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,
2515 Speedway, C1600 Austin, TX 78712-1192, USA*

(Received 27 August 2018; published 3 July 2019)

[This paper is part of the Focused Collection on Quantitative Methods in PER: A Critical Examination.] The education of physicists requires annual progress in mathematics through grades K–12, continuing on to high school physics, undergraduate instruction, and often many years of postgraduate study. The traditional domain of physics education research is the study of pedagogical techniques for a single lesson or a single course, but the physics and physics education research communities will need to study the complete educational system to increase performance and participation, especially for underrepresented groups. Analysis of long-term student outcomes and policy impacts requires predictive longitudinal techniques. We begin with concepts from fluid and statistical mechanics—trajectories and streamlines—to visualize longitudinal outcomes and make predictions. Drawing on coarse-graining procedures to model the movement of particles in a fluid, the longitudinal data are sorted into score bins to depict 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 explain the idea of RC streamlines through a toy model of freshman physics performance and then apply them to the realistic setting of 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 2 or 3 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’s Student Success Initiative.

DOI: [10.1103/PhysRevPhysEducRes.15.020109](https://doi.org/10.1103/PhysRevPhysEducRes.15.020109)

I. INTRODUCTION

The education and preparation of future physicists is an important topic within the physics and physics education research communities. Policies affecting K–12 education also have far-reaching implications within higher education. Recent policy changes in Texas, for example, eliminated Physics and Algebra 2 as required courses for high school graduation [1]. Associated changes in testing soon led to a significant drop in college readiness rates [2]; long-term implications for the interest and preparedness of undergraduate physics students may follow. 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 many years to reach fruition.

A common way to analyze longitudinal datasets is to use hierarchical linear modeling or structural equation modeling to compute outcomes with respect to observed variables or unobserved latent variables [3]. Marder and Bansal [4] and Bendinelli and Marder [5] developed an alternative approach that makes use of standard ideas from statistical mechanics. Trajectories can represent the position of any 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 subunits. It can be extended to fluid systems by using the Lagrangian description of fluid mechanics to coarse-grain the fluid at some point in

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

time and follow the motion of packets of fluid as the flow evolves [6]. 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 represented in the Eulerian description 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. The reason to perform this binning procedure for students is identical to the reason one performs it for fluids. Averaging together too many fluid parcels throws away almost all the information about how the fluid will behave over time. Similarly, students at different score levels not only tend to follow similar paths in time but also generally respond similarly to policy interventions. Thus the binning provides a much more detailed view of how students respond to changes in education than is possible with the full aggregate. When dealing with student scores, we will use the terms “coarse graining” and “binning” interchangeably.

One cost of obtaining this binned student score information, however, is a set of new technical problems. A consequence of the partially stochastic nature of observed scores and the distribution of observed scores is the regression (convergence) of the scores toward the mean score of the distribution. 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 2 or 3 years of data. In this way, RC streamlines can predict longitudinal outcomes for 5–10 years within a time suitable for informing policy; in addition, the comparison of

streamlines from different periods can identify the effects of intermediate interventions.

The benefits of RC streamlines are best demonstrated when analyzing large numbers of students over a long period of time; RC streamlines are well suited for K–12 studies. Predictive techniques are not necessary when studying outcomes in the short term (trajectories would be better suited for such studies) and grouping students by score is particularly useful when dealing with large numbers of students. The majority of the paper will discuss RTM and RC grouping within the context of K–12 mathematics scores, as it is the most relevant longitudinal K–12 information to physics education in our dataset. Nonetheless, the discussion of RTM and RC grouping also can apply to short-term studies; therefore, we first apply our methods to a toy model of freshman physics performance to briefly illustrate the effects of RTM and the adjustments from RC grouping in Sec. II.

