \end{cases}$$

$$\begin{cases} E(z_4|z_3 = c) = \rho_{z_3,z_4} c \\ E(z_5|z_3 = c) = \rho_{z_3,z_5} c \end{cases}$$

...

The differences of the two scores in each pair are used as the slopes for each of the vectors to construct the RC streamlines. A continuous streamline is strung together by using the previous score as the initial condition for the next pair. The initial conditions (b, c, etc.) can be calculated by setting the expectation values for the same exam equal to each other. For example, we would need to find b such that $E(z_3|z_1 = a) = E(z_3|z_2 = b)$. This gives $b = \frac{\rho_{z_1,z_3}}{\rho_{z_2,z_3}} a$, which then allows us to calculate $E(z_4|z_2 = b)$. Using this method iteratively gives a sequence of scores in the RC streamlines of

$$\begin{aligned}
&\varsigma_1 = a \quad \text{(Not shown on the plot)} \\
&\varsigma_2 = E(z_2|z_1 = a) = \rho_{z_1,z_2} a \\
&\varsigma_3 = E(z_3|z_1 = a) = \rho_{z_1,z_3} a \\
&\varsigma_4 = E(z_4|z_2 : E(z_3|z_2) = \varsigma_3) \\
&\quad = E(z_4|z_2 : \rho_{z_2,z_3} z_2 = \rho_{z_1,z_3} a) \\
&\quad = E(z_4|z_2 = \frac{\rho_{z_1,z_3}}{\rho_{z_2,z_3}} a) = \frac{\rho_{z_2,z_4}\rho_{z_1,z_3}}{\rho_{z_2,z_3}} a \\
&\varsigma_5 = E(z_5|z_3 : E(z_4|z_3) = \varsigma_4) \\
&\quad = E(z_5|z_3 : \rho_{z_3,z_4} z_3 = \frac{\rho_{z_2,z_4}\rho_{z_1,z_3}}{\rho_{z_2,z_3}} a) \\
&\quad = E(z_5|z_3 = \frac{\rho_{z_2,z_4}\rho_{z_1,z_3}}{\rho_{z_3,z_4}\rho_{z_2,z_3}} a) = \frac{\rho_{z_3,z_5}\rho_{z_2,z_4}\rho_{z_1,z_3}}{\rho_{z_3,z_4}\rho_{z_2,z_3}} a \\
&\ldots \\
&\varsigma_n = \frac{\rho_{z_{n-2},z_n}\rho_{z_{n-3},z_{n-1}}\rho_{z_{n-4},z_{n-2}}\cdots\rho_{z_1,z_3}}{\rho_{z_{n-2},z_{n-1}}\rho_{z_{n-3},z_{n-2}}\cdots\rho_{z_2,z_3}} a \\
&\quad = \frac{\prod_{p=1}^{n-2}(\rho_{p,p+2})}{\prod_{q=2}^{n-2}(\rho_{q,q+1})} a \quad (\forall n \geq 4)
\end{aligned}$$

Scores for Equivalent Correlation Coefficients			
	Trajectories	Streamlines	Regression-Corrected Streamlines
ς_3	a	a	a
ς_4	ρa	ρa	ρa
ς_5	ρa	$\rho^2 a$	ρa
ς_6	ρa	$\rho^3 a$	ρa
ς_7	ρa	$\rho^4 a$	ρa
ς_8	ρa	$\rho^5 a$	ρa
ς_9	ρa	$\rho^6 a$	ρa
ς_{10}	ρa	$\rho^7 a$	ρa
ς_{11}	ρa	$\rho^8 a$	ρa

TABLE II. Comparison between the score sequences within the trajectory, cohort streamline, and RC cohort streamline frameworks, assuming every pair of exams has the same correlation coefficient.

In this sequence, the coefficients of the scores are ratios of Pearson correlation coefficients between the exams. Again, if we were to assume the same correlation coefficients between all of the exams, the values in the sequence would be $a, \rho a, \rho a, \rho a$, etc., which is identical to the sequence for the trajectory.

