Algorithmic Bias in Education and Steps Towards Fairness[1]

Ryan S. Baker

University of Pennsylvania

June 3, 2021

Introduction

It's a great honor to have the opportunity to present in this session, in a conference honoring one of the greatest thinkers in the history of American education. Other speakers in this conference will speak of Professor Gordon's great accomplishments and about the contribution of his ideas. In this talk, I will discuss how some of the challenges he has spoken about through his career, towards establishing justice in education and creating an educational system that effectively educates all children, are appearing a new form in the increasing use of algorithms in K-12 education.

Algorithms – computational models that recognize something about a learner or drive some form of automated adaptation – have become commonplace in the computerized systems that we use in many areas of our lives today. Education is no exception. Algorithms are used in mastery learning, to determine whether a student has learned a key skill before they can move on (Ritter et al., 2016). Algorithms are used to predict which students are likely to drop out of high school, and why (Bowers, Sprott, & Taff, 2012). Algorithms are used to assess student essays and give feedback (Shermis & Burstein, 2013).

The use of these algorithms has led to better outcomes for many students, including results like improved learning outcomes (Ritter et al., 2016) and lower dropout (Milliron, Malcolm, & Kil, 2014).

It's worth discussing where these algorithms come from. In successful cases, they lead to better results, but are they effective for everybody? Not always.

In one particularly notorious case – the 2020 UK GCSE and A-Level Grading Controversy -- a low-quality algorithm was used to replace standardized tests. In that case, teacher predictions were taken and adjusted based on the "quality of the school". The algorithm – which was developed by hand by a few individuals at a regulatory agency -- explicitly lowered the grades for students at lower-income state schools compared to the grades for students at higher-income independent schools (Smith, 2020). After an outcry, the system was abandoned.

This leads to two points. One, there are good ways and bad ways to develop algorithms in education. Discussing how to develop a good algorithm, overall, is outside the scope of this talk, but I'll refer those interested to my free online course, *Big Data and Education* (Baker, 2020). In

[1] This paper represents an edited version of a talk given at the E.W. Gordon Centennial Conference, Teachers College Columbia University (virtually), June 3, 2021

general, though, developing an algorithm well involves collecting significant numbers of examples of what you want to infer, finding or creating a set of meaningful predictors from the data stream, and then finding a combination of predictors that map well to the examples you have. It’s possible to do this by hand, but these days there are a lot of computerized systems that can help develop these models in an automated fashion. It’s often faster and more reliable to fit your algorithm automatically – doing it by hand and doing it right is *hard*. Just making up a model in an afternoon isn’t going to be sufficient.

Second, whether an algorithm is developed by hand or with support from a computer, it needs to be checked for *algorithmic bias.*

Algorithmic Bias

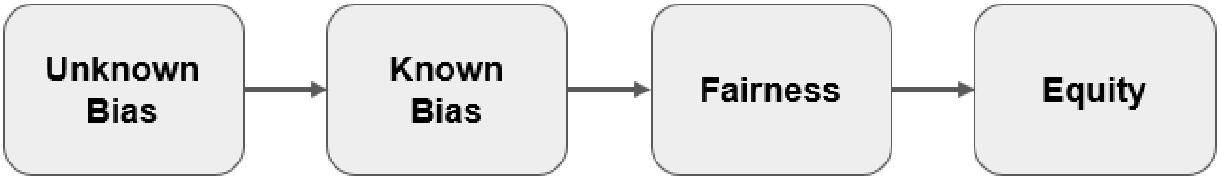
Algorithmic bias has been defined in several ways. One good definition, from Friedman and Nussbaum (1996) is that biased computer systems “*systematically* and *unfairly discriminate* against individuals or groups of individuals in favor of others.”

I’m a bit partial also to the definition in Baker and Hawn (2021), which identifies algorithmic bias as “cases where an algorithm’s performance is substantially better or worse across different groups of learners”.

If we want to fix educational algorithms, and remove bias, there’s a progression we have to follow, shown in Figure 1. At the beginning of this progression, we have unknown bias. We don't actually know how or where the algorithms might be failing. We have, perhaps, a sense of who might be affected, maybe from reports in other areas of algorithms or from just what we know about systemic inequities in society, but we don't actually know how it's playing out in our algorithm.

By doing research, we can move from that to known bias, where we start to know what's going wrong. From there we can work towards fixing the problems we find, moving towards fairness, and finally, towards creating equity.

Figure 1. The progression from algorithmic bias to equity.



