The Hidden Work of Implementing Technology in Education

Elijah Mayfield
Turnitin
Pittsburgh, PA
elijah@turnitin.com

INTRODUCTION

This paper lays out an example of the positive impact that data-driven technology products can have in education. It then highlights the drawbacks of heavy reliance on quantitative measures for predicting outcomes in schools and emphasizes the need for nuanced interpretation, then briefly introduces a framework for better evaluation of impact.

With the American Recovery and Reinvestment Act of 2009, the Department of Education provided $4.35 billion in additional grant funding to state and local school systems through the *Race to the Top* program [5]. Much of this supplemental funding went to educational technology, and led to further public and private investment in the industry. Foundations, venture capital, and private equity financing peaked at over $4 billion in 2015 [19], mostly from a small number of sources like the Bill & Melinda Gates Foundation.

Yet as products proliferate, there is a growing consensus that not all educational technology has positive impacts on the classroom. Technology has been divided by status and access since the early years of introduction into schools [18]. These dividing lines are often tied to race, class, and other non-academic factors [4]. Bill Gates himself stated that "We really haven't changed [students' academic] outcomes" [13]. As a result, 2017 has seen a surge in demand for proof of broad, equitable efficacy in educational technology. We know that today, "although Ed Tech developers value research to inform their products, they are not conducting rigorous research" [8]. This ties into a broader consensus, forming in parallel, that modern algorithmic products are learning from biased or unfair training data [3].

Part of the challenge in evaluating education research is that technology products behave differently in the classroom than they do in lab settings. Implementation challenges can introduce difficulties that are not present in controlled environments. For instance, the widely-used math education product made by Carnegie Learning is well-grounded in learning science theory [1] and shows large gains in student performance in controlled settings [9]. In contrast, school studies frequently find null results of the same product [2], resulting in skeptical, critical reporting [6]. In other cases, whole *categories* of educational technology interventions have been deemed suspect, as in a meta-analysis of "Brain Training" games that were found to have "little evidence that training enhances performance" and describes existing studies as having "major shortcomings in design or analysis that preclude definitive conclusions" [17].

	9th Grade			11th Grade		
School	2016	2017	Δ	2016	2017	Δ
1	22.0	33.7	+11.6	13.8	32.2	+18.4
2	29.4	35.7	+6.2	27.6	44.9	+17.4
3	22.3	21.2	-1.1	15.8	32.1	+16.3
4	43.0	55.7	+12.6	36.6	45.9	+9.3
5	14.8	23.8	+9.0	7.5	15.5	+8.0
6	10.2	26.3	+16.1	16.2	22.4	+6.2
7	2.1	14.3	+12.2	5.0	9.7	+4.7
8	7.3	19.0	+11.7	6.1	8.0	+1.9
9	75.6	88.1	+12.5	86.9	88.3	+1.4
10	92.7	95.5	+2.8	94.8	90.5	-4.3
Average	**25.0**	**34.3**	**+9.3**	**22.1**	**31.1**	**+9.0**
Georgia	*70*	*74*	*+4*	*65*	*70*	*+5*

Table 1. Changes in school English Language Arts passing rates after Revision Assistant intervention year, compared to statewide average.

A CASE STUDY OF SUCCESSFUL ED-TECH

In 2016, Turnitin released *Revision Assistant*, an educational technology product powered by data-driven algorithms. The product supports student writers with automated feedback based on automated essay scoring, a widely adopted machine learning implementation that affects millions of students each year through tests like the GRE and GMAT [14]. The product reproduces expert scoring of student essays reliably [16], generates feedback that improves student work [21], and results in improved outcomes for participating schools [12]. This contrasts with previous studies that have found automated essay scoring products to be useful, but "fallible" [7] and recent studies of similar products that see positive impacts on teachers but no improvement in student outcomes [20].

The outcomes of a district-wide implementation of *Revision Assistant* is profiled in Table 1, similar to the implementation described in [12]. The school district covered students in a medium-sized city in Georgia covering a wide range of socioeconomic status and pre-existing performance. High school students are assessed using the Georgia Milestones Assessment System in 9th and 11th grade English Language Arts; prior to the introduction of *Revision Assistant*, district passing rates in 11th grade ranged from 5% to 95% across ten high schools. The automated feedback product was embedded in writing curriculum across the district along with extensive administrator support, professional development and training for teachers. As a key part of a broader improvement in the curriculum, student performance improved drastically in the 2016-2017 school year, at nearly double the rate for other schools across Georgia.

This is good news for the participating school district. Interpretation of the results, however, shows a nuanced relationship between activity in the product and outcomes. Across the schools, active usage of the educational technology intervention was highly correlated with pre-implementation performance of the schools ($r = 0.70$ and 0.68 for 9th and 11th grade, respectively). High-performing magnet schools made extensive use of the product, while underperforming schools showed less activity. The highly active schools were not the schools that showed the highest growth, however. In fact, there was a slight negative correlation between quantity of usage and growth in performance ($r = -0.29$ and -0.24 for 9th and 11th grade). High usage produced no better gains than merely moderately active schools.

THE HIDDEN IMPACT OF EFFECTIVE IMPLEMENTATION

This evaluation is not a rigorous or complete study of the results in this Georgia county; it serves as an illustrative example. The addition of an algorithmic product to a school curriculum was correlated with improved performance; the highest-performing schools made the heaviest use of that intervention; and yet, a disconnect between heavy usage and the strongest gains was observed in the data.

