Measuring Ability-to-Learn Using Parametric Learning-Gain Functions

Chris Piech, Engin Bumbacher, Richard Davis
Stanford University
piech@cs.stanford.edu, rldavis@stanford.edu

ABSTRACT

One crucial function of a classroom, and a school more generally, is to prepare students for future learning. Students should have the capacity to learn new information and to acquire new skills. This ability to "learn" is a core competency in our rapidly changing world. But how do we measure ability to learn? And how can we measure how well a school has prepared their students to learn? In this paper we formally pose the problem, and introduce a grounded theory of how to measure ability to learn. Using simulations of students learning we provide initial evidence that this theory provides an elegant solution to this problem. We further validate our ideas using real world data from 70k middle-school students and show that our theory is more accurate and interpretable than current state-of-the-art models of learning gains. We consider our results a modest yet interesting first step for a novel type of test.

1. INTRODUCTION

Large-scale, standardized tests typically measure knowledge and skills that students already possess, such as reading comprehension and mathematical competency. However, these tests overlook students' abilities to acquire new knowledge and skills. Could we instead measure how well a student is able to *learn*? Measuring how well a school system has prepared a student for learning is a particularly hard challenge and as such it remains elusive. PISA (Programme for International Student Assessment – an international test run every three years to evaluate educational systems), has made it a goal of their 2024 innovative assessment to measure ability to learn. How could such a test be scored?

Early research has shown that measuring ability to learn is both important and difficult. Work by Schwartz et al. [18, 17, 19] has shown that assessments of students' ability to learn capture important information that assessments which simply measure what a student knows fail to capture. In these studies, students participated in two different educational interventions, one designed to teach students factual content in a manner that also prepared them for future learning, and one designed to teach students factual content using more traditional approaches. Standard measures of knowledge found that regardless of the intervention, students in both groups learned the same factual content. However, a second type of assessment designed to measure students' ability to learn uncovered significant predictive differences.

Despite the potential, to the best of our knowledge, there are no large-scale assessment that have attempted to measure students' ability to learn. In traditional tests, students get questions correct or incorrect — a single random variable that is traditionally modeled using Item-Response Theory (IRT). In a learning test, on the other hand, students work through learning experiences which produce two measurable values: a **prior (pre)** and **posterior (post)** ability. All learning experiences, especially relevant authentic ones, are impacted by what a student knows when they start. A useful model would enable measurement of student learning across countries, schools-districts, and millions of students as they engage in a necessarily wide variety of learning experiences. Without a useful model it is hard (if not impossible) to produce desired and important analyses such as: (a) inferring ability to learn from multiple learning experiences (b) discovering issues of fairness in learning experiences (c) reasoning about mixture effects within populations.

The prior-knowledge confound: Measuring learning-ability is particularly difficult because it requires us to reason about the impact of prior knowledge. For example, consider two populations where students have the exact same ability to learn but different levels of prior knowledge. Now imagine the two populations are given the same learning experience. Both populations will learn (recall they have the same learning ability) but will have different outcomes on the same exam. In practice most people model this relationship using a "linear" model [22]. However, research has shown that the impact of prior ability has important non linear properties [21, 15]. This is an instance of Simpson's Paradox.

A core insight of this paper is to think of the difference between prior and posterior ability as being governed by population specific parametric functions which we call **Learning-Gain Functions**. These learning gain functions are naturally incorporated in a fully Bayesian model of student responses on learning ability tests. The main contributions of this paper are:

1. We formalize, parametric Learning-Gain functions as a

way to model ability-to-learn tests.

