Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Big data in education and the models that love them

Zachary A Pardos

As the modal sources of data in education have shifted over the past few decades, so too have the modeling paradigms applied to these data. In this paper, we overview the principle foci of modeling in the areas of standardized testing, computer tutoring, and online courses from whence these big data have come, and provide a rationale for their adoption in each context. As these data become more behavioral in nature, we argue that a shift to connectionist paradigms of modeling is called for as well as a reaffirming of the ethical responsibilities of big data analysis in education.

Address
Univeristy of California at Berkeley, Berkeley, CA 94720, USA

Corresponding author: Pardos, Zachary A (zp@berkeley.edu)

Current Opinion in Behavioral Sciences 2017, **18**:107–113

This review comes from a themed issue on **Big data in the behavioural sciences**

Edited by **Michal Kosinski** and **Tara Behrend**

For a complete overview see the Issue and the Editorial

https://doi.org/10.1016/j.cobeha.2017.11.006

Introduction

The primary role of modeling in education has varied as data collection and analysis traverse the contexts of testing, tutoring, and online instruction at scale. Each context has brought its own unique critical foci of practice, thus forming related, but distinct, constituent academic research fields of study. The types of data produced have also differed, as a function of the context, and necessitated the development and adoption of different modeling paradigms. This paper offers a rationale and historical context for these differences and is intended to serve as an entry point for data modeling research in the adjacent fields of psychometrics and learning analytics.

Data

In this section, we will overview many of the modal sources of big data in education, the volume and character of the data, and historical context in which they were produced. Large scale standardized testing has been the original producer of high volume data in education. The SAT, originally the Scholastic Aptitude Test, was created by the not-for profit association of institutions called the College Board and first administered to high school seniors in 1926. The test consists of reading, writing & language, and math sections and produced 252 million item answer records (responses) from 1.63 million students in 2016.[1] In higher education, the Graduate Record Examinations (GRE) test, created by Educational Testing Service (ETS), was first administered in 1949 and presently covers topics in algebra, geometry, arithmetic, and vocabulary. The test, required by many graduate school programs, was taken by 584 000 respondents in AY 2015–2016,[2] producing around 24 million responses. The design objective [1] of these test providers is to craft items for the instruments, that are their tests, such that they reliably and accurately estimate abilities which correlate with post-secondary performance and which are of high relevance to college admissions offices. Student answers to items on the test are referred to in the field of measurement as dichotomous response data because the answers (responses) are scored with binary outcomes of correct or incorrect; although the GRE contains three essays and the SAT contains one optional essay which are scored on a continuous scale. These essays are scored by one human and one algorithmic rater [2]. If the algorithm does not agree with the human rater, a second human rater will score the essay to break the tie. Table 1 shows an example of this dichotomous (binary) response data collected from standardized tests.

The large test providers are in possession of these massive datasets which are not public and generally not shared with outside researchers. In practice, researchers in the field of Psychometrics most commonly use much smaller datasets on the order of thousands of respondents. A frequently cited source of data is Kikumi Tasuoka's fraction subtraction test [3], which is available by request. Synthetically generated datasets are also of high popularity in studying the properties of different approaches to estimation [4] in the various models used to represent ability. Aggregate results are provided by the Organization for Economic Co-operations and Development (OECD) for its Programme for International Student Assessment (PISA) test administered every three years since 2000, with 540 000 test takers across 72 participating countries in 2015.[3] Other sources of high volume data in

[1] https://secure-media.collegeboard.org/digitalServices/pdf/sat/total-group-2016.pdf
[2] https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2016.pdf
[3] https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Table 1

Example of data collected from standardized tests. These data are referred to as dichotomous responses because of their binary nature. A '0' represents an incorrect answer to an item (question) and '1,' a correct answer.