Table II shows the sequences for the special case where the correlation coefficients are the same between every exam. The trajectory and RC streamline sequences only have RTM between the 3rd and 4th grade exams, whereas the streamline sequence continuously regresses toward the mean. Despite the unlikely assumption of equivalent correlation coefficients, this is an indication that the RC streamlines may successfully reduce RTM. Table III shows the expressions and values for the expected scores in the trajectory, streamline, and RC streamline frameworks, using the Pearson correlation coefficients from the ERC dataset for the cohort of 2012. While the correlation decreases as the exams are further apart in time, the RTM in the trajectories and RC streamlines are nearly identical, whereas the uncorrected streamlines exhibit extreme RTM.

Figure 7 shows the RC cohort streamlines and the trajectories for the cohort of students that graduated in 2012. Overall, the two frameworks have very similar results. The slight differences between the two frameworks are meaningful; RC streamlines can include students who join the cohort after 3rd grade. Trajectories require students to have a 3rd and 4th grade score whereas RC streamlines require any three consecutive scores (grades i, $i+1$, and $i+2$) within the cohort to be included for a segment (from grade $i+1$ to $i+2$). Therefore, trajectories are more inclusive of joiners before 5th grade and RC streamlines are more inclusive after 5th grade. In the

Trajectories		Streamlines		Regression-Corrected Streamlines	
Expression	Value	Expression	Value	Expression	Value
$\varsigma_3 = a$	a	$\varsigma_3 = a$	a	$\varsigma_3 = a$	a
$\varsigma_4 = \rho_{3,4}a$	$.73a$	$\varsigma_4 = \rho_{3,4}a$	$.73a$	$\varsigma_4 = \rho_{3,4}a$	$.73a$
$\varsigma_5 = \rho_{3,5}a$	$.67a$	$\varsigma_5 = \rho_{4,5}\varsigma_4$	$.53a$	$\varsigma_5 = \rho_{3,5}a$	$.67a$
$\varsigma_6 = \rho_{3,6}a$	$.65a$	$\varsigma_6 = \rho_{5,6}\varsigma_5$	$.38a$	$\varsigma_6 = \frac{\rho_{4,6}}{\rho_{4,5}}\varsigma_5$	$.64a$
$\varsigma_7 = \rho_{3,7}a$	$.63a$	$\varsigma_7 = \rho_{6,7}\varsigma_6$	$.30a$	$\varsigma_7 = \frac{\rho_{5,7}}{\rho_{5,6}}\varsigma_6$	$.62a$
$\varsigma_8 = \rho_{3,8}a$	$.62a$	$\varsigma_8 = \rho_{7,8}\varsigma_7$	$.23a$	$\varsigma_8 = \frac{\rho_{6,8}}{\rho_{6,7}}\varsigma_7$	$.59a$
$\varsigma_9 = \rho_{3,9}a$	$.58a$	$\varsigma_9 = \rho_{8,9}\varsigma_8$	$.18a$	$\varsigma_9 = \frac{\rho_{7,9}}{\rho_{7,8}}\varsigma_8$	$.57a$
$\varsigma_{10} = \rho_{3,10}a$	$.58a$	$\varsigma_{10} = \rho_{9,10}\varsigma_9$	$.15a$	$\varsigma_{10} = \frac{\rho_{8,10}}{\rho_{8,9}}\varsigma_9$	$.56a$
$\varsigma_{11} = \rho_{3,11}a$	$.54a$	$\varsigma_{11} = \rho_{10,11}\varsigma_{10}$	$.12a$	$\varsigma_{11} = \frac{\rho_{9,11}}{\rho_{9,10}}\varsigma_{10}$	$.53a$

TABLE III. Comparison between the score sequences within the trajectory, cohort streamline, and RC cohort streamline frameworks, using the Pearson correlation coefficients computed from the data for the cohort of 2012.

