Flow and diffusion of high-stakes test scores

M. Marder and D. Bansal

Center for Nonlinear Dynamics and Department of Physics, University of Texas, Austin, TX 78712

Edited by Leo P. Kadanoff, University of Chicago, Chicago, IL, and approved August 18, 2009 (received for review December 2, 2008)

We apply visualization and modeling methods for convective and diffusive flows to public school mathematics test scores from Texas. We obtain plots that show the most likely future and past scores of students, the effects of random processes such as guessing, and the rate at which students appear in and disappear from schools. We show that student outcomes depend strongly upon economic class, and identify the grade levels where flows of different groups diverge most strongly. Changing the effectiveness of instruction in one grade naturally leads to strongly nonlinear effects on student outcomes in subsequent grades.

Fokker–Planck equation | convolution | education

Texas began testing almost every student in almost every public school in grades 3-11 in 2003 with the Texas Assessment of Knowledge and Skills (TAKS). Every other state in the United States administers similar tests and gathers similar data, either because of its own testing history, or because of the Elementary and Secondary Education Act of 2001 (No Child Left Behind, or NCLB). Texas mathematics scores for the years 2003 through 2007 comprise a data set involving more than 17 million examinations of over 4.6 million distinct students. Here we borrow techniques from statistical mechanics (1) developed to describe particle flows with convection and diffusion and apply them to these mathematical scores. The methods we use to display data are motivated by the desire to let the numbers speak for themselves with minimal filtering by expectations or theories.

The most similar previous work describes schools using Markov models. “Demographic accounting” (2) predicts changes in the distribution of a population over time using Markov models and has been used to try to predict student enrollment year to year (3, 4), likely graduation times for students (5), and the production of and demand for teachers (6). We obtain a more detailed description of students based on large quantities of testing data that are missing or had zero score the previous year minus those who were missing or had zero score the previous year. This flow, which represents a Fokker–Planck equation (1) that makes this idea precise.

The diffusive contribution to the flow due to guessing can be modeled mathematically: students invariably guess the answers to questions they do not know, because there is no penalty for guessing, and the fraction of correct guesses can be modeled by a binomial distribution. Each question has four responses, so the probability of guessing correctly is $f = \frac{1}{4}$, and students who know $n$ questions out of $N$ total questions on an exam will guess at the remaining $M = N - n$, resulting in a mean (normalized) score of $(n + fM)/N$ with a variance of $f(1 - f)M/N^2$. This variance provides a lower limit for the amount of diffusion in the absence of any other diffusive terms. The actual diffusion, as measured from the data, is several times greater than this lower limit in all cases, indicating that randomness due to guessing only provides a small part of the diffusion.

One consequence of diffusion is that not all students follow the path predicted by their flow arrows. Fig. 2A is a graphical representation of the amount students’ scores are raised above or below the main flow by diffusion. These differences are due both to guessing and to differences in education, school, etc. A more subtle consequence of diffusion is that following students from the future into the past is different from following them from the past into the future. One can obtain a different set of flow arrows by choosing students whose score fell into a range such as 80%-89% in 2007 and computing their average score the prior year. This flow, which is displayed as Fig. 1B answers the question “If I know students’ scores when they are in 11th grade, what is the most likely set of scores for them to have had coming from third?”

The upper and lower plots in Fig. 1 divide students into two groups according to their level of economic need. The upper plots show students not eligible for free and reduced meals (called “not low income”), and the bottom plots show those who are eligible (called “low income”). Many other groupings are possible, including those by race and gender, economic need according to school rather than by individual, or combining race and gender with economic need.

There are two additional phenomena that contribute to student flows. Students appear and disappear from one year to the next, and students can be required to repeat a grade. These contributions are reflected to some extent in the sizes of arrows in Fig. 1, but it is useful to bring them out directly. Vertical arrows in Fig. 2B show the number of students who appeared in each grade and were missing or had zero score the previous year minus those who

Author contributions: M.M. designed research; M.M. and D.B. performed research; M.M. and D.B. analyzed data; and M.M. and D.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

1 To whom correspondence should be addressed. E-mail: marder@mail.utexas.edu.
had been present and now vanish or get zero score. The horizontal arrows show the numbers of students who repeat a grade.