Section III discusses the setting used for the rest of the paper: the Texas State Longitudinal System and standardized testing in Texas. Section IV summarizes some techniques in education research that are currently used to analyze student performance over time. Section V lays the foundation for the RC streamlines by summarizing the trajectory and streamline techniques described in Bendinelli and Marder [5]. Section VI presents a simple theory for understanding the consequences of RTM, specifically in regards to trajectories and streamlines. Section VII describes the RC grouping technique and uses the same theory to demonstrate how RTM is reduced. Section VIII applies RC streamlines to identify the effects of a state-wide intervention called the Student Success Initiative, while Sec. IX explores additional applications including coarse-graining by alternate exams and future predictions of student performance. In Sec. X, we summarize our conclusions. Supplemental Material [7] containing pseudocode has been provided to assist others in constructing RC streamlines.

Much of the work in this paper is drawn from the Ph.D. dissertation of Stephens [8].

II. TOY MODEL OF FRESHMAN PHYSICS

To illustrate the problem posed by regression to the mean, and our solution to it, we begin with a toy model of freshman physics outcomes. Suppose we know student results from first and second semester undergraduate physics. Students are organized into groups based on their letter grades (90%–100%, 80%–90%, 70%–80%, 60%–70%, and < 60%) for the first of the courses; in this way, one could study the change in physics scores for A students compared to C students. One way to characterize performance in the first and second courses is through the scores on the final exams. Let us suppose that there is an additional measurement of student performance, such as the midterm exam scores in the first course. In a first version of analysis,

FIG. 1. Average changes in score for A, B, C, D, and F students from the first to the second semester of physics. Student letter grade categories are determined by the first semester physics final exam scores in the solid lines and by the first semester midterm exam scores in the dashed lines.

final exam scores from the first course are used to group students, and the change in final exam score after the second course is found for each letter grade group. Using the RC method, we would instead group students through their midterm scores, and then find the changes in final exam scores.

Let the midterm score for the first course, first final exam score, and second final exam score be S_m , S_1 , and S_2 respectively. We use the notation $\mathcal{N}(\mu, \sigma^2)$ to denote a normal distribution with mean μ and standard deviation σ . Using a sample size of 500 students, we set $S_m = \mathcal{N}(80, 10^2)$, truncating the scores at 0 and 100. The physics scores were generated in R using the `rnorm` and `rtorm` functions and the following expressions: $S_1 = S_m + \epsilon_1$ and $S_2 = (C)S_m + (1 - C)S_1 + \epsilon_2$ where $\epsilon_1 = \epsilon_2 = \mathcal{N}(0, 10^2)$ are error scores as in classical test theory [9] and $C = 0.5$ is a coefficient that determines the relative weight of the previous performances on the second semester physics score. The standard deviation $\sigma = 10$ in this toy example is somewhat smaller than the standard deviation for the actual exams we will study later. Both final exam scores were also truncated at 0 and 100, so that if a score was outside of these bounds, it was reset to equal the bound. The sample size, normal distributions, and weighting coefficient are all chosen to create fairly realistic score distributions but of course are arbitrary and could easily be adjusted. In this simple model we are assuming that the two courses are not affecting students' relative performance, so if there were no noise the student means for courses one and two would be the same. This model demonstrates the effects of RTM without obscuring the results with other effects.

Figure 1 shows the change in average physics final exam scores for A, B, C, D, and F students from the first to the

second semester. To represent the uncorrected version with RTM, students' letter grades are determined by their first semester final exams (solid lines). The RC version instead groups the students by their letter grade based on the midterm exams (dashed lines). Since the students are categorized by a single measurement, students who performed exceptionally well or poorly on that measurement are "miscategorized." Subsequent measurements will likely produce scores closer to their average performance. For example, "A" students on the final of Physics 1 are composed of the "real" A students ($S_m > 90$) plus the B students who got lucky, minus some "real" A students who got unlucky. As there are almost always more B students than A students in a typical score distribution, the addition of lucky students outweighs the loss of unlucky students. In Physics 2, the students who got the Physics 1 A's only through luck are likely to end up with B's. This is the source of regression to the mean. This regression is evident in the solid lines of Fig. 1. In the regression corrected (RC) version, RTM still occurs between the midterm scores and the other measurements, but the RTM between the final exam scores is much less, and the dashed lines appear straight. The essence of the RC grouping method is to form groups from measurements correlated with but independent of the ones whose changes one wishes to observe. The rest of the paper will address the analysis methods, RTM, and RC grouping in depth, using observed K–12 mathematics data.