Student ID	Item-1	Item-2	Item-3	Item-4
Janelle	0	1	1	1
Zeus	1	1	0	0
Erin	0	1	1	1

educational assessment include; the National Education Longitudinal Study (NELS) of 1998 and 2002,[4] Trends in International Mathematics and Science Study[5] (TIMSS) run every four years, and the Measures of Effective Teaching dataset, provisioned by the Bill & Melinda Gates Foundation [5].

While the field of Psychometrics is *primarily* concerned with the measurement of student abilities from low periodicity summative tests, the broad field of the learning sciences, including learning analytics, has focused on facilitating and measuring change (growth) in student ability and has used the mechanism of computer tutoring systems to scale instructional approaches toward that end. The first computer tutoring systems offering automated self-paced instruction were seen as early as 1960 [6]. Tutoring systems with more sophisticated adaptive qualities began to develop through the 70s and 80s and were branded intelligent tutoring systems [7]. These systems were in large part inspired by the efficacy of one-on-one human tutoring, later shown to produce learning gains up to two grade levels, or two standard deviations above that accomplished with traditional one-to-many classroom instruction [8]. Whereas standardized tests, because of their one-time summative nature, often measure large constructs like mathematics ability, tutoring systems, used throughout the school year, can measure finer-grained constructs. The terminology of 'skills' or 'knowledge components' is often used in tutoring system contexts, in place of 'constructs.' This granularity in tutoring systems matches their formative instructional design of measuring mastery of a set of skills before allowing the student to progress to the next section in the tutor. The legislation passed in the United States requiring every state to have a standardized test that students needed to pass in order to be awarded a high school diploma (No Child Left Behind, 2001[6]), in part, catalyzed the development and adoption of tutoring systems through the mechanism of federal grant funding. One such system, supported by the National Science Foundation's (NSF) Science of Learning Center grants, was an intelligent tutoring system called the Cognitive Tutor [9]. The Cognitive Tutor, still in operation as of this writing, had around 600 000 students using its various curricular systems in 2012, and in its most popular product, Bridge to Algebra, produced 1 billion events from students that year. These data are referred to as learner process data because they are produced from student engagement with material meant to facilitate learning. These data represent students' longitudinal interactions with problems in the system. An example of data from an intelligent tutoring system is shown in Table 2. These data are similar to standardized testing data in that they are response centric (one row per answer), but contain other information important to characterizing and measuring learning from the time series of responses. Other features of the response include the time when the answer was given, the skill associated with the question being answered, and the number of times the student had seen questions of that same skill thus far.

Other meta information about the student's interaction with the problem can also be included, such as if a hint was requested. In tutoring systems, students are often able to attempt a problem more than once but the response recorded for the question most commonly reflects the correctness of the answer given by the student on her first attempt. Thus, in the Cognitive Tutor, a row can represent all the interactions of a student with a particular problem and can include other meta information such as the number of total attempts made by the student to answer the problem correctly. This format is called step-rollup in Cognitive Tutor data as it is a summary of a student's interaction with each step. The word 'step' is used to refer to the fine-grained questions posed in the tutor.

Researchers have enjoyed a high degree of access to data from computer tutoring systems. The Cognitive Tutor in particular has made de-identified step-rollup level data publicly available. Many datasets from the Cognitive Tutor and other computer tutors can be found on an NSF sponsored project called DataShop.[7] The largest of the public datasets [10] was provided as the focus of a data mining competition whereby participants were given response data from students on the first portion of each lesson in the tutor and were tasked with predicting the correctness of the students' answers in the proceeding redacted portion of each lesson. This dataset[8] contained four separate datasets from two years of two different tutoring products; 'Algebra' and 'Bridge to Algebra,' a more remedial version. In total, 37.4 million student-step events are available in this combined dataset. The ASSISTments Platform [11] is another example of a tutoring system which has been generous with its dataset contributions to the research community. It has been fashioned to be more teacher oriented than its

[4] https://nces.ed.gov/surveys/els2002/
[5] https://nces.ed.gov/timss/
[6] http://files.eric.ed.gov/fulltext/ED447608.pdf
[7] https://pslcdatashop.web.cmu.edu/
[8] https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp