Replicating an "Astonishing Regularity in Student Learning Rate"

Mary Ann Simpson, Kole A. Norberg, Stephen E. Fancsali
Carnegie Learning, Inc
msimpson@carnegielearning.com

ABSTRACT

In a broad analysis of a large, diverse sample of students, we found robust support for the groundbreaking assertion that student learning rates in various educational technologies are "astonishingly" similar (Koedinger, Carvalho, Liu, & McLaughlin, 2023, "An astonishing regularity in student learning rate," Proceedings of the National Academy of Sciences). Koedinger et al. (2023) initial finding challenges long-standing beliefs about variability in students' learning rates. It suggests that the main requirement for learning within educational technologies is additional opportunities. The strength of this claim and its implications warrant deeper investigation. Here we replicate those original results leveraging much larger data sets with, collectively, over 15,000 students and 821,890 observations across 6 math topics. These data from the MATHia intelligent tutoring system reflect a diverse student user population learning in a "business as usual" context. Finally, we provide additional evidence in the form of confidence intervals around the variance to support this claim.

Keywords

learning rate, learning curves, growth modeling, logistic regression, individual difference.

1. INTRODUCTION

Recent advances in educational technology and modeling of individual differences [14] have invited renewed investigation into a long-standing debate regarding whether students show variability in their learning rates [17]. In response, Koedinger et al. [11] analyzed student learning rates across a variety of educational technology platforms and found an "astonishing regularity" in student learning rates across 27 research datasets gathered for cognitive modeling over the past several decades. They concluded that when students are provided with quality opportunities to learn (OTL), they learn at similar rates across each opportunity [11]. In consideration of the importance of replication for substantiating novel claims [9], we replicate and extend these findings with larger more recent samples, adding reliability and robustness to the initial claim.

Some early theories into the question of individual differences in learning rates suggested that given an ideal learning environment, students would learn at the same rate [4]. However, initial analytical investigation found substantial differences [1, 2] and subsequent work has focused on identifying the nature of those differences rather than determining the extent of the variability [3, 7, 9, 13, 17, 22, 24, 25]. The findings of Koedinger et al. [11] invite a reframing of this prior work and the role of individual differences.

1.1 Measuring Learning

Additive Factors models (AFMs) quantify the probability of a correct response in a task as a function of the additive contributions of the component cognitive skills [5]. Equation 1 gives the essential formulation of the AFM.

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i + \Sigma_{k=1}^{K} q_{jk}\beta_k + \Sigma_{k=1}^{K} q_{jk}\gamma_k T_{ik}, \qquad (1)$$

where θ_i is the student ability, q_{ik} the indicator flag for each skill, β_k the skill difficulty, γ_k the skill's increment or decrement to an overall learning rate, and T_{ik} the number of practice opportunities the student has had so far to learn the skill.

AFMs can be expressed either as fixed-effect models or as multilevel models [6]. AFMs are specified in their logistic regression form, but are isomorphs for compensatory, multi-skill IRT models [7, 11]. Initially, the multilevel AFMs were intended to evaluate task models and included random slopes for skills but not for students [5]. In recent years, researchers have expanded the model to include a fixed, overall learning rate, as well as random slopes for students to quantify individual differences in learning (See Equation 2). This modification creates bona fide growth models, which have the benefit of simultaneously measuring growth and initial status while avoiding the pitfalls of using gain scores [20]. This revised model is called the individualized-slope Additive Factors Model (iAFM) [14].