A key problem that we see a lot is convenience sampling. It’s essential to collect data to have examples that can be used to develop a model. But a lot of times when people build these models they just sample whoever they can easily get access to. As a result, the model performs less well for groups that are less well represented in the data used to develop the model. It should be obvious by 2021, but it doesn't seem to be, that -- for example -- suburban middle-class students are not the same as urban lower-income students. It should be obvious that learners of different racial or ethnic backgrounds are not identical, and yet we still see people developing machine learning models on populations that are overwhelmingly the students who are easiest to get access to.

A seeming solution to this – often proposed by computer scientists – is to simply include “everybody” in the sample. For example, one might take every single student using an online learning platform, or everyone in a college, and build a model for them. But even “complete” data won’t be enough if a group is rarely seen in the data set. To give an example of that, Anderson and colleagues (2019) took everybody in a university and tried to predict dropout among these learners. They demonstrated that there were so few students of Native American/First Nation background in their sample that it just wasn't possible to make a valid model for those learners. So even “everybody” isn’t good enough – we need purposeful sampling.

What We Know

Going beyond that: what do we know about the bias that impacts learners in common demographic categories? We’re nowhere near as far along as we should be. In one report by Holstein and colleagues (2019), algorithm developers in industry trying to build educational algorithms often struggled to anticipate which sub-populations and forms of unfairness they needed to consider for specific algorithms. Furthermore, Luc Paquette and colleagues showed in a recent report in the Journal of Educational Data Mining (2020) that most research on algorithms in education doesn't even *mention* learner demographics, much less investigate differential impacts or effectiveness for learners in different demographic groups.

There's some biases that are relatively well documented -- and by relatively well documented, I mean poorly-documented but more than nothing. There's been a small number of papers demonstrating algorithmic bias in terms of race and ethnicity, particularly in the context of predicting who will drop out of college or high school (Anderson et al., 2019; Hu & Rangwala, 2020; Lee & Kizilcec, 2020; Yu et al., 2020). There's been a little bit of work on national origin, particularly in linguistic algorithms for measuring language proficiency (Bridgeman et al., 2009, 2012; Ogan et al., 2015). There's been some work on gender across a few contexts (Kai et al., 2017; Anderson et al., 2019; Christie et al., 2019; Gardner et al., 2019; Hu & Rangwala, 2020; Lee & Kizilcec, 2020; Riazy et al., 2020; Yu et al., 2020). Even for these most well-studied categories, there's been insufficient research. We know that specific groups of learners are being affected but we don't know the span of contexts where they're being affected. Or exactly how.

And do we even know about all the groups that are impacted? There's been one study on second language learners (Naismith et al., 2018). Two studies considering learners with disabilities (Loukina et al., 2017; Riazy et al., 2020). Two studies comparing urban, rural, suburban differences between learners (Ocumpaugh et al., 2014; Samei et al., 2015). These studies were in two totally different domains. In one domain this variable made a big difference for the effectiveness of algorithm (Ocumpaugh et al., 2014). In the other it didn't (Samei et al., 2015).

Two studies on parental educational background (Kai et al., 2017; Yu et al., 2020). Two studies on socioeconomic status (Yudelson et al., 2014; Yu et al., 2020). One study on children in military families (Baker et al., 2020).

Many differences and many groups haven't even been studied at all. Intersectionality – these studies haven't even started to look at that. There's been so little work so far.
So if we want to go down the path from unknown bias to known bias, to fairness, to equity, we have a lot of work to do.

Obstacles to Overcoming Algorithmic Bias

There are currently several obstacles to overcoming bias. The first that I'll discuss is lack of data on group membership. These days there are many, many large-scale educational data sets, many of them publicly available to the research communities. Many of these data sets have all the interaction a student did over the course of a year, or they have all the courses a student took. They include a lot of information. And yet these data sets don't have any data on student demographics.

How did we get here? One answer is privacy risks. People are really worried about student privacy – and for good reason. Student privacy is important, but not thinking carefully about balancing privacy versus algorithmic bias means that we can't figure out if algorithmic bias is happening.

IRB processes. With the way the current legislation is set up, it's a lot easier to get IRB approval for studies if you just simply throw out everything identifiable about a learner, including variables that maybe actually aren't all that identifiable. If a school district has 700 students of the same race in a specific grade, is that information really identifiable? Probably not, but if a researcher just throws out that information, it's easier to get IRB approval.

Lack of transparency on bias and group-specific outcomes. There's strong commercial incentives against transparency and against group data collection. If a developer releases data for other people to inspect their algorithms, they risk getting accused of violating privacy. Who wants to risk being accused of bias *and* violating privacy? It's easier just to sweep equity issues under the rug. Privacy becomes an excuse for avoiding accountability.