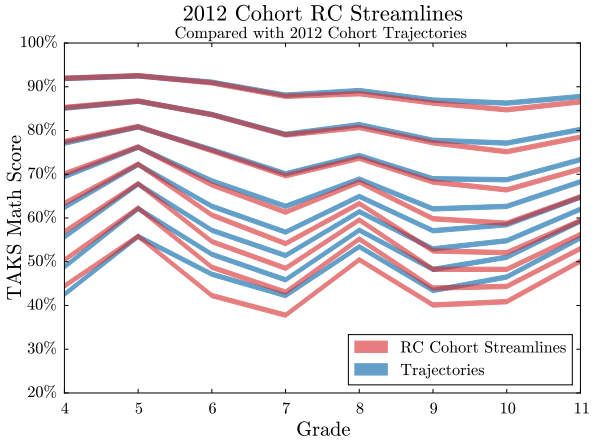

FIG. 7. RC cohort streamlines and trajectories for the cohort of 2012. The RC streamlines and the trajectories correspond very well. The slight differences are due to the students who join or leave the cohort. Before 5th grade, trajectory plots are slightly more inclusive of joiners. After 5th grade, RC cohort streamlines are more inclusive.

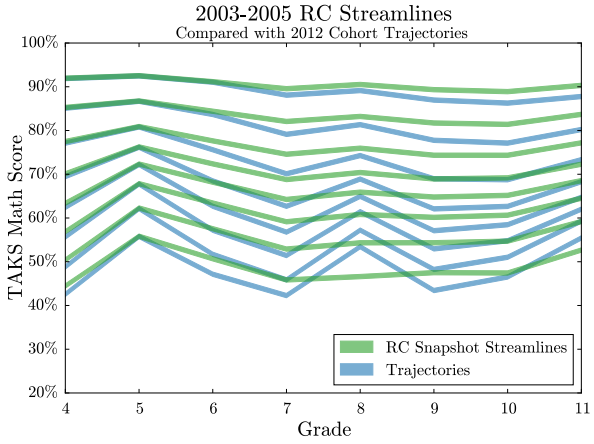

FIG. 8. RC snapshot streamlines for 2003-2005 and trajectories for the 2012 cohort. The 8th grade peaks observed in the lower-performing trajectories (2008), which are missing in the RC snapshot streamlines (2003-2005), are a result of the Student Success Initiative and the Accelerated Mathematics Instruction, which was provided to low-performing 8th grade students starting in 2008.

data, this corresponds with lower performance in the RC streamlines after 5th grade, as mobility has been shown to be correlated with lower performance [33]. Not only do RC cohort streamlines produce accurate longitudinal results, but they can better capture the performance of temporary cohort members.

RC streamlines can also utilize an Accelerated Longitudinal Design, reducing the required years of data to only three years. When using mathematics scores exclusively, the students are sorted by their score in a first year and then the changes in score are calculated between a second and third year. Figure 8 shows the RC snapshot streamlines for 2003-2005 compared to the trajectory for the cohort of 2012. While the two frameworks are fairly comparable, there is an 8th grade performance peak in the trajectories that is missing in the RC snapshot streamlines. Again, this difference is meaningful and it demonstrates a change in policy that affected the 8th graders by 2008 (trajectories) but not in 2005 (RC snapshot streamlines) called the Student Success Initiative.

VII. APPLICATION: SSI

The Student Success Initiative (SSI), established in 1999, encompasses several initiatives intended

to promote grade-level performance in mathematics and reading in Texas. One component of the SSI is a set of grade promotion requirements in 5th and 8th grade; if students fail the mathematics or reading exam three times then they will be retained in 5th or 8th grade for another year unless they are excused by a committee [2]. In addition, struggling students in every grade, especially in the high-stakes 5th and 8th grades, can receive additional instruction called Accelerated Mathematics/Reading Instruction (AMI/ARI) [34].

SSI was implemented gradually along with the cohort of 2012. In the 1999-2000 school year, SSI only impacted Kindergarten students, and the program expanded to include an additional grade each year [35]. Therefore, the first 5th grade retention and ARI/AMI occurred in the 2004-2005 school year, and the first 8th grade retention and ARI/AMI occurred in the 2007-2008 school year. This timeline of implementation may explain the similarities and differences in performance in Figure 8. The students in the trajectories were the first cohort of students to be affected by SSI in each grade. The students in the RC snapshot streamlines were only impacted by SSI up through 5th grade. Therefore, a performance peak can be seen as a result of the 5th grade high-stakes exam in both frameworks, but the 8th grade peak is only seen in the trajectories.