Fig. 3 shows how flow fields evolve over time for low-income students. The broad outlines of the flow pattern remain remarkably constant, while at the same time there are some systematic changes such as a rapid increase in the numbers of students obtaining commended scores at 10th grade.

Discussion

A characteristic pattern in Fig. 1 is a strong horizontal flow with arrows of decreasing size above and below it pointing towards the flow center. In fluids, this phenomenon results from the competition between fluctuations and dissipation: a particle that is moving much faster than those around it because it has just received a particularly large random kick is most probably going to slow down. In statistics this phenomenon is called regression to the mean (7) and explains why arrows above the center of the flow tend to point down. Regression to the mean can be caused by several factors, including the mathematics of guessing, or by the small likelihood of students having exceptional teachers several years in a row.

Educational outcomes for students from wealthy and poor families are very different in Texas. The flow fields show where the greatest divergences between these groups occur. The flow patterns in the top and bottom rows of Fig. 1 start out in nearly the same direction until the transition to middle school between fifth

Fig. 2. Student diffusion, disappearance, and retention. (A) Diffusion plots. These show the changes in student numbers due to random processes such as guessing on the exam (Eq. 11). To find the total change in numbers of students in every cell including contributions from diffusion, add these vertical arrows to the convective arrows of Fig. 1A. (B) Students appearing and disappearing from school or retained in a grade, deduced from mathematics exams administered in spring 2006 and spring 2007. Vertical arrows show the net result of students appearing and disappearing from school, whereas horizontal arrows show the numbers of students required to repeat a grade. Downward pointing arrows mean that more children are disappearing from a grade than appearing in it. Areas of arrows are proportional to the numbers of children involved. The scale of arrows and meaning of the colored bands is the same as in Fig. 1.
and seventh grade, when students from economically disadvantaged backgrounds flow downwards at a higher pace than their less disadvantaged counterparts and never recover. Ninth grade is another crucial time because students who are not passing the mathematics exams are forced to repeat a grade and consequently disappear from schools in large numbers. This effect is much stronger for those who are economically disadvantaged than for those who are not, as shown in Fig. 2B.

Flow fields address many questions about the educational system. There is a debate over the student variables that should be used to describe effects of teachers and schools. Sanders (8) states that “models should not include socio-economic or ethnic accommodations but should only include measures of previous achievement of individual students.” In this view, prior year scores contain everything one needs to know about the state of the students. However differences between flow directions have great statistical significance. For example, sixth graders not eligible for free and reduced meals and mathematics scores between 90% and 100% in 2006/2007 drop on average in score by 4.4% the next year, whereas those eligible for free and reduced meals drop in score by 7.0%. \( N \sim 30,000, \ t = 34, \ p < 10^{-7} \). Similar statistical significance applies to the differences between virtually all the arrows in the upper and lower rows of Fig. 1. Changes in scores depend strongly, reproducibly, and with high statistical significance, upon poverty level even after controlling for previous achievements of students. It is possible that this difference in score changes is entirely due to the lower quality of teachers assigned to the least affluent students. However, it is difficult to reach such a conclusion simply from test data; the conclusion that ineffective teachers are largely to blame for unsatisfactory student performance risks being circular (9) if ineffective teachers are defined to be those whose students’ test scores decrease (10). Drawing conclusions about school effectiveness from test data presents comparable difficulties (11).

Another claim is that the difficulty of items on the TAKS exams is carefully chosen so as to maintain students’ scores at the same level over time (12). This claim is only partly supported by the data. Most flow vectors are close to horizontal but the slopes are not negligible and, over the course of several years, students flow to very different regions from where they began, a change which depends strongly on variables such as students’ ethnic group and economic class.

Because the impetus to pass No Child Left Behind came from Texas, there have been debates on differing reasons Texas test scores have been rising. McNeill et al. (13) suggest that rises in Texas test scores can be attributed to an increasing pattern of retaining low-income students at ninth grade until they disappear. Data support this claim. Retention of low-income students (those eligible for free and reduced lunch) at ninth grade increased between 2003/2004 and 2006/2007 from 31,200 to >33,000, whereas the number of low-income students disappearing from ninth grade increased from 23,200 to 28,500.