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 [10]. The Texas Education Agency (TEA) provides the ERC with much of their student-level data, including demographics and standardized testing information. While the number of students in the analysis varies with the year and grade, each of the plots in this paper contain data for over 250 000 students. 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 grades, as well as periodic exams in other subjects [11]. 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 end-of-grade high school exams with end-of-course subject exams [12,13]. 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 [14]. Each year the exams were designed to be very similar, resulting in a consistent raw score metric. Second, the raw score conversion tables [15] were examined and the raw scores corresponding with a passing score within each grade for the mathematics TAKS exams were 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 [14], 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 of 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 [16]; this resulted in a large shift in scaled score ranges, without the ability to track longitudinal progress before 2008. Therefore, for simplicity, transparency, and consistency, we have chosen to use raw score percentages as our metric.

IV. 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 [17]. 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.

To minimize or isolate the potential sources of time-related change, studies may focus on many cohorts of students within a cross section of time or on a single cohort

of students over time. 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. Cross-sectional models are cost effective and efficient, but they provide no insight into the within-person changes over time. By contrast, single-cohort studies use the longitudinal data from a single graduating cohort of students to study within-person change [18]. 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, and the results may not extend to other cohorts. Accelerated longitudinal design (ALD) studies, also known as cross-sequential studies, are a compromise between cross-sectional studies and longitudinal single-cohort studies [19]. ALDs use data from multiple overlapping cohorts beginning at different grades to span a large grade range while using only a few years of data. For example, Ding and Zhang used the Colorado Learning Attitudes about Science Survey (CLASS) for six adjacent cohorts over seven grade levels to study the progression of epistemological beliefs of future physics teachers from the beginning of freshman year through the second year of in-service teaching [20]. ALDs reduce 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.

Most longitudinal studies utilize a statistical technique known as hierarchical linear modeling [21,22], or a similar technique called structural equation modeling [23–25], 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 nonparametric form.

This idea can be extended to grouped individuals through techniques such as group-based trajectory modeling [26] and latent class growth modeling [27]. 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 pretreatment 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 [28] and MPLUS [29], 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.

V. FOUNDATION

The methods detailed in Bendinelli and Marder [5] 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 [30]. 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.

FIG. 2. 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 271 500 students included in this analysis.

However, the process of grouping creates a set of technical problems that we have set ourselves in this paper to resolve.

Figure 2 shows the trajectories for the cohort of students who graduated high school in 2012. The thickness of each line in this figure is proportional to the number of students in that score bin. The average scores in each grade are connected by linear segments. The performance peaks in 5th and 8th grade will be discussed in Sec. VIII; otherwise, the trajectories are fairly flat. This is not surprising when averaging thousands of students, as prior performance is the strongest indicator of future performance.

In addition to grouping students by their observed scores, the students in trajectories and streamlines 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, socioeconomic status with respect to free or reduced lunch status, and race or ethnicity.

Students can also be disaggregated by course taking. Figure 3 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 School Program.

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

The high school science requirements in Texas changed in 2014, removing the basic physics requirement [31]. One interesting feature of this figure is that while trajectories are typically flat, supporting the idea that previous scores are a very strong indicator of future performance, initially poor performing AP physics takers in particular show considerable upward movement in their trajectories, especially after the 5th grade intervention. We believe this warrants further study.

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.

FIG. 4. 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 average 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.

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 Fig. 4, 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 regrouped 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). In this figure, and most of those that follow, line thicknesses are constant, rather than representing the numbers of students as in Fig. 2.

Snapshot streamlines, as in Fig. 5, use an ALD with data from eight sequential cohorts, following each for 2 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 year, the students from each cohort are sorted into score bins in their grade and the score change is calculated

FIG. 5. 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.

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 2 years of data. Assuming that different cohorts perform similarly, snapshot streamlines can be used to predict longitudinal results from a single cohort.

Figures 6 and 7 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 point 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.