2. We introduce an interpretable single-parameter Bayesian family of Learning-Gain functions.
3. We show that this model is able to near-perfectly recover learning ability in a complex, simulated dataset.
4. We demonstrate that this model outperforms other single- and multi-parameter models on two real-world datasets.
5. We show the practical value of this model by comparing real-world schools on their "ability to learn" as estimated by our model

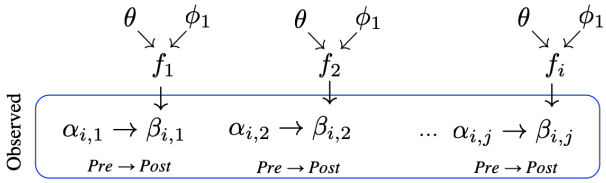
This work is a first attempt at addressing the need for models of student "ability to learn" that can be employed in large-scale assessments such as PISA 2024. The initial results are promising, and we hypothesize that the model will generalize broadly to different learning tests.

1.1 Population Learning-Ability Tests

Learning-ability-tests are built to directly measure the "ability to learn" of a population. The most straightforward format for such an exam has learners complete a set of learning tasks and for each task j, the learner i is given a pre and post test – these fence-post the learning gains. We define alpha ($\alpha_{i,j}$) to be a student's prior ability on the task and beta ($\beta_{i,j}$) to be the student's posterior ability.

We seek to measure ability-to-learn for a population (or an individual as a singleton population) as a number, which we call θ. This measure should generalize and explain ability-to-learn of the population on a different learning-task. In order to learn a generalizable θ we must learn to separate ability-to-learn from task specific effects (such as if the task is easier for beginners to learn than for advanced students etc). We use the notation phi (ϕ_j) to represent task specific parameters for task j.

We propose that when a student engages with a learning task, the learning-ability of the student (θ) interacts with task-specific-parameters (ϕ_j) to produce a **learning-gain-function** (f_j) which determines how prior-abilities will map to post-abilities. As such a function oriented probabilistic model of a single student, from a population with learning-ability θ, working on a series of learning tasks would look like the following:



Learning-ability tests stand in contrast to Intelligence Quotient (IQ) exams as measurement takes place on either end of a learning experience. IQ tests on the other hand measure aptitude, and while this often requires learners to engage in complex tasks the goal is to measure ability on the task.

1.2 Prior work

This work builds on a rich and broad literature of work on measuring ability-to-learn which extends for decades [5, 13, 9]. Evaluation of students' ability to learn is often treated as equivalent with change in knowledge over time, typically with a pretest and posttest. Common approaches include comparison of raw gain scores (posttest minus pretest), analysis of posttest scores with pretest scores as a covariate, and analysis of gains scores with pretest scores as a covariate. Each of these methods has strengths and weaknesses, although there is evidence that analysis of gain scores with pretest scores as a covariate is the best of these methods when certain assumptions are met [6]. As such, we included this model (Linear Multi-Theta) in our model comparison on real-world data and find that it doesn't fit as well. Additionally, while the intercept and slope parameters in the Linear Multi-Theta model can be interpreted as describing a population's ability to learn, it is not immediately clear how they might be used to compare different populations. Both the Learning-Gain-Decay and Learning-Gain-Bump models estimate ability to learn with a single parameter, avoiding this problem. Taken together, these factors suggest that it would be prudent to move away from the Linear Multi-Theta model if our goal is to estimate ability to learn.

Another approach to estimating student ability to learn is to characterize "learning curves" [7]. This requires repeated sampling over time so the learning rate can be determined from the shape of the curve, where students with higher ability to learn are characterized by steeper learning curves, and students with lower ability to learn are characterized by shallower curves. However, the shape of a learning curve does not reveal the full interaction between prior and posterior knowledge. We would expect two students with the same ability to learn but different levels of prior knowledge to progress at significantly different rates. Additionally, collecting enough data to plot a learning curve requires repeated measurement that is infeasible in most educational settings.

NWEA has looked into how to quantify learning gains [12, 11] and most recently [21]. Their contemporary models project student abilities into norm grade levels. [16, 8]. Anderman et al make initial steps into translating learning-gain research into a bayesian model [1]

Significant research has focused on the promise and perils of using student gain data as an outcome—as a good indicator of teacher effectiveness. There is a book on the subject of evaluating teachers by measuring their value added: Evaluating Value-Added Models for Teacher Accountability [10]. We remind the reader that it is necessary to be careful and accurate in measuring student learning.