Figure 9 shows the RC snapshot streamlines from 2007-2009, after the SSI impacted students up to 8th grade. The 8th grade performance peak is evident in both the trajectories and the RC snapshot streamlines. It should be mentioned that the funding for SSI has been reduced dramatically in recent years, dropping from over $150 million annually in 2011 [35] to $5.5 million in the 2018-2019 biennium budget [36].

Overall, the RC snapshot streamlines and trajectories are remarkably similar despite the significant difference in the number of years of testing data required for the plot: trajectories use 9 years of data whereas RC snapshot streamlines only use 3 years. Therefore, in the absence of policy changes, RC snapshot streamlines can make accurate longitudinal predictions and in the event of a policy change, RC snapshot streamlines before and after can be compared to identify the effects of the intermediate intervention. RC snapshot streamlines could be a powerful tool for education researchers, particularly those who work at State Education Agencies and are required to produce rapid results to influence policy [37].

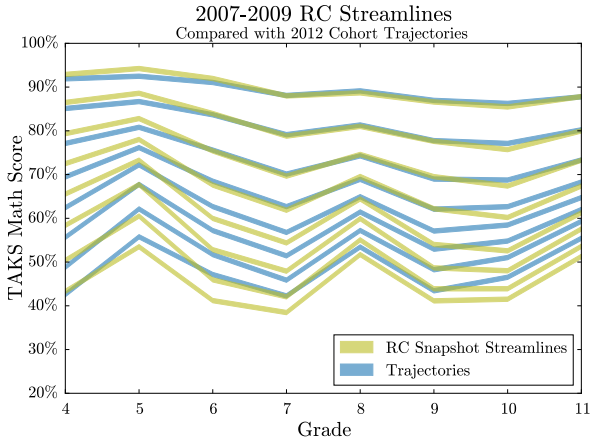

FIG. 9. RC snapshot streamlines for 2007-2009 and trajectories for the 2012 cohort. The double peaks in 5th and 8th grade, due to SSI and AMI, are observed in both the trajectories and the RC snapshot streamlines. These interventions were provided to low-performing K-8th grade students.

VIII. OTHER APPLICATIONS

A. Disaggregation

In addition to grouping students by their observed scores, the students in RC streamlines and trajectories can be grouped according to other variables of interest. For example, students could be disaggregated by demographic variables to study equity and performance disparities. In the Texas Longitudinal Data System, some of the available demographic variables include gender, socio-economic status (SES) with respect to free or reduced lunch status, and race/ethnicity.

Figure 10 shows the trajectories for the cohort of 2012 disaggregated by gender. Male and female students perform similarly, with female students performing slightly better on the TAKS mathematics exams. Figure 11 shows the trajectories for the low-income students who receive free or reduced lunch, disaggregated by race/ethnicity. The trajectories for each score bin are separated into subplots to highlight the within-bin differences.

Students can also be disaggregated by course taking. Figure 12 shows the trajectories for the students in the cohort of 2012, disaggregated by the highest level of physics that they took in high school. In Texas, the students could take advanced placement (AP) physics, basic physics, or integrated physics and chemistry (IPC). For the cohort of 2012, the students were required to take at least basic physics to satisfy the Recommended High School Program, although IPC was sufficient for the Minimum High

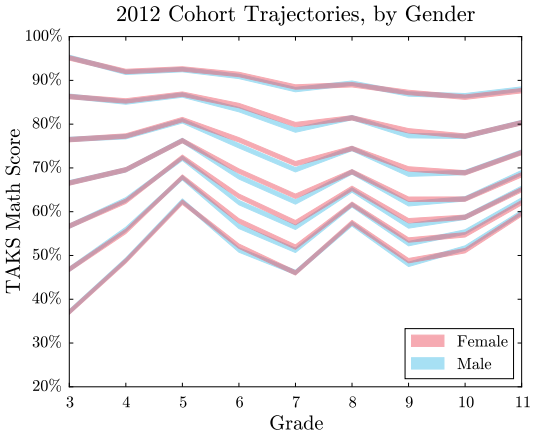

FIG. 10. Trajectories for the cohort of 2012 disaggregated by gender. There are minimal performance disparities associated with differences in gender.