Universalist models of effectiveness like the What Works Clearinghouse (n.d.) and Evidence for ESSA (Slavin, 2017). There's no question that these initiatives have done a lot to promote better evidence for educational effectiveness, but there are some unintended consequences. These clearinghouses try to collect the evidence on what's effective *in general,* not evidence on what's effective where and for who. Their websites generally treat a curriculum as having evidence for being effective or not effective overall. A lot of curricula will work in some places – for some learners, in some contexts of use -- and not in others. These clearinghouses don't take this subtlety into account.

Where Do We Go From Here?

So where do we go from here? One key step is improved data collection. We need to collect data on group membership. We need to collect as many variables about learner identity as we can. We don't yet have enough evidence to really make broad recommendations about all of the categories that data needs to be collected for, but we know some categories already – see the discussion earlier in this talk. We need to collect those variables.

As a prerequisite towards that, we need to encourage regulators, IRBs, other privacy officers to balance the risk of privacy violations with the risk of missing out on algorithmic bias.

There are potential ways to reduce privacy risks while still being able to use fuller information. There are steps we can take like data obfuscation (Bakken et al., 2014), where the data steward takes categories that are sufficiently rare that one could re-identify students, and merges them together until categories have a sufficient number of members that it becomes infeasible to re-identify specific students.

Alternatively, data may be made available for specific uses and under oversight (Meyer et al., 2012). For instance, a data enclave can be configured such that a researcher can enter an analysis and then get information on the results of their analysis. But they can't actually see the data they're analyzing in terms of potentially identifiable variables – they can conduct analyses using those variables, but can't see the values for specific students (Gardner et al., 2018).

A third option is legal agreements for access to data like the ASSISTments Project uses (ASSISTments, 2014), which requires anyone analyzing their data to sign a legal agreement about how they'll use it first, including agreeing not to attempt to re-identify students.

Beyond attempts to make demographic variables available, we also need to make sure that the right data is available. We need to work on creating practices for making sure the training sets we use to develop algorithms are representative. Beth Tipton, formerly of Teachers College, heads The Generalizer Project, which helps researchers select samples of students so that key groups of learners are sufficiently represented (Tipton, 2014). Not all the variables that are necessary are included in that system – in part because more research needs to be done to determine which variables should be included -- but it's a great start. Through this, we can address underrepresentation of key groups in data sets.

A related key step is to figure out how much data we need from the groups we want to include in our data sets. Statisticians have power analysis to answer this question; machine learners and data miners largely still don't. We need to have methods that provide a good way to know how much data is enough data to know that our algorithms are less likely to be biased. Machine learners don't tend to think about sample size issues because the data sets that are used in areas other than education are often extremely large. Data sets aren't always as big in education, and even a large data set might not include enough data for a specific subgroup of importance.

Another key step is to facilitate and incentivize openness. Scientific societies and journal editors and publishers have a role to play here – by creating guidelines for openness around data and algorithmic bias in academic publication. Similarly, we should start moving as a field towards conducting more regular review of algorithmic bias. Today, reviews are used to check for bias in research on new medicines (Ciociola et al., 2014) – if a similar process were adopted by clearinghouses in education, it would make a significant positive impact.

A final recommendation is to broaden the community. Data science is a notoriously non-diverse field. Members of the communities being affected can always do a better job of advocating for their perspective than well-meaning outsiders, who often fail to fully understand all the constraints and factors that must be considered for a solution to be successful. We need to consider the biases and the blind spots in the people developing algorithms if we're to fix the biases and blind spots in the algorithms (cf. Holstein et al., 2019). One step towards this is to broaden the pipeline of talent going into data science; another step is to create tools that make it possible for a broader range of people to be involved in doing data science work, even if they don't have a background in data science. If anything, we are going in the wrong direction – far more investment has gone into toolkits that involve programming in Python than in updating and enhancing easy-to-learn graphical data science tools like RapidMiner.

Conclusions

In this talk, I've given a brief overview of the problem of algorithmic bias in education. I've discussed the existing evidence for algorithmic bias in education – as limited as it is. I've discussed where more research is needed. I've then talked about key steps that could enhance the field's capacity to discover and address algorithmic bias.

It's possible to envision a trajectory that, if we follow it, can make education fairer, more equitable, more just. This path leads from unknown bias – where we mostly are today -- to known bias, to fairness, to equity and justice. There's a lot of work to do, but at the end of it we can the potential to have algorithms that are verifiably fair and that help realize education's promise to improve students' futures equitably.

References

Anderson, H., Boodhwani, A., Baker, R. (2019) Assessing the Fairness of Graduation Predictions. Poster paper. *Proceedings of the 12th International Conference on Educational Data Mining*, 488-491.