There is a rich mathematical history of reasoning about functional mappings. This field of mathematics draws from domains as diverse as 3D geometry [14, 3] to neocortical circuitry [20]. This is, to the best of our knowledge, the first use of functional maps in measuring learning.

1.3 Learning Gain Functions

In traditional IRT, each interaction between a student and a question (aka item) produces a single number. In a learning test, each learning-experience produces two numbers ($\alpha_{i,j}$

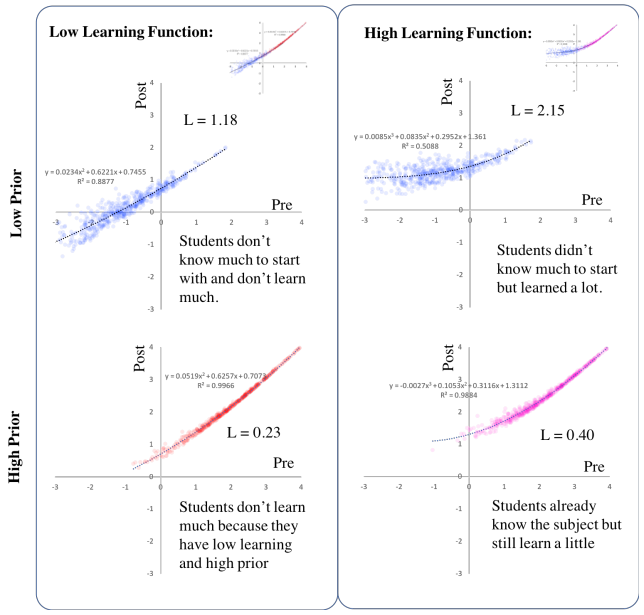


Figure 1: Simulations of four populations on the same task. Each graph represents pre/post abilities for one population. Each point represents one student. Countries in the columns have the same learning function. L = μ post minus pre.

and $\beta_{i,j}$). This poses a modelling challenge. How do we model learning, in a way that elegantly considers the effect of prior knowledge ($\alpha_{i,j}$)?

We found it natural to resolve this problem by thinking of the learning experience as being a reflection of an underlying *function* which we call learning-gain-function. A learning-gain-function is a population wide mapping of student pre-conditions to post-conditions. In a learning-ability-test, we would like to compare populations on their ability to learn, and as such it seems like it would be best to compare the countries by their functional mapping. Thinking of the mapping of prior-ability to post-ability is incredibly useful if we want to build a Bayesian model of learning-ability.

To articulate this point, consider the four different populations learning on the exact same learning task (Figure 1). For all four populations we plot prior abilities and posterior abilities. The two populations on the left both have the same learning "function" on this task. If you had two students with the same prior ability, after the learning-task they would have (within noise) the same post ability. As a confound, they have different prior ability distributions. Typical measures of learning gains would compare these two populations based on the average difference in post ability versus pre ability (L shown on the figure). *As such they would look very different even though the two populations have the same learning-gain-function.* The same is true for the two populations on the right. They also have the exact same learning-gain-function, but as a result of different prior knowledge distributions, typical metrics make them seem quite different. By modelling a learning-gain-function we neither benefit, nor penalize populations for having different prior distributions. Instead we compare learning in a way that is agnostic to previous knowledge.

The learning function f is "parameterized" by the ability-to-learn parameter θ and task specific parameters, ϕ:

$$f_{\theta_i,\phi_j}(\alpha_{i,j}) \rightarrow \beta_{i,j}$$

In the case of PISA, this theta should represent "ability to learn" for a specific population. The function, importantly, does not have to be linear – and in fact ample evidence shows that it should not be. Note that, $\alpha_{i,j}$ and $\beta_{i,j}$ can be estimated using standard item-response theory.

This formalization lends some insight into how we can deal with the different levels of prior knowledge between populations. At this point we haven't made any claims about what the function looks like. What is an appropriate parametric form of a learning-gain function?