Finally, we note that linear modeling is pervasive in analysis of educational data (14), but we see many effects that are inherently nonlinear. For example, suppose that through some form of improved instruction it is possible to increase the score gains of low-income students in sixth and seventh grades. This will have the effect of diverting the entire flow pattern slightly upward. In 10th grade, the number of low-income students in the highest score bracket (90%-100%) constitutes the exponentially small tail of a distribution centered at around 60%. Thus small motions of the distribution upward result in exponentially varying gains at the top. Such gains are evident. For example, consider the low income students scoring >90% in 10th grade circled in Fig. 3. Between 2003 and 2006, the number of these students grew by 15-50% per year, starting with 2,150 students in 2003 and ending with 7,088 in 2006. The model presented here is nonparametric and makes minimal assumptions about the form of the underlying probability distributions for student score changes.

It should be possible to use our convection and diffusion models in order to predict quantitatively how improvements at lower grades affect flow of students at higher grades. These predictions would apply to the average behavior of large numbers of students although individuals would display considerable variation. The success of these predictions will partly hinge on the extent to which score changes in successive years are statistically independent. Our preliminary examination of this question indicates that knowledge of two prior year’s scores only improves prediction by \( \approx 10\% \), so the assumption of independence is acceptable.
Materials and Methods

Dataset. We obtained all TAKS-related data that the Texas Education Agency, the government agency charged with administering and evaluating standardized tests in Texas, is able to release. Each row of the dataset contains information about a single examination of a single student, including student demographic details as well as their responses to each question of the examination along with their score. Each student has a globally unique, anonymized identifier that allows us to follow them through time as they move between schools. Students are described by race, gender, and eligibility for free and reduced-price meals, which is an indication of whether family income is low or high. Every answer they have bubbled on each test is provided, and can be compared with correct answers. The schools and districts in which students take the exams are named. There are some things we would have liked to know that are not included; there is no indication of whether students have changed schools in the middle of the year. The State probably does not know. More puzzling, there is no information on the teacher under whose care the student took the exam. The State certainly does have this information, because it returns to each teacher a record of their students’ scores, and provides reports on the performance of every teacher to each district. However they must not retain the data, because the Texas Education Agency is required release all information in their possession in accord with the Texas Freedom of Information Act, and data linking students to teachers are not available.

The data set has defects, some of which can partially be remedied and some of which cannot. There are over 27,000 students with invalid records who end up coded with the same unique identifier and must be removed. Such defects appear to involve tens of thousands of students but there is nothing to be done about them. Out of a population of millions we do not believe that these defects are likely to distort our results.

We transformed the dataset to make analysis more manageable, but without changing any entries. We created normalized sets of tables to speed up searches in MySQL. We also produced condensed files containing all relevant information, because it returns to each student on one line, useful for analysis with Python scripts. We normalize all students’ scores by dividing by the maximum possible score for that student for that exam.

Statistical Methods. Let \( n^b_a \) and \( n^w_a \) be the numbers of students in years \( b \) and \( w \) with score \( S \). \( N \) will always depend upon some other variables as well, such as the grade level, perhaps economic need or race of the student, but we suppress for the moment so as to focus on the primary variables of scores and time. Let \( R^b_{ab} \) be the number of students with score \( S \) in year \( a \) who score \( S \) in year \( b \). The master equation is

\[
N^w_a - N^b_a = \sum_{S} \left[ R^b_{ab} S_{ab} - R^w_{ab} S_{ab} \right] = \sum_{S} n^b_a R^b_{ab} - \sum_{S} n^w_a R^w_{ab} \quad [1]
\]

Transitions to and from the state with score zero have to be treated separately because they correspond to students who were sick, absent for other reasons, left school, left the country, or have an invalid exam. We do not distinguish between students who show up in the dataset with a zero score and those who do not appear at all. We define

\[
\Delta S = R^a_0 - R^b_{ab} \quad [2]
\]

to be the disappearance of students with score \( S \) between years \( a \) and \( b \). To obtain a Fokker–Planck equation, assume that \( R \) is slowly varying as a function of \( S \), although not slowly varying as a function of \( a \). Then, to second order in score changes,

\[
R^b_{ab} S_{ab} \approx R^b_{ab} S_{ab} - \frac{\Delta S}{\Delta S} \left( \frac{S}{\Delta S} \right)^2 \frac{\Delta S}{\Delta S} \quad [3]
\]