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta + \theta_i + \Sigma_{k=1}^{K} q_{jk}\beta_k + \Sigma_{k=1}^{K} q_{jk}\left(\delta + \delta_i + \gamma_k\right) T_{ik}, \qquad (2)$$

where θ is the overall intercept, the average initial ability, θ_i

M. A. Simpson, K. A. Norberg, and S. E. Fancsali. Replicating an "astonishing regularity in student learning rate". In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 420–425, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729850

Table 1: Datasets selected from Koedinger et al. [2023]

DS	Topic	*N* Records	*N* Students	*N* Skills
99	High School Geometry	17,419	51	39
104	College Physics	6,024	104	12
253	Geometry, High School Geometry, Area	14,875	41	22
392	Middle School Geometry, Area	41,602	123	38
1007	College Computer Science	5,063	49	4

Note. Numbers in the DS column refer to dataset identifiers in LearnSphere DataShop, available at `http://pslcdatashop.web.cmu.edu`[10].

the student initial ability, q_{ik} the indicator flag for each skill, β_k the skill intercept, i.e., difficulty, δ the overall growth or learning rate, δ_i the student's increment or decrement to that rate, and γ_k, the skill's increment or decrement to that rate, and T_{ik} the number of practice opportunities the student has had so far to learn the skill. The ln of the odds for p_{ij} transforms results to the logit scale, or log odds scale an interpretable, continuous score scale.

1.2 Present Study

Given the long history of finding variance in learning gains and rates and evidence that these are attributable to individual differences [1, 2, 3, 8, 9, 13, 17, 22, 24, 25] confidence in Koedinger et al's findings will require additional replication and investigation into the models. Therefore, we set out to replicate Koedinger et al.'s 2023 study on datasets from Carnegie Learning's MATHia intelligent tutor system (ITS; formerly Cognitive Tutor) [18]. These data offered advantages of greatly increased sample size, increased diversity among students (i.e., an approximately nationally representative sample), more observations per student, operational (i.e., "live") context, and recent provenance, i.e., 2022-2023 school year.

2. STUDY 1: DIRECT REPLICATION

We first sought to verify the implementation of our models by using them to replicate the findings of Koedinger et al. [11]. We then extended their analysis to determine how skill model choice affected the results. See Table 1 for details about the data sets. Following the same exclusion criteria (i.e., dropping all skills for which there are not 10 students with more than 3 observations each), we recovered the same interquartile range (i.e., middle 50%, IQR) reported in the original paper to the thousandth decimal place. We additionally report two new statistics which further support the robustness of the original findings for these datasets. First, the variance of the random effects was small suggesting low variation in student slopes even among students outside the IQR. Second, we found the IQR for all candidate skill models and report small deviations in IQR across models within a DS. Maximum IQR found across all five replicated data sets here did not meaningfully exceed those reported in the paper. The low range in IQR across models demonstrates that choice of skill model would not have affected the interpretation of the Koedinger et al. findings. See Table 2 for statistics.

Table 2: Directed Replication Statistics

DS	Original IQR	Replicated IQR	Student Slope Variance	Range in IQR across all Skill Models
99	0.012	0.012	0.00015	(0.002, 0.017)
104	0.012	0.012	0.00034	(0.010, 0.057)
253	0.015	0.015	0.00020	(0.009, 0.016)
392	0.013	0.0137	0.00042	(0.006, 0.015)
1007	0.047	0.047	0.00200	(0.010, 0.105)

Note. Original IQR reflects the IQRs reported by Koedinger et al. [11]. Student slope variance reflects the un-shrunken variance.

3. STUDY 2: REPLICATION ON MATHIA

Study 2 extends the findings from Koedinger et al. [11] by applying the same model to larger data sets from a more diverse pool of participants who are engaging with an ITS (MATHia) in a "business as usual" context. MATHia's student population is broadly representative of the US K-12 population.

3.1 Data

Domains were chosen to closely align with the area and geometry topics covered in DS99, DS253, and DS392. Six MATHia topics or workspaces were identified. We considered student event or action-level data (sometimes referred to as "records") from the 2022-2023 school year. Records were limited to first attempts at steps within each problem. Keeping with Koedinger et al.'s practice, skills with fewer than 10 students and students with fewer than 2 records were removed [11]. As in Koedinger et al., data were converted to the student-step-rollup format, where multiple skills corresponding to the same student step are concatenated into a single, combined skill. Tables A1 in Appendix A shows the total number of records, students, and concatenated skills eligible for use in the study as well as the number of records available for each student. Data management was conducted with scala version 2.12, python version 3.12, and pandas, version 1.4 [21, 23, 16].