ASSISTments (2014) ASSISTments Data Terms of Use for Using Data. Retrieved June 12, 2021 from https://sites.google.com/site/assistmentsdata/termsofuseforusingdata

Baker, R.S. (2020) Big Data and Education. 6th Edition. Philadelphia, PA: University of Pennsylvania.

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. edarXiv manuscript PBMVZ.

Baker, R. S., Berning, A., & Gowda, S. M. (2020). Differentiating Military-Connected and Non-Military-Connected Students: Predictors of Graduation and SAT Score. EdArXiv manuscript CETXJ.

Bakken, D. E., Rarameswaran, R., Blough, D. M., Franz, A. A., & Palmer, T. J. (2004). Data obfuscation: anonymity and desensitization of usable data sets. IEEE Security & Privacy, 2(6), 34–41.

Bowers, A. J., Sprott, R., & Taff, S. A. (2012). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 77-100.

Bridgeman, B., Trapani, C., & Attali, Y. (2009). Considering fairness and validity in evaluating automated scoring. [Paper presentation]. Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA, United States.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education, 25*(1), 27–40.

Christie, S. T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. (2019). Machine-Learned School Dropout Early Warning at Scale. *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019),* 726–731.

Ciociola, A. A., Cohen, L. B., Kulkarni, P., & FDA-Related Matters Committee of the American College of Gastroenterology. (2014). How drugs are developed and approved by the FDA: current process and future directions. *The American Journal of Gastroenterology, 109*(5), 620–623.

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems, 14*(3), 330–347.

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. S. (2018). MORF: A Framework for Predictive Modeling and Replication At Scale With Privacy-Restricted MOOC Data. 2018 IEEE International Conference on Big Data (Big Data), 3235–3244.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge,* 225–234.

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–16.

Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020),* 431–437.

Kai, S., Andres, J. M. L., Paquette, L., Baker, R. S. ., Molnar, K., Watkins, H., & Moore, M. (2017). Predicting Student Retention from Behavior in an Online Orientation Course. *Proceedings of the 10th International Conference on Educational Data Mining*, 250–255.

Lee, H., & Kizilcec, R. F. (2020). Evaluation of Fairness Trade-offs in Predicting Student Success. ArXiv E-Prints, arXiv:2007.00088. https://arxiv.org/abs/2007.00088

Loukina, A., & Buzick, H. (2017). Use of Automated Scoring in Spoken Language Assessments for Test Takers With Speech Impairments. *ETS Research Report Series,* 2017(1), 1–10.

Meyer, P. S., Robinson, E. S., & Madans, J. (2012). Protecting confidentiality in a data enclave. *Statistical Journal of the IAOS*, *28*(1, 2), 25-30.

Milliron, M. D., Malcolm, L., & Kil, D. (2014). Insight and Action Analytics: Three Case Studies to Consider. *Research & Practice in Assessment*, *9*, 70-89.

Naismith, B., Han, N.-R., Juffs, A., Hill, B., & Zheng, D. (2018). Accurate Measurement of Lexical Sophistication with Reference to ESL Learner Data. *Proceedings of 11th International Conference on Educational Data Mining,* 259–265.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology, 45*(3), 487–501.

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards Understanding How to Assess Help-Seeking Behavior Across Cultures. *International Journal of Artificial Intelligence in Education, 25* (2), 229–248.

Paquette, L., Ocumpaugh, J., Li, Z., Andres, J.M.A.L., Baker, R.S. (2020) Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining, 12* (3), 1-30.

Riazy, S., Simbeck, K., & Schreck, V. (2020). Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments. *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020), 1,* 15–25.

Ritter, S., Yudelson, M., Fancsali, S. E., & Berman, S. R. (2016, April). How mastery learning works at scale. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 71-79).

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., & Graesser, A. (2015). Modeling Classroom Discourse: Do Models That Predict Dialogic Instruction Properties Generalize across Populations? *Proceedings of the 8th International Conference on Educational Data Mining,* 444–447.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Slavin, R. E. (2017). Evidence-based reform in education. *Journal of Education for Students Placed at Risk (JESPAR)*, *22*(3), 178-184.

Smith, H. (2020). Algorithmic bias: should students pay the price? *AI & SOCIETY*, *35* (4), 1077–1078.

Tipton, E. (2014). Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments. *Evaluation Review, 37*(2), 109–139.

What Works Clearinghouse. (n.d.). What Works Clearinghouse. Institute of Education Sciences. Retrieved from http://ies .ed.gov/ncee/wwc/

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020),* 292–301.

Yudelson, M. V., Fancsali, S. E., Ritter, S., Berman, S. R., Nixon, T., & Joshi, A. (2014). Better Data Beat Big Data. *Proceedings of the 7th International Conference on Educational Data Mining,* 205–208.