2. SIMULATING LEARNING

To begin the process of understanding the family of functions for how much students learn during a task, we built a series of simulators in python that attempt to match as realistically as possible the process of learning during a task. The simulator has fake students learn through the process of working on fake items, where the learning and progress at each minute is governed by the interaction between a student's prior knowledge and the difficulty of the items (an assumption loosely based on the zone of proximal development). This simulation is not perfect, but it provides us with a starting point for building a theory of ability to learn. It is simple, and makes it possible to observe all the factors that impact changes in knowledge, including variables which are often unobservable like learning ability.

These simulations have the added benefit of building a falsifiable condition for any model which tries to estimate ability to learn. I.e., any good model should be able to describe this synthetic data. While ability to describe synthetic data is evidence in support of a theory, it is a necessary but not sufficient condition. The final test would be to show that it also works with real world data.

Figure 1 shows a simulation of 2,000 students learning in four countries via a single task which is heavily biased towards "beginners learn more". The countries in the right column both have the exact same learning function, but because their students have different priors, they are very hard to distinguish that they have the same learning ability.

The main take away at this point is to confirm what we believed from prior work: average "ability" gain is not a very useful metric. Even for countries with the exact same learning rate we observe very different average gains (L) when priors are different.

3. THEORY OF LEARNING FUNCTIONS

If we could come up with an equation for that function (aka the form) we could formalize our measurement of ability-to-learn. In the example from Figure 1 it feels like a "polynomial" fits f well – but that turns out to be a bit misleading. The learning-experience in that figure represents one where "beginners learn more." If we change the learning task to be one where "medium level students learn more" the function is not well fit by a polynomial.

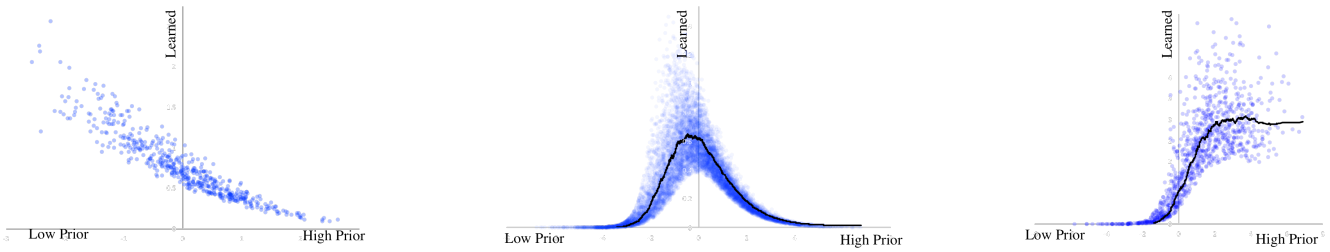


Figure 2: The same population on three different tasks. Left: a task on which beginner students learn more. Middle: a task on which medium students learn more. Right: top students learn more. Points are simulations pre-vs-learn of different students.

If we revisit our simulations and consider "pre vs learn-delta" as opposed to "pre vs post" we can gain insight into the functional form. Here we define learn-delta to be an individuals improvement from post to pre on the learning-task ($\beta_j - \alpha_j$). Figure 2 shows the "pre vs learn-delta" for three different tasks, produced by the simulation, one where beginners learn more, one where medium students learn more, and one where students with advanced prior knowledge learn more.

The graph in the middle (medium students learn more) resembles a "Gaussian" bump, whereas the graphs on the left and right look like exponential decay and growth, respectively. However, upon further inspection, we note that all of the graphs can be represented by an equation with a Gaussian bump. The exponential graphs could be considered to be left and right legs of a bump.

In order to build a functional form that matches all three scenarios (while appreciating that the "beginners learn more is much more common") we propose a simple parametric form which can describe all three, the learning-gain function family:

Learning-Gain-Bump Family: The function of a student i from population j with learning ability θ, learning on task k with parameters ϕ is:

$$f_{\theta,\phi}(\alpha) = \alpha + \theta \cdot e^{-\frac{(\alpha-\phi_1)^2}{\phi_2}} = \beta \qquad (1)$$

Where:
α is prior ability of student i on task k
β is posterior ability of student i on task k
θ is "ability to learn" of population j
ϕ is a vector of two task k specific constants.

