Title
Big data in education and the models that love them

Author
Pardos, ZA

Publication Date
2017-12-01

DOI
10.1016/j.cobeha.2017.11.006

Peer reviewed

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

Table 2

Example of data from an intelligent tutoring system. One row is generated per student per response with meta information pertinent to modeling their mastery of a skill.

Student ID	Time	Opportunity count	Skill	Response
Zeus	6/5/2017 10:01	1	Parallelogram perimeter	0
Zeus	6/5/2017 10:09	2	Parallelogram perimeter	1
Zeus	6/5/2017 10:13	1	Rectangle Area	1
Erin	6/5/2017 10:00	2	Circle Circumference	1
Erin	6/5/2017 10:05	1	Circle Area	1

predecessors, albeit less intelligent [12]. Its full 2012–2013 academic year public dataset[9] comprises 6.1 million rows of student-problem level data with additional meta information on the exact answer the student gave, an anonymized school and teacher ID, among other meta information. This computer tutoring system was also supported by funding from the federal government in a testing climate pushing states, and their teachers, to see students well prepared to pass their mandatory standardized state test. Early versions of the ASSISTments Platform used previously released standardized test items from its state of operation, Massachusetts, as its primary source of problem content. The majority of content now in ASSISTments is not from tests but instead generated in-house or by teachers.

The approach of intelligent tutoring systems and their derivates has largely been a problem-first pedagogy, holding as their primary instructional hypothesis that problem solving, with immediate feedback and scaffolded supports, serves as a more effective method of learning a new subject than the relatively passive consuming of lectures and texts [13•]. However, the human deconstruction of a task into fine-grained components, a process called cognitive task analysis, is an expensive one, and while this process has seen success in its application to courses in statistics [14], the scalability of this manual coding process to the expanse of all subject areas in higher education is daunting. Furthermore, there is an incompatibility between the problem-first approach of tutors and the current lecture-first pedagogical approach of most post-secondary courses. With a high number of knowledge components needed to cover the tasks in college level curricula, the number of assessment problems typically given in a course would have to substantially increase. Retrofitting ITS to higher education, at least when strictly adhering to ITS domain model principles, seems intractable.

Forty years after intelligent tutoring systems were introduced to high-schools, higher-education experienced its own digital renaissance in the form of Massive Open Online Courses (MOOCs) after two Stanford computer science professors offered the first MOOC in 2012, drawing over 100 000 enrollments. After realizing the demand for such freely accessibly, high quality instructional material, they formed a for-profit company called Coursera. Shortly after, a non-profit venture between Harvard and MIT, called edX, was created to also provide high-quality, freely accessible, online courses but with a glance toward facilitating research in education. MOOCs, both as instructional resources and as the subject of research, differ from the e-Learning and distance education movements that proceeded it in that those movements were based on the waning hypothesis that the costs of a community college education could be driven down through the scaling of low cost online instruction while maintaining the same level of instructional value. MOOCs are a presentation of university level courses, many from prominent institutions, using high production value user interfaces, software design, and video lectures. A response to apparent popular demand, MOOCs did not and still do not represent a unified pedagogical hypothesis being tested. Likewise, MOOCs do not share a single critical pedagogy, a reflection of their residentially taught course inspirations, but instead tout universal access as their primary innovation. The relevance of MOOCs to the big data landscape is that they produced detailed clickstream behavior of, in some cases, tens of thousands of learners' interactions with the various elements of a single course. The 290 courses offered by MIT and Harvard in the first four years of edX produced 2.3 billion logged events from 4.5 million learners [15•]. Unlike intelligent tutoring systems, the events logged from MOOCs went beyond problem solving and encompassed video play, pause, and scrub events, page navigations, problem answer text, and social discussion board data. Table 3 depicts an example of these event log data. Additionally, exported datasets[10] made available to edX university partners included learners' optionally provided demographic information such as age, country of origin, gender, highest level of education, and stated motivations for signing-up for an edX account.