School Program. The high school science requirements in Texas changed in 2014, removing the basic physics requirement [38].

B. Binning by Alternate Exam

One unfortunate consequence of using the RC grouping process for snapshot streamlines is the necessity of a third year of data. However, RC snapshot streamlines can be constructed using only two years of data if an alternate exam is used for binning the students. Students in Texas were assessed in both mathematics and reading annually; therefore, the reading exam is a consistent alternate exam for sorting the students into score bins. Figure 13 shows the RC snapshot streamlines of 2008-2009, showing the flow in mathematics scores with students sorted by their reading scores in 2008. This demonstrates that accurate RC snapshot streamlines can be constructed with only two years of data; the quantitative improvements to the streamlines through the RC process do not require additional data collection time.

C. Future Predictions

To make future predictions of student performance in Texas it is necessary to use data from a new exam whose roll-out started in 2011-2012, the State of Texas Assessments of Academic Readiness (STAAR). Not enough time has passed since STAAR was first implemented to construct a full trajectory plot, although RC snapshot streamlines can be constructed. Figure 14 shows the partial trajectory for

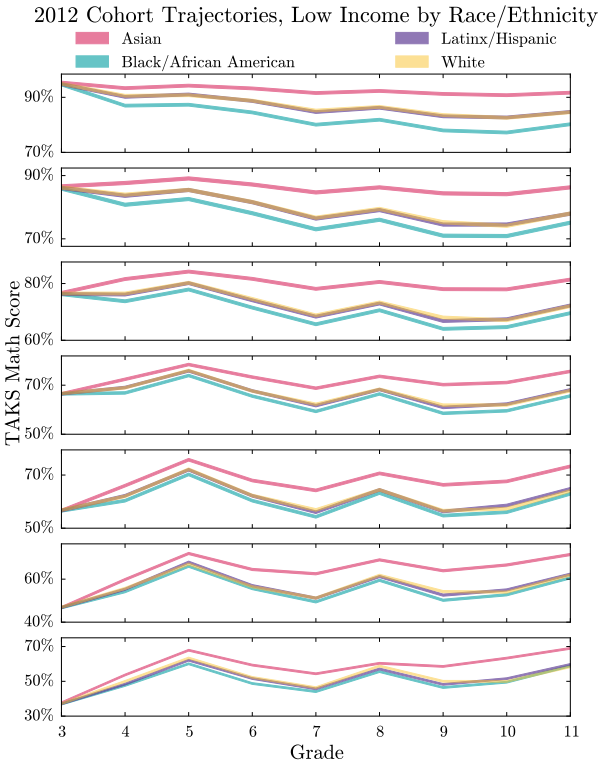

FIG. 11. Trajectories for the low-income students in the cohort of 2012, disaggregated by race/ethnicity. Latinx/Hispanic and White low-income students perform very similarly.

the first cohort of students who took STAAR in 3rd grade (6th graders in 2015) and the RC snapshot streamlines from 2012-2014. The dip in performance in 2015 compared to the predictions may be real. The STAAR mathematics performances in 2015 were so low compared to the State's expectations that the Texas Education Commissioner decided not to use the mathematics exams for any accountability measures [39]. The cause for the drop in performance could be a change in mathematics curriculum standards that was implemented in the 2014-2015 school year [40], affecting the spring 2015 scores, or it could be the delayed result of budget cuts that started in 2011 [41].

IX. CONCLUSIONS

We have developed a new method that can visualize, analyze, and predict longitudinal student testing data. RC streamlines improve on the streamlines created by Bendinelli and Marder [9] by reducing the effects of RTM. Consequently, the RC streamlines can be used to predict long-term student test-

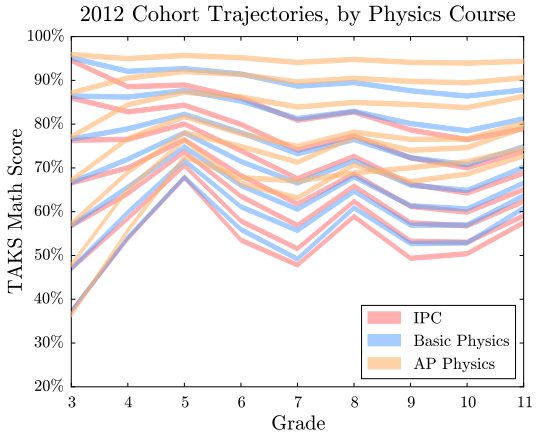

FIG. 12. The trajectories for the cohort of 2012 disaggregated by highest level of physics course taking. Despite having similar 3rd grade scores, AP physics students outperform their basic physics and IPC counterparts.