In this model larger values of learning-ability (θ) scale up the Gaussian shaped bump.

We note that in practice most learning experiences tend to have the property that "beginners learn more" and as such an exponential decay function should often work well in practice. As such we also consider the Learning-Gain-Decay Family: $f_{\theta,\phi}(\alpha) = \alpha + (\alpha + \phi_1)^{-1} + \phi_2$

Inference is performed using a PyTorch implementation of the model, and Adam optimization to minimize the Mean Squared Error in predicting posterior (β) abilities.

4. EVALUATION

While the Learning-Gain function family seems reasonable as a hypothesis. In order to test its utility as a basis for item response theory on learning-ability, we evaluate on both simulated data with known learning-abilities and real-world data.

4.1 Simulated Evaluation

To evaluate we generated two tasks, and for each task simulated 2000 students from eight countries with a range of parameters: most importantly a single parameter which represented the latent ability to learn of a student from that population.

To evaluate, we build an inference algorithm to take the observed data produced by the simulations (the pre/post abilities of each student) and attempt to infer single value θ_j for each population j using the generative model in Equation 1. Recall that the simulations are **not** generated from our assumed function, rather it is a product of a zone-of-proximal development rather-complex simulation.

The Bayesian model, which estimates learning-ability via learning-gain-functions, is able to perfectly back-out "population ability to learn" from such simulated data (For both tasks with eight countries, $R^2 > 0.99$). In contrast a linear function was not able to fit the data nearly as well. For the task that was good for beginners it performed reasonable ($R^2 = 0.92$) whereas for the task that was good for medium prior knowledge the model was predictably unable to fit the data ($R^2 = 0.81$). While this is impressive result especially considering the complexity of the simulation, in order to consider this model useful we would like it to be able to make predictions on real-world data.

4.2 Evaluation on Real-World Data

We trained the Learning-Gain-Fn model on two real-world datasets: NWEA and ECDL. The NWEA dataset contains 69612 students from 330 schools in Grade 7 whose reading level was assessed twice (pre test and post test) using item-response theory, once in Winter and again in Spring 2017. The ECDL dataset contains data from 379 undergraduate students at the University of Alcalá (Spain) [4]. Scores for each student include four pretests and four posttests corresponding to distinct learning modules.

We compared the Learning-Gain-Fn model to a number of other plausible models: a linear model, a second-order polynomial, an exponential-decay model, and a linear model

Table 1: Results on Real-World Data

Model	Parameters per Population	Formula	NWEA Test-Set MSE	ECDL Test-Set MSE
Linear	1	$\Delta_i = \alpha_i\phi_1 + \theta_j$	56.9	0.45
Polynomial	1	$\Delta_i = \phi_1\alpha_i^2 + \phi_2\alpha_i + \theta_j$	56.0	0.47
Learning-Gain-Decay	1	$\Delta_i = \theta_j(\alpha_i + \phi_1)^{-1} + \phi_2$	54.9	**0.44**
Learning-Gain-Bump	1	$\Delta_i = \theta_j \cdot e^{-\frac{(\alpha_i-\phi_1)^2}{\phi_2}}$	**53.1**	**0.44**
Linear Multi-Theta	2	$\Delta_i = \alpha_i\theta_{j_1} + \theta_{j_2}$	55.1	**0.44**