Each edX University partner has access to the full event log and demographic datasets for their courses, past and present. The logs are updated on a daily basis. A research data exchange (RDX) program allows participating edX

[9] https://sites.google.com/site/assistmentsdata/home/2012-13-school-data-with-affect

[10] http://edx.readthedocs.io/projects/devdata/en/latest/internal_data_formats/index.html

Table 3

Example of event level data from a learner's interactions with an online course. In the example, the learner loaded the intro course page and then, two minutes later, began playing a video on the intro page. She paused the video after one minute then loaded the next page, which was a quiz. She submitted her answer, '3.1415,' to the first question of the quiz five minutes later, which was the correct answer.

Student ID	Time	Event type	Resource	Meta info
Janelle	6/5/2017 10:01	Load_page	/lec1/intro.html	–
Janelle	6/5/2017 10:03	Play_video	/lec1/welcome.vid	0:00
Janelle	6/5/2017 10:04	Pause_video	/lec1/welcome.vid	1:00
Janelle	6/5/2017 10:08	Load_page	/lec1/quiz.html	–
Janelle	6/5/2017 10:13	Answer	/lec1/quizQ1	'3.1415' [Correct]

partners to request de-identified data from courses of other partners participating in the program. Coursera has data sharing agreements only at the individual instructor level, making institution based research more difficult to facilitate as researchers retrieve the data from the instructors rather than centrally through a university data administrator contact. No event level MOOC datasets have been made public; however, researcher requests for MOOC event level data under a memorandum of understanding (MOU) is available from Stanford,[11] MITx,[12] and HarvardX.[13] A student-course level aggregated dataset was made public by MIT and Harvard [16] summarizing learner interactions with the first 17 courses offered by the two institutions. The dataset was heavily de-identified with perturbations of the data made to ensure re-identification was not possible. The bias introducing side-effects of this form of anonymization were well documented [17] and greatly reduced the utility of the dataset to be mined for novel inferences and correlational findings.

Models

In this section, we will overview the canonical modeling frameworks applied to each of the three modal source of big data described in the previous section. In standardized testing, the Rasch model [18] has been the cornerstone statistical framework for estimating students' abilities and item difficulties from responses to a test, a paradigm called Item Response Theory (IRT).

$$P(Y_{ij}=1|\theta_j)=\frac{e^{\theta_j-\beta_i}}{1+e^{\theta_j-\beta_i}} \tag{1}$$

As shown in Equation 1, IRT is based on a standard logistic function, where the dependent outcome variable Y is the dichotomous response (correct or incorrect) for a student j on test item i conditioned on the ability θ of the student. The probability of a correct response evaluates to 0.50 when the ability of the student equals the difficulty of the item β_i. For example, if θ_{j} and β_i are both 2, then the numerator becomes one (e^0) and the denominator is two, resulting in a probability of correct of 0.50. As the difficulty or ability changes by one, relative to one another, the probability of correct changes approximately with percentages corresponding to standard deviations. That is, if the student's ability is three greater than the difficulty of the item, the probability of correct will be 0.9526. If a student's θ were estimated to be 3, they would be reported to be in the top five percentile of test takers. The abilities of all test takers are estimated after responses are collected; however, the parameters of the items are typically estimated beforehand on smaller sample cohorts during the development of the test. Item difficulties estimated during test development are used to select the items that will comprise the test to make sure the desired variety of item difficulty is represented.

Large scale standardized tests are focused on the measurement of a construct [19] at a single point in time, known as *summative assessment*. In the environment of computer tutoring systems, continuous measurement that informs instructional decisions, known as *formative assessment*, is called for. Longitudinally oriented models and finder-grained decomposition of constructs are therefore used by tutoring systems to track students' mastery in these constructs and adaptively guide them through lessons. Variations on the IRT logistic model have been introduced which consider growth with respect to the number of problems a student has attempted of a particular skill [20,21]. The statistical framework of Bayesian Networks, however, has been the implementation of choice inside of the Cognitive Tutor, primarily due to the affordances of Bayes theorem which allows for an update to the inference of student mastery of a skill based on new evidence, such as the correctness of the last response to a problem of the same skill, without needing to re-estimate model parameters.