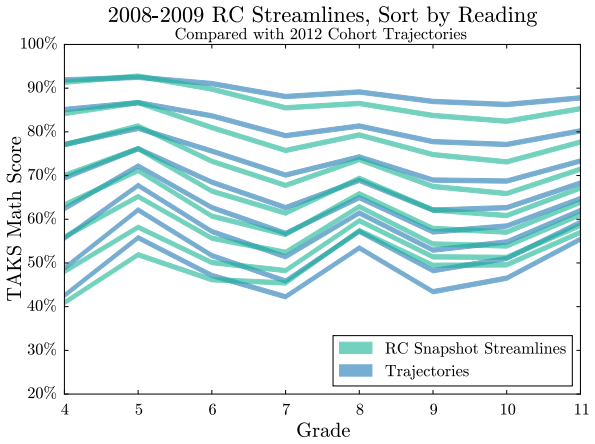

FIG. 13. RC snapshot streamlines for 2008-2009 and trajectories for the 2012 cohort. The students are sorted into score bins by their 2008 reading exams and the changes in score are calculated between their 2008-2009 mathematics exams.

ing performance with only two or three consecutive years of data. RC streamlines can also be used to determine the outcomes of interventions by comparing predictions derived before and after the intervention. This technique has the potential to analyze outcomes rapidly enough for the results to influence policy decisions.

Trajectory plots are a visualization tool to observe long-term trends in the scores for groups of students that are determined by the initial scores. Trajectories are a fundamental longitudinal analysis technique, as they show the average observed scores over time for the students in each initial score bin. By coarse-graining the students into score bins,

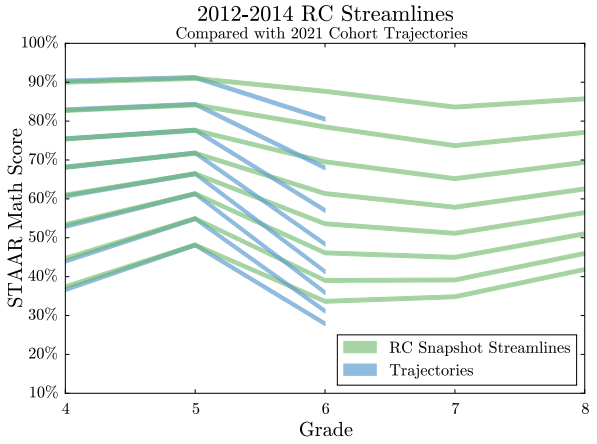

FIG. 14. Partial trajectories for the students who were 6th graders in 2015 and RC snapshot streamlines for 2012-2014. The dip in observed performance in 2015 caused the Texas Education Commissioner to cancel accountability measures related to the mathematics scores [39].

trajectories provide a compromise between the overwhelming information at the individual student level and the loss of information in the full aggregate, which groups completely different types of students together. Trajectories act as a source of comparison for longitudinal prediction methods, such as snapshot streamlines. The longitudinal nature of the trajectory plots causes some limitations. The trajectories are not inclusive of students who join the cohort after the initial grade, and similar to other longitudinal techniques, trajectories suffer from attrition. Full trajectories require almost a decade of data to follow the scores of students throughout primary and secondary school; therefore, they are only possible to assemble after the students have already graduated.

Streamline plotting, as described in Bendinelli and Marder [9], is a tool inspired by techniques in statistical mechanics that models the movement of particles in a fluid. Streamlines in an education context use the changes in score for students in each score bin and for each grade transition to represent the flow of scores through the grades. The streamlines can have a single-cohort design, known as cohort streamlines, or they can utilize an ALD with two years of data, called snapshot streamlines. However, the repeated sorting of students into score bins results in an exaggerated RTM, rendering the streamlines quantitatively inaccurate.