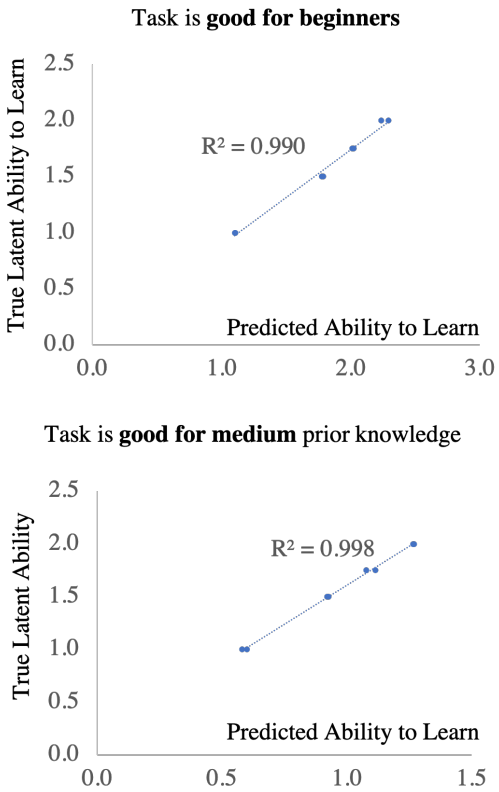


Figure 3: The simple model proposed by the Learning Gain Bump Family allows for very accurate prediction of latent ability to learn from the complex simulation.

with two parameters per population. (See Table 1 for model details.) Our primary goal was to identify a model that could best capture "ability to learn" in a *single parameter* across a variety of populations and testing scenarios. Estimation with a single parameter is important because it is more-easily interpreted—a higher value corresponds to a higher ability to learn. Each of the models we evaluated estimates "ability to learn" with a single parameter where higher values correspond to better learning ability. The exception to this rule is the Linear Multi-Theta model, which estimates ability to learn using two parameters. (See Related Work for an explanation of why this model was included.)

To compare models, we held out 10% of the data from each dataset and computed the mean-squared error when different models made predictions about the missing data.

Notably, the Learning-Gain-Bump model outperforms all models on predicting held-out data, including the Linear Multi-Theta model. Full results are reported in Table 1. This suggests that "ability to learn" in these two cases followed a parametric form best explained by a more nuanced learning-gain function. While the gains in MSE are modest, we hypothesize that for some datasets, especially ones where the learning tasks most benefit medium strength students, the linear model will break down. We also note that the Learning-Gain-Decay and the Learning-Gain-Bump function performed very similarly – which indicates that all the tasks in this data were ones where ones where beginners learned the most.

Figure 4 shows the shape of the learning-gain-fn for different grade levels in the NWEA dataset between Winter and Spring. For every one of the 330 schools in the dataset we can now compute the ability-to-learn (θ) of the students in their population. We note that, as shown in Figure 4(b), the distribution of θs appears to be Gaussian. Figure 4(a) also includes the learning-gain-fn for two of the top schools in the NWEA dataset. We note that it is impressive how much of ability-to-learn can be explained by which school a student went to. In the top schools (by learning-ability) students with low, medium and high prior ability substantially improve between the pre and post test.

These results are preliminary. The robust model of ability-to-learn presented in this paper will open up deeper analysis into learning in a wide range of contexts: from short tests of learning ability to evaluations of ability-to-learn in schools.

5. LIMITATIONS

Ability to learn is unlikely to be a single parameter: It is highly unlikely that a student's ability to learn can be captured in a single number. However, this simplifying assumption proves to be convenient and useful. Often, the amount of data available to estimate parameters is small, making a model with few parameters attractive. Additionally, estimating ability to learn with a single parameter results in a model that is maximally-interpretable—the higher the number, the better the ability to learn.

There is a three-month gap between testing periods in the NWEA data: Our hypotheses about student ability to learn are based on a simulation of student learning that occurs over the course of a day. In testing these hypotheses we relied on real-world datasets that measured learning over