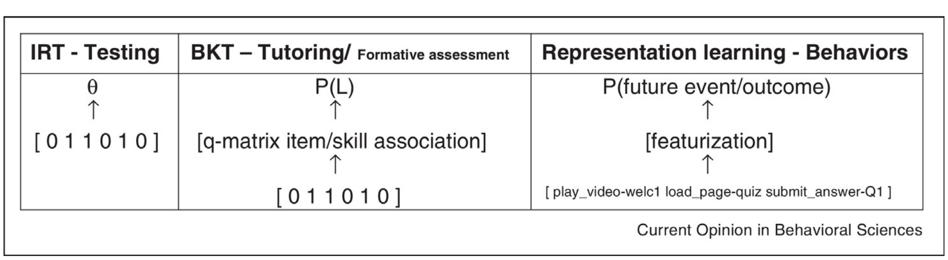
$$\begin{aligned}P(K|Q)&=\frac{P(Q|K)P(K)}{P(Q)}\\&=\frac{(1-P(S))P(L_t)}{(1-P(S))P(L_t)+(1-P(L_0))P(G)}\end{aligned} \tag{2}$$

[11] https://iriss.stanford.edu/carol/research
[12] http://web.mit.edu/ir/mitx/
[13] https://vpal.harvard.edu/access-vpal-research-data

The Bayes' theorem equation is shown above and assumes the positive value of all the binary variables ($Q = 1$ and $K = 1$). On the right, is the theorem expressed in terms of the parameters of the Bayesian Knowledge Tracing model [22]. This expression is the updated probability (posterior) that a student has mastered the skill K given observation of a correct response on a problem associated with skill K. This is the probability that the correct answer was observed because the student knows the skill, the compliment of the probability of slip $P(S)$, answering incorrectly even though the skill is known, multiplied by the current estimate of knowledge $P(L_t)$ divided by that same value plus the probability that the answer was observed because the student did not know the skill and guessed $P(G)$. Since students are in a tutoring context and feedback and hints can be given between opportunities to answer questions, growth is assumed and represented by parameter $P(T)$ which increases the probability of knowledge between opportunities with $P(L_t) = P(L_{t-1}) + (1 - P(L_{t-1}))P(T)$. Forgetting, in the canonical form of the model, is assumed to be negligible (zero) within the short lesson segments in which problems of a skill appears. On the very first opportunity to answer a question of a skill, the probability of knowledge at the previous opportunity is not known $P(L_0)$ and so this initial prior probability value serves as a fourth parameter to be estimated. The association of questions to skills is conducted by subject matter experts through a process called cognitive task analysis. A strong assumption is made with this model that all skills are independent of one another, therefore a different set of parameters are learned for each skill independently from data collected from past cohorts of students using the tutor. Within the tutor, if the inference of $P(L)$ is greater than 0.95, the student is assumed to have mastered that skill and is no longer asked to practice it in the tutor. This dynamic prescription of practice is the primary aspect of adaptivity in intelligent tutoring systems.

Different layers of abstraction are present in models applied to data from testing, tutoring, and online course contexts (Figure 1). In IRT, item responses are input and used directly to estimate ability. In intelligent tutoring models, the question responses go through a look-up layer where the question is associated with a skill and the data point then represents a response to the skill, with the item abstracted away. In this case, this look-up, or Q-matrix, is hand specified but can be iteratively refined from data [21]. In an online course context, the available input includes event stream (or clickstream) data which features inputs of mixed type. The output can be any logged outcome, such as certification in the course [23], stop-out [24,25] or predicting what action the learner is going to take next in the course [26•]. In this case, behavior is the input and the output. What is difficult about this scenario is that theoretical foundations, and thus intuitions for explaining behavior, are often lacking which subject matter experts need to bring to bear to construct effective features. This lack of intuition about what governs the behaviors of students in MOOCs is in part what has fueled the numerous approaches to visualize MOOC data in order to better understand it [27]. Instead of hand engineering features a priori, neural networks, which have distributed featurization at the core of their generalizing principle, can be used in place to induce features. Featurization approaches have been demonstrated prominently outside of education in autoencoders used to reconstruct the bag of words of a document from a lower dimensional featurization of the same document [28], and more recently in word2vec approaches which featurize words with respect to the contexts they appear in [29]. These simpler, linear representation learning models embed inputs into a vector space which can be interrogated to de-bias sexist language in large text corpora [30] or to reason about the compositionality of academic departments [31••] and problems in a tutoring system [32]. While interpretation of neural network hidden states is an area of continuing research, dimensionality reduction techniques which lend themselves to visually surfacing regularities in their representations [33] have opened up their inner workings to productive scrutiny [34]. Still, the application of neural networks does not guarantee superior results when compared to classical approaches in classical data scenarios. Piech *et al.*'s [35••] application of neural networks to the knowledge tracing scenario demonstrated the pre-requisite inference utility of

Figure 1



Model abstraction layers for Item Response Theory, Bayesian Knowledge Tracing, and representation learning (neural network) models applied to learner process data from online courses.

modeling all skills at once, unlike BKT which treats them as independent; however, its predictive performance was shown to be on par with modern extensions of the BKT model [36•]. Connectionist models most distinguish themselves when applied to more complex data from less traditionally structured behavioral contexts.