This gives

\[
N^w_a - N^b_a \approx -\Delta S \left( \frac{S}{\Delta S} \right) \frac{\Delta S}{\Delta S} \quad [4]
\]

where the forward flow \( v_f \) and the forward diffusion \( D_f \) are defined by

\[
v_f = \sum_{S} \frac{R^b_{ab} S_{ab} - R^w_{ab} S_{ab}}{N^b_a} \quad [5]
\]

\[
D_f = \sum_{S} \left( \frac{S}{\Delta S} \right)^2 \frac{\Delta S}{\Delta S} \quad [6]
\]

The forward flow \( v_f \) gives the average score change of students with score \( S \) in year \( a \) who also have a (nonzero) score in year \( b \). The diffusion coefficient \( D_f \) sets the magnitude of random variations in scores. One can repeat the derivation of Eq. 4 but Taylor expand the second term of Eq. 1 rather than the first. This leads to the reverse flow \( v_r \) and reverse diffusion \( D_r \) :

\[
v_r^w_a - v_r^b_a = \sum_{S} \frac{R^b_{ab} S_{ab} - R^w_{ab} S_{ab}}{N^b_a} \quad [7]
\]

\[
D_r = \sum_{S} \left( \frac{S}{\Delta S} \right)^2 \frac{\Delta S}{\Delta S} \quad [8]
\]

\[
N^w_a - N^b_a = \Delta S \left( \frac{S}{\Delta S} \right) - \frac{\Delta S}{\Delta S} \frac{\Delta S}{\Delta S} \left( \frac{S}{\Delta S} \right)^2 D_r N^b_a \quad [9]
\]

The reverse flow \( v_r \) answers the question “If a student has a score between 80% and 89% in 12th grade, what is the most likely path to have been followed since third grade?” The average of the forward and reverse flows is a current that predicts changes in student numbers without diffusion:

\[
N^w_a - N^b_a + \Delta S = \Delta S \left( \frac{S}{\Delta S} \right) - \frac{\Delta S}{\Delta S} \frac{\Delta S}{\Delta S} \left( \frac{S}{\Delta S} \right)^2 D_r N^b_a \quad [10]
\]

Subtracting Eq. 4 from Eq. 9 gives the identity

\[
J_d = -v_d^a N^a_a - v_d^b a N^b_a = -\frac{1}{D_r} \left( D_r a N^a_a + D_r b N^b_a + \right) \quad [11]
\]

which means that the difference between the forward and reverse flows or the sum of the forward and reverse diffusions provides a measure of total diffusion.

We now can interpret more precisely what we have plotted. Fig. 1A displays the vector \((it, \Delta S)\), where \(it = 1\) is the horizontal distance from one grade level to the next, with the vector scaled in the vertical direction so that \(i = 1\) corresponds to the height difference between 0 and 100%. The flow plots show streamlines of the most likely future path of students. Fig. 1B displays the vector \(-\Delta S, v_f a\). Fig. 2A plots triangles of height \(J_d\), and whose width is proportional to \( D_r a N^a_a + D_r b N^b_a \). Note that \(J_d = v_f a = J_d\), so the vertical components of the vectors plotted under Flow and Diffusion sum to the average score changes of all students including both the effects of convective flow and diffusion. The vertical arrows in Fig. 2B are the disappearance rate \(\Delta S\). The horizontal arrows are computed similarly, and are obtained from the total number of students in two consecutive years found to be repeating a grade.

Acknowledgments. We thank Phil Croner for helping to put the data in normalized form, Stephen Stigler for posing stimulating questions following a presentation, the Texas Education Agency for supplying preliminary data, and the University of Texas Dallas Educational Research Center and the Ray Marshall Center at the University of Texas at Austin for access to final unfiltered data. This work was supported by National Science Foundation Grant DMR 0701373.