There are multiple explanations that can be drawn from this. A simple explanation is the impact of ceiling and floor effects on school performance. The most active, heaviest usage schools already exceeded 90% passing rates; as a result, they had less room for growth than underperforming schools. No intervention can improve passing rates beyond 100%. On the other extreme, a more aggressive hypothesis would propose that the purchase of educational technology acts as a forcing function for better collaboration and professional development among teachers and administrators, within and across schools. This improved communication may lead to a more aligned curriculum, better professional development, and better student outcomes, even if the intervention technology itself is ineffective or minimally used.

A more moderated stance may be that careful classroom implementation is a necessary and critical component of educational technology. Professional development, training workshops, close adherence to best practices, and availability of technology are all contributing subcomponents in this work. Lab studies fail to capture the impact of these factors on school performance; similarly, objective optimization functions based only on in-app behavior and engagement fail to capture prominent variables that lead to school success.

IMPLICATIONS

These factors are invisible from in-application statistics. Efforts to scale education without these implementation factors, especially online, have seen completion rates lower than 10% [22]. Learning science research has "troublingly" shown that student engagement with algorithmic products is highest when tasks are at their easiest and most repetitive [11]. And as algorithmic measures of engagement and effectiveness grow in complexity and age, their reliability and ability to be interpreted decays [15]. In general, heavily reliance on quantitative measures of success poses a challenge for designers and developers of educational systems; for researchers evaluating the impact of the products they develop; and for practitioners making informed decisions about what supplemental supports to implement in their classrooms.

In light of this complex interplay between engagement and outcomes, each of these roles may benefit from a mixed-methods approach that ties quantitative evaluation to a richer qualitative understanding of an implementation. Product developers should work to define clearer outcome goals and metrics that better align to learning goals promised by the product; engagement may not be a correlate to learning. Both product vendors and implementing classrooms should ensure that professional development and highly communicative involvement of practitioners is a critical component of a technology adoption. Researchers would benefit from a mixed-methods approach that gives considerable weight to lived experience of practitioners, as is already being emphasized in industry applications of algorithmic products [10]. And purchasing administrators, in turn, will benefit from more nuanced reading of research results that lean less on individual metrics, especially of engagement, and more on the holistic impact of an implementation. For all stakeholders, more work is needed to build a refined understanding of the behaviors within an algorithmic product that lead to transferable outcomes for students and teachers.

REFERENCES

1. Aleven, V., and Koedinger, K. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science 26*, 2 (2002), 147–179.
2. Cabalo, J., Ma, B., and Jaciw, A. Comparative effectiveness of carnegie learning's cognitive tutor bridge to algebra" curriculum: A report of a randomized experiment in the maui school district. research report. *Empirical Education* (2007).
3. Caliskan, A., Bryson, J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (2017), 183–186.
4. Charity-Hudley, A., and Mallinson, C. *Understanding English language variation in US schools*. Teachers College Press, 2015.
5. Duncan, A. Fundamental change: Innovation in america's schools under race to the top.
6. Gabriel, T., and Richtel, M. Inflating the software report card. *The New York Times* (2011).
7. Grimes, D., and Warschauer, M. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment 8*, 6 (2010).
8. Hulleman, C., Burke, R., May, M., Charania, M., and Daniel, D. Merit or marketing?: Evidence and quality of efficacy research in educational technology companies. *White paper produced for the EdTech Academic Efficacy Symposium* (2017).

9. Koedinger, K., Corbett, A., and Ritter, S. Carnegie learnings cognitive tutor: Summary research results. *Carnegie Learning* (2000).

10. Lee, M. K., Kusbit, D., Metsky, E., and Dabbish, L. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM (2015), 1603–1612.

11. Lomas, D., Patel, K., Forlizzi, J. L., and Koedinger, K. R. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 89–98.

12. Mayfield, E., Adamson, D., Woods, B., Miel, S., Butler, S., and Crivelli, J. Beyond automated essay scoring: Forecasting and improving outcomes in middle and high school writing. In *Proceedings of the ACM Conference on Learning Analytics and Knowledge* (2018).

13. Molnar, M. Bill gates: Ed tech has underachieved, but better days are ahead. **https://marketbrief.edweek.org/marketplace-k-12/bill-gates/**. Accessed November 27, 2017.

14. Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., and Bridgeman, B. Evaluation of the e-rater® scoring engine for the gre® issue and argument prompts. *ETS Research Report Series 2012*, 1 (2012).

15. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems* (2015), 2503–2511.

16. Shermis, M. D. State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing 20* (2014), 53–76.

17. Simons, D., Boot, W., Charness, N., Gathercole, S., Chabris, C., Hambrick, D., and Stine-Morrow, E. Do brain-training programs work? *Psychological Science in the Public Interest 17*, 3 (2016), 103–186.

18. Warschauer, M., Knobel, M., and Stone, L. Technology and equity in schooling: Deconstructing the digital divide. *Educational Policy 18*, 4 (2004), 562–588.

19. Watters, A. The business of ed-tech 2016. **http://2016trends.hackeducation.com/business.html**. Accessed November 27, 2017.

20. Wilson, J., and Czik, A. Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education 100* (2016), 94–109.

21. Woods, B., Adamson, D., Miel, S., and Mayfield, E. Formative essay feedback using predictive scoring models. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining* (2017).

22. Yang, D., Sinha, T., Adamson, D., and Rosé, C. P. 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.