Synthesizing research

Scarcity of scholarly interaction between model researchers in the fields of psychometrics and learning analytics is in part attributable to differences in disciplinary training. Asserting distributional assumptions that improve model fit and reliability of testing instruments is in the bailiwick of the statistician, which makes up the majority background of researchers in Psychometrics, while the computer scientist and engineer is at home developing adaptive learning technologies with models validated through predictive generalization. The cognitive scientist finds residence in either area. There is evidence, however, of the two fields becoming more interested in aspects of modeling long studied by the other. Psychometric research is increasingly interested in developing longitudinal models of fine-grained constructs [37,38], while learning analytics is increasingly interested in the validity and reliability of their models, studying their convergence properties [39,40] and the metrics and criterion used to arrive at the intended measure [41]. A shared research pursuit may be on the horizon. Standardized testing fatigue is setting in, politically, with students being tested an average of 20–25 hours per year in the US [42]. Learner process data may be turned to as the source for continuous assessment [43] which will require both statistical assurances of reliability and algorithmic capabilities for extracting information from complex behavioral data.

Privacy implications

It is the nature of privacy for its essence and boundaries to be questioned and debated [44]. With the amount of information contained in behavioral data, the privacy implications of its analysis in educational contexts is similarly contested and at times met with hesitation and uncertainty by organizations and government [45]. US law has provided little in the way of guidance on how to proceed, inheriting an antiquated concept of what constitutes the protected student 'record' and lacking due caution with respect to publicly subjecting de-identified data to the contemporary dangers of re-identification [46]. The current practice of the educational analytics community has been to rationally weigh the tangible harms of big data analysis with the benefits [47], acknowledging an opportunity cost to *not* innovating with data in the learning sciences and in practice [48]. The debate around big data analysis can be analogized to ethnography. Should the observer stay far enough away so as not to affect the behaviors of the observed or is authentic collaboration and engagement with the minutia of their everyday lives a necessity to the understanding needed to shape practices for the better [49]? The answer will likely hinge on the degree to which these activities are perceived as serving the interests of the learner.

Conflict of interest statement

Nothing declared.

Acknowledgement

This work was supported in part by a grant from the National Science Foundation (Award #1547055).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Mislevy RJ, Haertel GD: **Implications of evidence-centered design for educational testing**. *Educ Meas Issues Pract* 2006, **25**:6-20 http://dx.doi.org/10.1111/j.1745-3992.2006.00075.x.

2. Leacock C, Chodorow M: **C-rater: automated scoring of short-answer questions**. *Comput Humanit* 2003, **37**:389-405.

3. Tatsuoka KK: **Toward an integration of item-response theory and cognitive error diagnosis**. *Diagn Monit Skill Knowl Acquis* 1990:453-488.

4. Baker FB, Kim S-H: *Item Response Theory: Parameter Estimation Techniques*. CRC Press; 2004.

5. Kane TJ, McCaffrey DF, Miller T, Staiger DO: *Have we Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. MET Project*. Bill & Melinda Gates Foundation; 2013.

6. Smith SG, Sherwood BA: **Educational uses of the PLATO computer system**. *Science* 1976, **192**:344-352.

7. Corbett AT, Koedinger KR, Anderson JR: **Intelligent tutoring systems**.*Handbook of Human-Computer Interaction* . 1997:849-874.

8. Bloom BS: **The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring**. *Educ Res* 1984, **13**:4-16.

9. Koedinger KR, Anderson JR, Hadley WH, Mark MA: **Intelligent tutoring goes to school in the big city**. *Int J Artif Intell Educ* 1997, **8**:30-43.