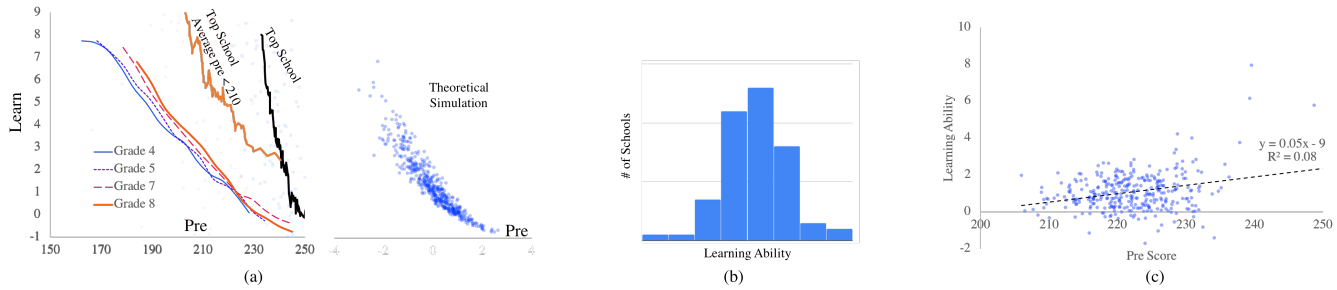


Figure 4: (a) NWEA pre-vs-learn graphs for different grade levels. Note that the distribution matches the theoretical simulated learning-task. The graph includes the pre-vs-learn for the school with the highest θ and the school with the highest θ for students with low prior ability. (b) shows the full histogram of learning abilities for different schools. (c) shows the relationship between student abilities and ability-to-learn

significantly-longer time period. For example, the two tests in the NWEA dataset that we used to measure ability to learn occurred approximately three months apart. Despite this fact, our model still fit the real-world data better than any alternative model, providing a measure of reassurance.

6. DISCUSSION

Modelling individual students' ability to learn: The models in this paper estimate ability to learn at the group level. However, there are many cases where estimating individual students' abililty to learn would be useful as well. Due to the small number of datapoints per student, this could prove challenging. However, a hierarchical model that assumed individual students' abilities to learn were governed by a strong group-level prior could overcome this problem.

Incorporating Pre/Post Tests: Given the function, we can incorporate this ability to learn into the traditional IRT process before and after. Specifically, the probability that a student i gets an item k right on the pre test should be, under the IRT-2PL: $p_{ik} = \sigma(\alpha_i - d_k)$ where d_k is the difficulty of item k and α_i is the same $alpha_i$ that we used in our learning model. σ is the sigmoid function. Similarly the probability that student i gets a item k correct on the post test would be: $p_{ik} = \sigma(\beta_i - d_k)$ where β_i is the posterior ability of the student after the learning task. In the case where pre-post tests are real valued, we can use the logit-normal IRT proposed by Arthurs et al [2].

Fairness and Mixture Models: A Bayesian model of learning-gain-functions can do much more than simply infer ability-to-learn from pre-post tests. It would also allow for researchers to disentangle mixture distributions. This would allow researchers to identify sub-population effects within a larger population. Similarly, a robust model of ability-to-learn can be the basis of ensuring that a learning-task, and/or an education system is fair to different demographics.

Learning The Learning-Gain Function: In this paper we have modeled ability to learn as a parameter in a family of learning functions. This family of functions is Gaussian-like, a choice that was informed by observing the outcomes of a theoretically-grounded simulation. While this choice proved to have the lowest error, it is likely that another choice could offer improvements. Rather than trying a number of models, each with its own assumptions, an alternative approach would be to use a small neural network to learn the model directly from the data.

Neural networks are universal function approximators, which means a small neural network should be able to learn the function family that serves as the best model that incorporates θ, ϕ, and α. Fears that neural networks are black-box algorithms that lack interpretability do not apply in this case—since the number of parameters is small, the learned function can be visualized directly across all values of the parameters. This approach would combine the flexibility of neural networks with the transparency and interpretability of the current models.

7. CONCLUSION

"Learning how to learn" is considered an essential skill for the 21st century [23]. Given the rapid pace of technological development, this is one of the most valuable skills an educational system can provide for its students. In recognition of this fact, the PISA 2024 test will contain an experimental section that has been explicitly designed to measure students' ability to learn. However, few assessments have been explicitly designed to gauge this abililty, meaning that the community lacks models that are capapble of directly estimating this skill. In this paper we introduce a model that estimates student ability to learn using a single parameter. This model is more accurate at estimating student change in knowledge than other competing single- and multi-parameter models on two real-world datasets. Additionally, it is able to perfectly recover "ability to learn" from a complex, theoretically-grounded simulation of student learning over time. We present this work to demonstrate the value in explicitly modeling this skill, and we propose this model as a first step towards a more complete theory of understanding ability to learn.