RC streamlines use a separate exam to create the score bins from the exams used to calculate the changes in score. In this way, the RTM is reduced considerably, resulting in acceptably accurate depic-

tions of score flow. RC cohort streamlines are an alternative to trajectories, as both follow a single cohort of students throughout school; however, the RC cohort streamlines include additional students who join the cohort after the initial exam. RC snapshot streamlines use three consecutive years of data to make predictions of the score flow throughout the grades; the first year is used to sort the students into score bins and the second and third years are used to calculate the change in score. RC snapshot streamlines can also be constructed from only two years of data, as long as the sorting exams are not used in the change-in-score calculations. We demonstrated this by using the reading exams to form score bins and the mathematics exams for the changes in score.

Deviations between the RC snapshot streamlines and trajectories are indications of a policy change. Similarly, differences between RC snapshot streamlines from different years identify the effects of intermediate interventions. This was demonstrated by identifying the effects of SSI in the Texas State Longitudinal Data System. SSI was implemented along with the cohort of students who graduated in 2012 and it mandated especially high-stakes mathematics and reading exams in 5th and 8th grade. By comparing the RC snapshot streamlines from 2003-2005 to those from 2007-2009, the effects of the 8th grade SSI policies can be identified.

RC grouping, trajectories, and RC streamlines can be expanded in several ways. Students can be further disaggregated by other variables, such as demographics or course taking, to identify the between-person differences in within-person changes over time. Furthermore, RTM is not only a concern when creating streamline plots. Texas is implementing new school accountability measures beginning partially in the 2018-2019 school year and fully by the 2019-2020 school year that use the relative score classifications of the students from one year to the next as a component of the metric [42]. As demonstrated with the streamlines and trajectories, classifying students based on observed scores leads to RTM. Therefore, RTM can be expected to affect the new school accountability measures and to have potentially profound consequences for Texas schools.

X. ACKNOWLEDGEMENTS

This work was supported by the National Math and Science Initiative and by the National Science Foundation Grants No. 1557410, No. 1557273, No. 1557295, No. 1557294, No. 1557276, No. 1557278, No. 1557286, and No. 1557290. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[1] Editorial Projects in Education Research Center. Issues a-z: No child left behind: An overview. education week. http://www.edweek.org/ew/section/multimedia/no-child-left-behind-overview-definition-summary.html/, April 2015.

[2] Texas Education Agency. Student Success Initiative Manual (2017). http://tea.texas.gov/student.assessment/SSI/.

[3] Cory Koedel, Kata Mihaly, and Jonah Rockoff. Value-added modeling: A review. *Economics of Education Review*, 2015.

[4] Eric Hanushek. The economic value of higher teacher quality. *Economics of Education Review*, 30(3):266–479, 2011.

[5] Steven Glazerman, Ali Protik, Bing ru Teh, Julie Bruch, and Jeffrey Max. *Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment*. United States Department of Education, 2013.

[6] Texas Education Code. Title 2, Subtitle H, Chapter 39: Public School System Accountability.

[7] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, editors. *Longitudinal Data Analysis*. Chapman and Hall/CRC, 2009.

[8] M. Marder and D. Bansal. Flow and diffusion of high-stakes test scores. *Proceedings of the National Academy of Sciences*, 106(41):17267–17270, 2009.

[9] Anthony Bendinelli and Michael Marder. Visualization of Longitudinal Student Data. *Physical Review Special Topics - Physics Education Research*, 8(020119), 2012.

[10] Lesa Hoffman. *Longitudinal Analysis: Modeling Within-Person Fluctuation and Change*. Routledge, 2015.

[11] Stephen Raudenbush and Anthony Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, 2 edition, 2002.

[12] Judith Singer and John Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurence*. Oxford University Press, 2003.

[13] John McArdle and John R. Nesselroade. *Longitudinal Data Analysis Using Structural Equation Models*. American Psychological Association, 2014.