The initial data sets were too large to perform computational modeling in a convenient time frame. Therefore, we sampled 3,000 students per workspace at random and included all eligible records for those students. One workspace had only 570 students, all of whom were included in the model. No skills were lost in sampling, and students had a large number of records for growth modeling, $m = 55$ records, $range =$ (24.59, 91.02), (see Table 3).

3.2 Results

Separate models were fit for each MATHia workspace. Table A2 in Appendix A shows fit statistics for each workspace. We adopted a liberal rule of requiring at least 10 skills for reporting skill-level random effects. Rules of thumb for such inclusion vary greatly, e.g., from 5, 30, to 50 units. Consequently, skill-level random effects were not reported for the Determining Parts of Quadrilaterals and Parallelograms workspace given its small number of skills ($n = 5$). Table

Table 3: Number of Records, Students, and Skills for Sampled Data by MATHia Workspace

Workspace	*N* Records	*N* Students	*N* Skills	Mean Records per Student
W1. Using Measures of Circles	39,948	570	57	70.08 (14.96)
W2. Calculating Area of Composite Figures	273,048	3,000	12	91.02 (43.47)
W3. Calculating Circumference and Area of Circles	145,145	3,000	22	48.38 (15.27)
W4. Calculating Area of Various Figures	179,362	3,000	12	59.79 (17.21)
W5. Calculating Area of Rectangles	110,955	3,000	12	36.99 (16.07)
W6. Determining Parts of Quadrilaterals & Parallelograms	73,762	3,000	5	24.59 (5.68)

Note. Parentheses reflect standard deviation from the mean.

4 shows the estimates for fixed effects for each workspace and their associated Wald 95-percent confidence intervals. All fixed slopes were positive, which suggests that students were learning. The confidence intervals for the slopes were reasonably small, which suggests their estimation was good. The confidence intervals for the overall intercepts were surprisingly wide, which suggests this effect was poorly estimated.

Table 4: Fixed Effects Coefficients with 95-Percent Confidence Intervals

Workspace	Effect	Coef	LL.025	UL.975
W1	Intercept	1.15	0.88	1.41
	Slope	0.48	0.40	0.55
W2	Intercept	0.09	-0.35	0.53
	Slope	0.16	0.09	0.22
W3	Intercept	0.81	0.48	1.15
	Slope	0.45	0.36	0.53
W4	Intercept	-0.31	-1.02	0.40
	Slope	0.30	0.22	0.38
W5	Intercept	-0.13	-0.62	0.35
	Slope	0.39	0.33	0.46
W6	Intercept	1.69	0.65	2.73
	Slope	0.45	0.34	0.55

Regarding uncertainty in parameter estimation, we found that confidence intervals for the student intercepts, student slopes, and fixed slopes in the CL data were very small (max range was .18 logits for student intercepts and student slopes .0013 and skill slopes .0306, fixed slopes .08 logits). This finding suggests the iAFM model provides precise estimation for most of its random effects and fixed slopes.

Estimation was much less precise for the skill intercepts (min range = .55 logits). This is most likely due to the relatively small number of skills in each dataset. It is unclear why the skill slopes appear unaffected (Table A3 in Appendix A).

A problem emerged with the estimation of the fixed intercepts. The confidence intervals were very wide for several workspaces (max range = 1.04 logits). It's puzzling that the iAFM model seemed to have trouble estimating this effect. Typically, fixed effects are easily estimated and with much less data required than for random effects [12, 15]. However, some tinkering with the multilevel model for the Calculating Circumference and Area of Circles workspace revealed that the presence of skill slopes in the model led to the large uncertainty in estimates for the fixed intercept. Low sample size does not appear to be much involved, because this workspace has a respectable number of units, i.e., 22 skills. We are still investigating why including random skill-level slopes in the model is having this effect on estimation.

Table 5 shows the IQRs and their bootstrapped 95-percent confidence intervals. As in Koedinger at al. [11], these IQRs are quite small. The largest student slope IQR was .116 for Properties of Parallelograms. However, even this IQR was only slightly larger than the largest IQR that Koedinger et al. identified. Their largest student slope IQR was .102, for a dataset covering acquisition of Chinese vocabulary. Table 6 shows the bootstrapped means, empirical standard errors, and 95-percent confidence intervals for the shrunken variance components of the student random effects for each workspace's model.

Table 5: Student-Level IQRs with 95-Percent Bootstrapped Confidence Intervals

Workspace	Effect	Mean IQR	LL.025	UL.975
W1	Intercepts	1.0271	0.9332	1.1308
	Slopes	**0.0355**	**0.0305**	**0.041**
W2	Intercepts	1.0698	1.0255	1.1129
	Slopes	**0.0056**	**0.0054**	**0.0058**
W3	Intercepts	1.3238	1.2618	1.3784
	Slopes	**0.0666**	**0.0627**	**0.0705**
W4	Intercepts	0.8525	0.8174	0.8908
	Slopes	**0.0136**	**0.013**	**0.0142**
W5	Intercepts	1.0568	1.0156	1.1042
	Slopes	**0.0797**	**0.0739**	**0.0850**
W6	Intercepts	0.7474	0.7034	0.8021
	Slopes	**0.1182**	**0.1129**	**0.1246**

As with the IQRs, the student slope variances were quite small. This is especially apparent in comparison to the variance components for skill-level random slopes, which were much larger than the student slopes (Table A3, Appendix A). The confidence intervals for most random effects were reasonably small, which suggests good estimation. Importantly, the student intercept variances had a much wider range suggesting students did not start each workspace with similar levels of content knowledge.

4. GENERAL DISCUSSION

4.1 Summary and Conclusions

Table 6: Student-Level Shrunken Variance Components with 95-Percent Bootstrapped Confidence Intervals by MATHia

Workspace	Effect	Mean IQR	LL.025	UL.975
W1	Intercepts	0.6503	0.5656	0.7388
	Slopes	**0.0015**	**0.0012**	**0.0017**
W2	Intercepts	0.5492	0.5233	0.5763
	Slopes	**<.0001**	**<.0001**	**<.0001**
W3	Intercepts	0.8102	0.7699	0.8507
	Slopes	**0.0035**	**0.0033**	**0.0037**
W4	Intercepts	0.3809	0.3604	0.4029
	Slopes	**0.0001**	**0.0001**	**0.0001**
W5	Intercepts	0.6047	0.5761	0.6349
	Slopes	**0.0044**	**0.0042**	**0.0047**
W6	Intercepts	0.321	0.3013	0.3435
	Slopes	**0.0087**	**0.0081**	**0.0094**

The most important “take-aways” from this project are

- The extremely small student learning rate variance that Koedinger et al. [11] identified persists in similar modeling in similar mathematical topics on much larger datasets involving more diverse student populations. This finding held whether the spread in student learning rates was measured by the IQR or by variance components.
- This finding supports the idea that EdTech products reduce inequity with the additional OTLs and scaffolding they offer [11]. Earlier research into decreased variance in student learning rates in mastery learning revealed a “mixed” picture [1]. Today’s researchers might find wider support for the equity hypothesis.

4.2 Additional Findings

Our research produced incidental findings likely of interest to the learning science communities. The positive overall slopes from the MATHia models’ fixed effects (Table 4) indicate that students are learning. Uncertainty for most parameter estimates was acceptably small except for the fixed intercepts and random skill intercepts.

4.3 Skill Model Selection in Koedinger et al.’s Data

In reanalysis of Koedinger et al.’s [11] data, we found that selection of the skill model had little effect on variance suggesting that the finding concerning student slope variance is robust and not sensitive to other parameters in the model.

4.4 Implications for Teaching and Learning

Education has long been concerned with the Matthew effect, a finding that students who start with more prior knowledge or better resources not only remain ahead but learn at faster rates exacerbating knowledge gaps [24]. The results here offer hope for reframing this effect. Students who start further behind can learn at the same rate if provided with quality OTLs, but in order to catch up to their peers, these students will need to be provided with more opportunities than their peers. Under this paradigm, ensuring students are provided with the opportunities they need becomes critical. Indeed, one reason these results may have been detected among students using EdTech tools is that these tools often provide students with extra OTLs, hints, instructional materials, adaptive learning, and other supports considered foundational to learning [4, 19]. Thus, these results highlight the importance of equity in access to opportunities and the critical role this plays in learning.

4.5 Limitations

Users should remember that the iAFM is a very new model and will require subsequent evaluations before researchers should unqualifiedly accept its estimates. The large uncertainty in estimation around the fixed intercepts and skill slopes suggests that all is not well with the iAFM applied to the MATHia data. Researchers have, however, already ruled out a substantial body of potential artifacts, e.g., possible over-shrinkage of slope variance components and possible confounding of skill and student variances[11]. Secondly, the research covered a only handful of topics in mathematics. Researchers should extend this modeling to more topics and varieties of assessment and pedagogical data to learn if the findings generalize.

4.6 Next Steps

We plan to run similar models on MATHia data where we hypothesize either very little student slope variance or large student slope variance and examine the results. For instance, in a topic area where very few students have had pre-exposure, we would expect to find greatly decreased student slope variance. On the other hand, in the MATHia Concept Builder workspaces, where all students have the same number of OTLs, we would expect to find increased student slope variance.

Additionally, in such studies we could examine the relationship between student intercepts and student slopes. Student intercepts represent pre-knowledge and can have an effect on the student slopes.

As part of the learning science community’s continuing efforts to evaluation the iAFM models, we plan to conduct a simulation study and assess bias and spread in parameter recovery.

5. ACKNOWLEDGMENTS

The authors thank Paulo Caravalho for his help in setting up our MATHia data for use with the iAFM model, Ken Koedinger and CMU for access to the Datashop datasets, and Sergey Kirschner for his advice on statistical estimation and the mathematics underling the iAFM model.

6. REFERENCES

[1] M. Arlin. Time, equality, and mastery learning. *Review of Educational Research*, 54(1):65–86, 1984.

[2] M. Arlin. Time variability in mastery learning. *American Educational Research Journal*, 21(1):103–120, 1984.

[3] D. P. Ausubel and D. Fitzgerald. Organizer, general background, and antecedent learning variables in sequential verbal learning. *Journal of educational psychology*, 53(6):243, 1962.

[4] B. S. Bloom. *Human characteristics and school learning.* McGraw-Hill, 1976.

[5] H. Cen, K. Koedinger, and B. Junker. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International conference on intelligent tutoring systems*, pages 164–175. Springer, 2006.

[6] H. Cen, K. Koedinger, and B. Junker. Comparing two irt models for conjunctive skills. In *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings 9*, pages 796–798. Springer, 2008.

[7] P. De Boeck. *Explanatory item response models: A generalized linear and nonlinear approach.* Springer Science & Business Media, 2004.

[8] C. S. Dweck and E. L. Leggett. A social-cognitive approach to motivation and personality. *Psychological review*, 95(2):256, 1988.

[9] Editorial Board. Replicating scientific results is tough — but essential. *Nature*, 600(7889):359–360, Dec 2021.

[10] K. R. Koedinger, R. S. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A data repository for the edm community: The pslc datashop. *Handbook of educational data mining*, 43:43–56, 2010.

[11] K. R. Koedinger, P. F. Carvalho, R. Liu, and E. A. McLaughlin. An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, 120(13):e2221311120, 2023.

[12] W.-C. Lee, B. A. Hanson, and R. L. Brennan. Procedures for computing classification consistency and accuracy indices with multiple categories. act research report series. 2000.

[13] R. Liu and K. R. Koedinger. Variations in learning rate: Student classification based on systematic residual error patterns across practice opportunities. *International educational data mining society*, 2015.

[14] R. Liu and K. R. Koedinger. Towards reliable and valid measurement of individualized student parameters. *International Educational Data Mining Society*, 2017.

[15] C. J. Maas and J. J. Hox. Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58(2):127–137, 2004.

[16] W. McKinney and P. Team. Pandas-powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, 1625, 2015.

[17] National Research Council. *Learning and understanding: Improving advanced study of mathematics and science in US high schools.* National Academies Press, 2002.

[18] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett. Cognitive tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, 14:249–255, 2007.

[19] H. L. Roediger III and J. D. Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3):249–255, 2006.

[20] D. Rogosa, D. Brandt, and M. Zimowski. A growth curve approach to the measurement of change. *Psychological bulletin*, 92(3):726, 1982.

[21] Scala Team. Scala programming language, version 2.12. `https://www.scala-lang.org`, 2023.

[22] S. Tobias. *Overcoming math anxiety.* WW Norton & Company, 1993.

[23] G. van Rossum and P. C. Team. Python language reference, version 3.12. `https://www.python.org`, 2023. Accessed: 2024-05-01.

[24] H. J. Walberg and S.-L. Tsai. Matthew effects in education. *American educational research Journal*, 20(3):359–373, 1983.

[25] C. L. Zerr, J. J. Berg, S. M. Nelson, A. K. Fishell, N. K. Savalia, and K. B. McDermott. Learning efficiency: Identifying individual differences in learning rate and retention in healthy adults. *Psychological science*, 29(9):1436–1450, 2018.

APPENDIX

A. APPENDIX

This appendix provides supplementary tables to support interpreta- tion of the paper's text.

Table A1: Number of Records, Students, and Skills for Source Data by MATHia Workspace

Workspace	*N* Records	*N* Students	*N* Skills	Mean Records per Student
W1	39,948	570	57	70.08 (14.96)
W2	3,947,106	43,589	12	90.55 (43.53)
W3	3,597,403	74,657	22	48.19 (15.19)
W4	3,043,161	50,987	12	59.69 (17.43)
W5	2,293,361	62,658	12	36.60 (15.92)
W6	379,385	15,381	5	24.67 (6.01)

Note: Parentheses reflect standard deviation from the mean.

Table A2: Fit Statistics for GLMM Models by MATHia Workspace

Workspace	BIC	Log Likelihood
W1	28,152.10	-14,033.67
W2	240,696.00	-120,297.93
W3	117,928.44	-58,916.68
W4	112,535.62	-56,219.42
W5	97,074.56	-48,490.81
W6	55,678.93	-27,794.63

Table A3: Skill-Level Shrunken Variance Components with 95-Percent Bootstrapped Confidence Intervals by MATHia Workspace

Workspace	Effect	Mean IQR	LL.025	UL.975
W1	Intercepts	0.7350	0.4819	1.035
	Slopes	0.0298	0.0208	0.0408
W2	Intercepts	0.5983	0.1077	1.3237
	Slopes	0.0122	0.0043	0.0209
W3	Intercepts	0.5347	0.2676	0.8792
	Slopes	0.0303	0.0191	0.0478
W4	Intercepts	1.503	0.2847	3.2994
	Slopes	0.0161	0.0031	0.0337
W5	Intercepts	0.6659	0.2202	1.1660
	Slopes	0.0134	0.0060	0.0218
W6	Intercepts	†	†	†
	Slopes	†	†	†

†*Fewer than 10 skills*