10. Stamper J, Pardos ZA: **The 2010 KDD Cup Competition Dataset: engaging the machine learning community in predictive learning analytics**. *J Learn Anal* 2016, **3**:312-316 http://dx.doi.org/10.18608/jla.2016.32.16.

11. Heffernan NT, Heffernan CL: **The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching**. *Int J Artif Intell Educ* 2014, **24**:470-497.

12. Baker RS: **Stupid tutoring systems, intelligent humans**. *Int J Artif Intell Educ* 2016, **26**:600-614.

13. • Koedinger KR, Kim J, Jia JZ, McLaughlin EA, Bier NL: **Learning is not a spectator sport: doing is better than watching for learning from a MOOC**. In *Proceedings of the Second ACM Conference on Learning@ Scale*. ACM; 2015:111-120.

Drew attention to the contrast between the problem-first pedagogical approach of tutors and the lecture-first approach of online courses. Their application of a causal inference tool to observational clickstream data suggested a higher educational utility to problem solving than video watching.

14. Lovett M, Meyer O, Thille C: **JIME — the open learning initiative: measuring the effectiveness of the OLI statistics course in accelerating student learning**. *J Interact Media Educ* 2008.

15. Chuang I, Ho AD: *HarvardX and MITx: Four Years of Open Online*
• *Courses — Fall 2012-Summer 2016*. 2016 http://dx.doi.org/10.2139/ssrn.2889436.
A definitive white paper on the activities and demographics of MOOC users as told through a series of nine pointed descriptive statistical analyses.

16. MITx, HarvardX: *HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0*. Harvard Dataverse, V10; 2014 http://dx.doi.org/10.7910/DVN/26147.

17. Angiuli OM: *The effect of quasi-identifier characteristics on statistical bias introduced by k-anonymization*. 2015. (Doctoral dissertation).

18. Rasch G: **On general laws and the meaning of measurement in psychology**. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4*. 1961:321-333.

19. Wilson M: *Constructing Measures: An Item Response Modeling Approach*. Routledge; 2004.

20. Pavlik PI, Cen H, Koedinger KR: **Performance factors analysis — a new alternative to knowledge tracing**. *Proceeding of the International Conference on Artificial Intelligence in Education*; *Amsterdam, The Netherlands: 2009:531-538*.

21. Koedinger K, McLaughlin E, Stamper J: **Automated student model improvement**. In *Proceedings of the 5th International Conference on Educational Data Mining*; *Chania, Greece: 2012:17-24*.

22. Corbett AT, Anderson JR: **Knowledge tracing: modeling the acquisition of procedural knowledge**. *User Model User Adapt Interact* 1994, **4**:253-278.

23. Coleman CA, Seaton DT, Chuang I: **Probabilistic use cases: discovering behavioral patterns for predicting certification**. In *Proceedings of the Second ACM Conference on Learning@ Scale*. ACM; 2015:141-148.

24. Yang D, Sinha T, Adamson D, Rosé CP: **Turn on, tune in, drop out: anticipating student dropouts in massive open online courses**. In *Proceedings of the 2013 NIPS Data-driven Education Workshop, vol. 11*. 2013:14.

25. Boyer S, Veeramachaneni K: **Transfer learning for predictive models in massive open online courses**. *International Conference on Artificial Intelligence in Education*. Cham: Springer; 2015, 54-63.

26. Tang S, Peterson J, Pardos Z: **Predictive modelling of student**
• **behaviour using granular large-scale action data**. In *The Handbook of Learning Analytics,* edn 1. Edited by Lang C, Siemens G, Wise AF, Gaevic D.Alberta, Canada: Society for Learning Analytics Research (SoLAR); 2017:223-233.
Demonstrates that there is predictive signal and structure to the click-stream behaviors of learners in a MOOC.

27. Emmons SR, Light RP, Börner K: **MOOC visual analytics: empowering students, teachers, researchers, and platform developers of massively open online courses**. *J Assoc Inf Sci Technol* 2017, **68**:2350-2363.

28. Hinton GE, Salakhutdinov RR: **Reducing the dimensionality of data with neural networks**. *Science* 2006, **313**:504-507.

29. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J: **Distributed representations of words and phrases and their compositionality**. *Advances in Neural Information Processing Systems*. 2013:3111-3119.

30. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT: **Man is to computer programmer as woman is to homemaker? Debiasing word embeddings**. *Advances in Neural Information Processing Systems*. 2016:4349-4357.

31. Pardos ZA, Nam A: **The school of information and its**
•• **relationship to computer science at UC Berkeley**. In *iConference 2017 Proceedings*; *Wuhan, China: 2017:309-316*.
Authors apply word vector models to 'sentences' of course enrollments and reason about courses and departments and their relationship to one another in vector space.

32. Pardos ZA, Dadu A: **Imputing KCs with representations of problem content and context**. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. Bratislava, Slovakia: ACM; 2017, 148-155.

33. Maaten LVD, Hinton G: **Visualizing data using t-SNE**. *J Mach Learn Res* 2008, **9**:2579-2605.

34. Wang L, Sy A, Liu L, Piech C: **Learning to represent student knowledge on programming exercises using deep learning**. In *Proceedings of the 10th International Conference on Educational Data Mining*; *Wuhan, China: 2017:324-329*.

35. Piech C, Bassen J, Huang J, Ganguli S, Sahami M, Guibas LJ,
•• Sohl-Dickstein J: **Deep knowledge tracing**. *Advances in Neural Information Processing Systems*. 2015:505-513.
Using tutor data from Khan Academy and ASSISTments, the authors show recurrent neural networks can perform the knowledge tracing task. The ability of the model to learn representations over skills is demonstrated with an offline inferencing of pre-requisite relationships based on model predictions.

36. Khajah M, Lindsey RV, Mozer M: **How deep is knowledge**
• **tracing?** In *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016).* Edited by Barnes T, Chi M, Feng M. *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)* 2016:94-101.
Authors re-evaluate the results of Piech *et al.* (2015) and find that mild adjustments to BKT, such as allowing for a forget probability, and allowing for individualized student parameters, brings BKT up to parity with DKT in terms of predictive performance, overall.

37. Rupp AA, Templin J, Henson RA: *Diagnostic Assessment: Theory, Methods, and Applications*. New York: Guilford; 2010.

38. Kaya Y, Leite WL: **Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance**. *Educ Psychol Meas* 2017, **77**:369-388.

39. Pardos ZA, Heffernan NT: **Navigating the parameter space of Bayesian knowledge tracing models: visualizations of the convergence of the EM algorithm**. In In *Proceedings of the 3rd International Conference on Educational Data Mining (EDM).* Edited by Baker RSJd, Merceron A, Pavlik Jr PI. *Proceedings of the 3rd International Conference on Educational Data Mining (EDM)* Pittsburgh, PA; 2010:161-170.

40. van De Sande B: **Properties of the Bayesian knowledge tracing model**. *J Educ Data Min* 2013, **5**:1-10.

41. Pelánek R, Řihák J: **Experimental analysis of mastery learning criteria**. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. Bratislava, Slovakia: ACM; 2017, 156-163.

42. Hart R, Casserly M, Uzzell R, Palacios M, Corcoran A, Spurgeon L: *Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis*. Council of the Great City Schools; 2015.

43. Shute VJ, Rahimi S: **Review of computer-based assessment for learning in elementary and secondary education**. *J Comput Assist Learn* 2017, **33**:1-19 http://dx.doi.org/10.1111/jcal.12172.

44. Mulligan DK, Koopman C, Doty N: **Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy**. *Philos Trans R Soc A* 2016, **374**:20160118.

45. Drachsler H, Greller W: **Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics**. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM; 2016:89-98.

46. Daries JP, Reich J, Waldo J, Young EM, Whittinghill J, Ho AD, Chuang I: **Privacy, anonymity, and big data in the social sciences**. *Commun ACM* 2014, **57**:56-63.

47. Ho A: **Advancing educational research and student privacy in the "big data" era**. *Workshop on Big Data in Education: Balancing the Benefits of Educational Research and Student Privacy*. Washington, DC: National Academy of Education; 2017, .

48. Alamuddin R, Brown J, Kurzweil M: *Student Data in the Digital Era An Overview of Current Practices. Report. ITHAKA S+R*. 2016 http://dx.doi.org/10.18665/sr. 283890.

49. Monahan T, Fisher JA: **Benefits of 'observer effects': lessons from the field**. *Qual Res* 2010, **10**:357-376.