[14] Patrick Curran and Bengt Muthen. The application of latent curve analysis to testing developmental theories in intervention research. *American Journal of Community Psychology*, 27(4), 1999.

[15] Bengt Muthen and Siek-Toon Khoo. Longitudinal studies of achievement growth using latent variable modeling. *Learning and Individual Differences*, 10(2):73–101, 1998.

[16] Daniel Nagin. Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, 4(2):139–157, 1999.

[17] Heather Andruff, Natasha Carraro, Amanda Thompson, Patrick Gaudreau, and Benoit Louvet. Latent Class Growth Modelling: A Tutorial. *Tutorials in Quantitative Methods for Psychology*, 5(1):11–24, 2009.

[18] S. Raudenbush, A. Bryk, and R. Congdon. *HLM 7.01 for Windows [Computer Software]*. Scientific Software International Inc., Skokie, IL, 2013.

[19] L.K. Muthen and B.O. Muthen. *Mplus User's Guide*. Muthen and Muthen, Los Angeles, CA, 8 edition, 2017.

[20] Norman B. Ryder. The cohort as a concept in the study of social change. *American Sociological Review*, 30(6):843–861, December 1965.

[21] Richard Bell. Convergence: An Accelerated Longitudinal Approach. *Child Development*, 24(2):145–152, June 1953.

[22] Yasuo Miyazaki and Stephen Raudenbush. Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods*, 5(1):44–63, 2000.

[23] The University of Texas at Austin. Texas education research center. https://research.utexas.edu/erc/.

[24] Student Assessment Division. Technical Digest 2009-2010, 2010.

[25] Texas Education Agency. State of Texas Assessment of Academic Readiness. http://tea.texas.gov/student.assessment/staar/.

[26] Texas Education Agency. House Bill 3 Transition Plan. http://tea.texas.gov/student.assessment/hb3plan/.

[27] Student Assessment Division. Technical Digest 2014-2015, 2015.

[28] Texas Education Agency. Taks raw score conversion tables. https://tea.texas.gov/student.assessment/taks/convtables/.

[29] Matthew Guthrie. *Grouping and Comparing Texas High Schools Through Machine Learning and Visualization Techniques*. PhD thesis, University of Texas at Austin, 2018.

[30] John R. Nesselroade, Stephen M. Stigler, and Paul B. Baltes. Regression toward the mean and the study of change. *Psychological Bulletin*, 88(3):622–637, 1980.

[31] Melvin R. Novick. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3:1–18, 1966.

[32] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publication Corporation, 1968.

[33] Simon Tidd and Sonia Dominguez. Chronic absenteeism and student mobility. http://e3alliance.org/wp-content/uploads/2017/04/E3-3D-Mobility-Chronic-Absence-040417.pdf, 2017.

[34] Texas Education Agency. The Student Success Initiative: An Evaluation Report (2009). https://tea.texas.gov/WorkArea/DownloadAsset.aspx?id=2147490898.

[35] Texas Education Agency. The Student Success Initiative: 2009-2010 Biennium Evaluation Report, 2011.

[36] Texas 85th Legislative Session. SB1 General Appropriations Bill. https://capitol.texas.gov/tlodocs/85R/billtext/pdf/SB00001F.pdf, 2017.

[37] Carrie Conaway, Venessa Keesler, and Nathaniel Schwartz. What Research Do State Education Agencies Really Need? The Promise and Limitations of State Longitudinal Data Systems. *Educational Evaluation and Policy Analysis*, 37(1S):16S–28S, May 2015.

[38] Texas Education Agency. State Graduation Requirements. http://tea.texas.gov/graduation.aspx.

[39] Jeffrey Weiss. Three things you should know about the 2015 STAAR results. *The Dallas Morning News*, May 2015.

[40] Texas Education Agency. Mathematics Texas Essential Knowledge and Skills. https://tea.texas.gov/index2.aspx?id=2147499971.

[41] M. Marder and Chandra K. Villanueva. Consequences of the texas public school funding hole of 2011-2016. https://forabettertexas.org/images/EO_2017_09_SchoolFinance_ExecSum.pdf, October 2017.

[42] Texas Education Agency. *2015-16 A-F Ratings*, December 2016.