VI. REGRESSION TO THE MEAN

Regression to the mean 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

FIG. 6. 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 disparities, 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.

observed scores on a single exam, the random component of their score may force the student into a different bin than their achievement might warrant. 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

FIG. 7. 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.

FIG. 8. 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 Sec. VIII. The relationship between the scores for most exams is roughly linear.

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, as in the work of Nesselrode, Stigler, and Baltes on RTM [32], we can better understand the significance of RTM in the coarse-graining procedures. We invoke classical test theory [9,33] 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 8

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 partial floor effects due to a program called Student Success Initiative (see Sec. VIII). 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

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

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

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

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

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

The sequence of anticipated theoretical scores is shown in Table I and depends on the correlation coefficients between 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 $a, \rho a, \rho^2 a, \rho^3 a$, 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 the exams tend to be less correlated with time (see Table III), 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:

TABLE I. Anticipated theoretical 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.

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	$\varsigma_7 = \rho_{6,7}\rho_{5,6}\rho_{4,5}\rho_{3,4}a$	$\rho^4 a$

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

$$E(z_3|z_2 = b) = \rho_{z_2,z_3}b,$$

$$E(z_4|z_3 = c) = \rho_{z_3,z_4}c,$$

...

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

$$\varsigma_1 = a,$$

$$\varsigma_2 = \rho_{z_1,z_2}a,$$

$$\varsigma_3 = \rho_{z_2,z_3}\varsigma_2 = \rho_{z_2,z_3}\rho_{z_1,z_2}a,$$

$$\varsigma_4 = \rho_{z_3,z_4}\varsigma_3 = \rho_{z_3,z_4}\rho_{z_2,z_3}\rho_{z_1,z_2}a,$$

...

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, Fig. 7 exhibits RTM mainly in the first segment of the trajectories but throughout the streamlines, aligning with the observations from the theory.

VII. REGRESSION-CORRECTED GROUPING

The sequence of anticipated theoretical 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 transition. Streamlines are essentially several short trajectories strung together; each segment is the first 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 transition. Therefore, if the streamlines instead utilized the second 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 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 Sec. IX and it demonstrates that RC streamlines can be constructed with only 2 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

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

$$E(z_3|z_2 = b) = \rho_{z_2,z_3}b, \quad E(z_4|z_2 = b) = \rho_{z_2,z_4}b,$$

$$E(z_4|z_3 = c) = \rho_{z_3,z_4}c, \quad E(z_5|z_3 = c) = \rho_{z_3,z_5}c,$$

...

The differences of the two scores in each pair are used as the slopes for each of the vectors to construct the RC streamlines within the theoretical framework. For a score bin with average score a on exam 1, the slope of the vector

between exams 2 and 3 would be $(\rho_{z_1, z_3} - \rho_{z_1, z_2})a$. 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 $b = (\rho_{z_1, z_3}/\rho_{z_2, z_3})a$, which then allows us to calculate $E(z_4|z_2 = b)$. Using this method iteratively gives a sequence of theoretical scores in the RC streamlines of

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

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 theoretical 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 theoretical 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

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

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 III. Comparison between the theoretical 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.

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

streamlines are nearly identical, whereas the uncorrected streamlines exhibit extreme RTM.

While this simple theoretical framework is helpful for understanding the role of RTM in the trajectories, streamlines, and RC streamlines, it does not depict the scores that are used in the plots. The trajectories only use observed average scores but both types of streamlines use interpolation functions to determine the slopes of each segment from the observed average scores. Let $f_{g,b}(s)$ equal the interpolated increase in score in grade g from a score s in grade $g-1$, based on the observed change in average scores from grade $g-1$ to grade g when binned by the scores from grade b and let $\bar{s}_{g,b}$ be the observed average score in grade g when binned by the scores from grade b . The actual sequence of values in the trajectory, streamline,

TABLE IV. Comparison between the observed/computed score sequences within the trajectory, cohort streamline, and RC cohort streamline frameworks. These are the values that are used when creating the plots.

Trajectories	Streamlines	Regression-corrected streamlines
$s_4 = \bar{s}_{4,3}$	$s_4 = \bar{s}_{4,3}$	$s_4 = \bar{s}_{4,3}$
$s_5 = \bar{s}_{5,3}$	$s_5 = s_4 + f_{5,4}(s_4)$	$s_5 = \bar{s}_{5,3}$
$s_6 = \bar{s}_{6,3}$	$s_6 = s_5 + f_{6,5}(s_5)$	$s_6 = s_5 + f_{6,4}(s_5)$
$s_7 = \bar{s}_{7,3}$	$s_7 = s_6 + f_{7,6}(s_6)$	$s_7 = s_6 + f_{7,5}(s_6)$
$s_8 = \bar{s}_{8,3}$	$s_8 = s_7 + f_{8,7}(s_7)$	$s_8 = s_7 + f_{8,6}(s_7)$
$s_9 = \bar{s}_{9,3}$	$s_9 = s_8 + f_{9,8}(s_8)$	$s_9 = s_8 + f_{9,7}(s_8)$
$s_{10} = \bar{s}_{10,3}$	$s_{10} = s_9 + f_{10,9}(s_9)$	$s_{10} = s_9 + f_{10,8}(s_9)$
$s_{11} = \bar{s}_{11,3}$	$s_{11} = s_{10} + f_{11,10}(s_{10})$	$s_{11} = s_{10} + f_{11,9}(s_{10})$

and RC streamline plots from grade 4 to grade 11 can be seen in Table IV. The segment from grade 4 to grade 5 for the trajectories and RC streamlines would be identical if not for the slight differences in the students that are included in that segment. The interpolated changes in score used in the RC sequence produce much more reasonable estimates as they do not incorporate as much RTM.

Figure 9 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 generally require students to have a 3rd grade score (and a 4th grade score in this case as the plot begins in 4th grade) whereas RC streamlines require any three consecutive scores

FIG. 9. 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. 10. 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.

(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. The number of students in the trajectories drops from about 271 500 in 3rd grade to 218 700 in 9th grade, whereas the number of students in the streamlines start at 250 500 and drop to 236 400. The difference in population corresponds with lower performance in the RC streamlines after 5th grade, as mobility has been shown to be correlated with lower performance [34]. 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 3 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 10 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. We discuss this next.

VIII. APPLICATION: SSI

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

FIG. 11. RC snapshot streamlines for 2007–2009 and trajectories for the 2012 cohort. The double peaks in 5th and 8th grades, 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.

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 grades; 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 [35]. 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) [36].

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 [37]. 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 Fig. 10. 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 11 shows the RC snapshot streamlines from 2007 to 2009, after the SSI impacted students up through 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 [37] \$5.5 million in the 2018–2019 biennium budget [38].

Most of the comparisons between two sets of trajectories and/or streamlines have been by visual inspection, or by making the thickness of the lines equal to the standard error. When trying to quantitatively compare two sets of curves, one value may not be sufficient, as the difference between the curves may change with grade or bin. Ultimately, a matrix of differences could be used to quantify the differences for each grade and bin, and averages along the grade dimension and/or bin dimension could be used. The overall average of the absolute value of the difference between the trajectory and the streamline for each bin and grade combination in Figs. 9, 10, and 11 are 2.108, 3.077, and 2.150, respectively. The 2007–2009 RC streamlines predict the 2012 cohort trajectories better than the 2003–2005 streamlines.

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 [39].

IX. OTHER APPLICATIONS

A. Binning by alternate exam

One challenge 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 2 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 that could be used to sort the students into score bins. Figure 12 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 2 years of data; the quantitative improvements to the streamlines through the RC process do not require additional data collection time.

B. 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. Not enough time has passed since STAAR was first implemented to construct a full trajectory plot, although RC snapshot streamlines can be constructed.

FIG. 12. 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 scores are calculated between their 2008–2009 mathematics exams.

Figure 13 shows the partial trajectory for the first cohort of students who took STAAR in 3rd grade (6th graders in 2015) and the RC snapshot streamlines from 2012 to 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 [40]. The cause for the drop in performance could be a change in mathematics curriculum standards that was implemented in the 2014–2015 school year [41], affecting the spring 2015 scores, or it could be the delayed result of budget cuts that started in 2011 [42].

FIG. 13. Partial trajectories for the cohort of 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 [40].

X. 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 [5] by reducing the effects of RTM. Consequently, the RC streamlines can be used to predict long-term student testing performance with only 2 or 3 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.

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, 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 has 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 [5], is a tool inspired by techniques in statistical mechanics that model 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. Despite slight changes in the population due to the additional measurement, the RTM is reduced considerably, resulting in acceptably accurate depictions 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 3 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 2 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. RC grouping, trajectories, and RC streamlines can also be expanded by further disaggregating students by other variables, such as demographics or course taking, to identify the between-person differences in within-person changes over time.

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 grades. By comparing the RC snapshot streamlines from 2003 to 2005 to those from 2007 to 2009, the effects of the 8th grade SSI policies can be identified.

For short-term longitudinal studies with only three sets of scores, trajectories and streamlines are identical; therefore, some of the novelty of RC streamlines is lost on short-term studies. The effects of RTM should not be ignored, however, as demonstrated in our toy model for freshman physics performance, and can be overcome by the techniques we have described. Grouping students into score bins is a useful tool when the number of students is large, such that individual scores or overall average scores

are cumbersome or oversimplified. For smaller sample sizes, score bins may be unnecessary as patterns in the individual trajectories could be discernible. While the discussion of RTM and RC grouping was fairly technical in this paper, we believe that the outcomes are simple and intuitive, especially for researchers with a physics background.

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 [43]. 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.

ACKNOWLEDGMENTS

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] Texas Education Agency, House Bill 5, Foundation high school program, <https://tea.texas.gov/graduation-requirements/hb5.aspx>.
 - [2] M. Marder, College readiness drops back ten years in Texas, 2017, <https://www.tribtalk.org/2017/06/05/college-readiness-drops-back-ten-years-in-texas/>.
 - [3] *Longitudinal Data Analysis*, edited by G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Chapman and Hall/CRC, London, 2009).
 - [4] M. Marder and D. Bansal, Flow and diffusion of high-stakes test scores, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 17267 (2009).
 - [5] A. Bendinelli and M. Marder, Visualization of longitudinal student data, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020119 (2012).
 - [6] P. Kundu, I. Cohen, and D. Dowling, *Fluid Mechanics*, 6th ed. (Academic Press, New York, 2015).
 - [7] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.020109> for pseudo-code describing algorithms for the main methods in this paper.
 - [8] S. Stephens, Longitudinal predictions using alternative binning to reduce regression to the mean, Ph.D. thesis, University of Texas at Austin, 2017.
 - [9] M. R. Novick, The axioms and principal results of classical test theory, *J. Math. Psychol.* **3**, 1 (1966).
 - [10] The University of Texas at Austin, Texas Education Research Center, <https://research.utexas.edu/erc/>.
 - [11] Texas Education Agency, Student Assessment Division, Technical Digest 2009–2010 (2010), <https://tea.texas.gov/student-assessment/techdigest/yr0910/>.
 - [12] Texas Education Agency, State of Texas assessment of academic readiness, <http://tea.texas.gov/student-assessment/staar/>.
 - [13] Texas Education Agency, House Bill 3, Transition plan, <http://tea.texas.gov/student-assessment/hb3plan/>.
 - [14] Texas Education Agency, Student Assessment Division. Technical Digest 2014–2015 (2015), https://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2014-2015/.
 - [15] Texas Education Agency, TAKS raw score conversion tables, <https://tea.texas.gov/student-assessment/taks/convtables/>.