8. REFERENCES

[1] E. M. Anderman, B. Gimbert, A. A. O'Connell, and L. Riegel. Approaches to academic growth assessment. *British Journal of Educational Psychology*, 85(2):138–153, 2015.

[2] N. Arthurs, B. Stenhaug, S. Karayev, and C. Piech. Grades are not normal: Improving exam score models using the logit-normal distribution. In M. C. Desmarais,

C. F. Lynch, A. Merceron, and R. Nkambou, editors, *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019*. International Educational Data Mining Society (IEDMS), 2019.

[3] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[4] L. de Marcos, E. García-López, and A. García-Cabot. Dataset on the learning performance of ecdl digital skills of undergraduate students for comparing educational gaming, gamification and social networking. *Data in brief*, 11:155–158, 2017.

[5] R. Glaser and A. J. Nitko. Measurement in learning and instruction. 1970.

[6] L. J. Hendrix, M. W. Carter, and J. L. Hintze. A comparison of five statistical methods for analyzing pretest-posttest designs. *The Journal of Experimental Education*, 47(2):96–102, 1978.

[7] B. Jovanovic and Y. Nyarko. A bayesian learning model fitted to a variety of empirical learning curves. *Brookings Papers on Economic Activity. Microeconomics*, 1995:247–305, 1995.

[8] N. M. Laird and T. A. Louis. Empirical bayes ranking methods. *Journal of Educational Statistics*, 14(1):29–46, 1989.

[9] Y.-J. Lee, D. J. Palazzo, R. Warnakulasooriya, and D. E. Pritchard. Measuring student learning with item response theory. *Physical Review Special Topics-Physics Education Research*, 4(1):010102, 2008.

[10] D. F. McCaffrey, J. Lockwood, D. M. Koretz, and L. S. Hamilton. *Evaluating Value-Added Models for Teacher Accountability. Monograph.* ERIC, 2003.

[11] M. S. McCall, C. Hauser, J. Cronin, G. G. Kingsbury, and R. Houser. Achievement gaps: An examination of differences in student achievement and growth. the full report. *Northwest Evaluation Association*, 2006.

[12] M. S. McCall, G. G. Kingsbury, and A. Olson. Individual growth and school success. a technical report from the nwea growth research database. *Northwest Evaluation Association*, 2004.

[13] J. P. Meyer and S. Zhu. Fair and equitable measurement of student learning in moocs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment*, 8:26–39, 2013.

[14] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (TOG)*, 31(4):1–11, 2012.

[15] M. Ramscar, P. Hendrix, C. Shaoul, P. Milin, and H. Baayen. The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, 6(1):5–42, 2014.

[16] D. D. Ready. Associations between student achievement and student learning: Implications for value-added school accountability models. *Educational Policy*, 27(1):92–120, 2013.

[17] D. L. Schwartz, J. D. Bransford, D. Sears, et al. Efficiency and innovation in transfer. pages 1–51.

[18] D. L. Schwartz, R. Lindgren, and S. Lewis. Constructivism in an age of non-constructivist assessments. pages 34–61.

[19] D. L. Schwartz and T. Martin. Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. 22(2):129–184.

[20] A. M. Thomson and C. Lamy. Functional maps of neocortical local circuitry. *Frontiers in neuroscience*, 1:2, 2007.

[21] Y. M. Thum and C. H. Hauser. Nwea 2015 map norms for student and school achievement status and growth. *Portland, OR: NWEA*, 2015.

[22] E. Weber. Quantifying student learning: how to analyze assessment data. *The Bulletin of the Ecological Society of America*, 90(4):501–511, 2009.

[23] C. E. Weinstein. Learning how to learn: An essential skill for the 21st century. *Educational Record*, 66(4):49–52, 1996.