- [16] Texas Education Code, Title 2, Subtitle H, Chapter 39: Public school system accountability, <https://statutes.capitol.texas.gov/Docs/ED/htm/ED.39.htm>.
- [17] L. Hoffman, *Longitudinal Analysis: Modeling Within-Person Fluctuation and Change* (Routledge, New York and London, 2015).
- [18] N. B. Ryder, The cohort as a concept in the study of social change, *Am. Sociological Rev.* **30**, 843 (1965).
- [19] R. Bell, Convergence: An accelerated longitudinal approach, *Child Development* **24**, 145 (1953).
- [20] L. Ding and P. Zhang, Making of epistemologically sophisticated physics teachers: A cross-sequential study of epistemological progression from preservice to in-service teachers, *Phys. Rev. Phys. Educ. Res.* **12**, 020137 (2016).
- [21] S. Raudenbush and A. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. (Sage Publications, Thousand Oaks, California, 2002).
- [22] J. Singer and J. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (Oxford University Press, New York, 2003).
- [23] J. McArdle and J. R. Nesselroade, *Longitudinal Data Analysis Using Structural Equation Models* (American Psychological Association, Washington, DC, 2014).
- [24] P. Curran and B. Muthen, The application of latent curve analysis to testing developmental theories in intervention research, *Am. J. Community Psychol.* **27**, 4 (1999).
- [25] B. Muthen and S.-T. Khoo, Longitudinal studies of achievement growth using latent variable modeling, *Learning and individual differences* **10**, 73 (1998).
- [26] D. Nagin, Analyzing developmental trajectories: A semi-parametric, group-based approach, *Psychol. Methods* **4**, 139 (1999).
- [27] H. Andruff, N. Carraro, A. Thompson, P. Gaudreau, and B. Louvet, Latent class growth modelling: A tutorial, *Tutorials in Quantitative Methods for Psychology* **5**, 11 (2009).
- [28] S. Raudenbush, A. Bryk, and R. Congdon, *HLM 7.01 for Windows [Computer Software]* (Scientific Software International Inc., Skokie, IL, 2013).
- [29] L. K. Muthen and B. O. Muthen, *MPLUS User's Guide*, 8th ed. (Muthen and Muthen, Los Angeles, 2017).
- [30] M. Guthrie, Grouping and comparing Texas high schools through machine learning and visualization techniques, Ph.D. thesis, University of Texas at Austin, 2018.
- [31] Texas Education Agency, State graduation requirements, <http://tea.texas.gov/graduation.aspx>.
- [32] J. R. Nesselroade, S. M. Stigler, and P. B. Baltes, Regression toward the mean and the study of change, *Psychol. Bulletin* **88**, 622 (1980).
- [33] F. M. Lord and M. R. Novick, *Statistical Theories of Mental Test Scores* (Addison-Wesley, Reading, MA, 1968).
- [34] S. Tidd and S. Dominguez, Chronic absenteeism and student mobility, 2017, <http://e3alliance.org/wp-content/uploads/2017/04/E3-3D-Mobility-Chronic-Absence-040417.pdf>.
- [35] Texas Education Agency, Student Success Initiative Manual, 2017, <http://tea.texas.gov/student.assessment/SSI/>.
- [36] Texas Education Agency, The student success initiative: An evaluation report, 2009, <https://tea.texas.gov/WorkArea/DownloadAsset.aspx?id=2147490898>.
- [37] Texas Education Agency, The student success initiative: 2009–2010 biennium evaluation report, 2011, <https://tea.texas.gov/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147495699&libID=2147495696>.
- [38] Texas 85th Legislative Session, SB1 general appropriations bill, <https://capitol.texas.gov/tlodocs/85R/billtext/pdf/SB00001F.pdf> (2017).
- [39] C. Conaway, V. Keesler, and N. Schwartz, What research do state education agencies really need? The promise and limitations of state longitudinal data systems, *Educ. Eval. Policy Anal.* **37**, 16S (2015).
- [40] J. Weiss, Most Texas schools meet eased rating criteria, but racial gaps remain, The Dallas Morning News, 2015, <https://www.dallasnews.com/news/education/2015/08/07/most-texas-schools-meet-eased-rating-criteria-but-racial-gap-remains>.
- [41] Texas Education Agency, Mathematics Texas essential knowledge and skills, <https://tea.texas.gov/index2.aspx?id=2147499971>.
- [42] M. Marder and C. K. Villanueva, Consequences of the Texas public school funding hole of 2011–2016, https://forabettertexas.org/images/EO_2017_09_SchoolFinance_ExecSum.pdf (2017).
- [43] Texas Education Agency, A-F resources (2019), <https://tea.texas.gov/A-F/>.