Qualifier Question

Prompt

You've identified several potential sources of bias in education. What quantitative methods have been used to detect the bias in such a way as to avoid inadvertently picking up other correlated factors and demonstrate causality? How can you apply this to the topics you mention?

Introduction

Bias in education is a growing area of research, especially as the societal push for equality in the education system becomes stronger. However, humans still tend to be especially skeptical of research findings on bias, because they conflate the idea of admitting the existence of bias with admitting that they themselves are biased (Saul, 2013). While maintaining a healthy dose of skepticism is perfectly reasonable in many scientific scenarios, bias in traditional classroom environments has been experimented and evidenced beyond the scope of reasonable doubt. The goal of my research is to perform several quantitative experiments in the field of bias in online education to add to the growing amount of evidence indicating existence of bias in online education. In this paper, I will introduce several popular multidisciplinary quantitative research and experimental methodologies, explain how they have previously been used to establish statistically significant findings within both human and algorithmic bias, discuss specific examples of these methodologies within the field of education, and describe how I plan to use these methodologies within my own research.

Quantitatively Measuring Human Bias

In the decades that human bias has been explored in a classroom context, researchers have utilized increased amounts of educational data and advances in technology to provide stronger statistical evidence to support their claims. As such, there are many different quantitative techniques that a researcher can choose to use in their study. Making the decision to use certain techniques over others primarily depends on whether the bias is detected through a controlled experiment that generates new evidence or through a novel analysis of existing data. Secondarily, the quality of the data being generated or analyzed also helps to determine the most effective quantitative technique to use for statistical analyses.

Controlled experiments can be used as a method of gathering data to detect and analyze human bias in education. While controlled experiments aren't always grounded in quantitative evidence, they provide a strong definition for causality and a structured framework to test causality. To demonstrate causality through a controlled experiment, researchers must show statistically significant evidence supporting a hypothesis, or they must show similarly statistically significant evidence negating the contrapositive, which is the partial or full opposite of the hypothesis (Cohen & Nagel, 1934). Cohen and Nagel state that experimental evidence can be considered causally valid if every variable and invariable aspect of the experiment is controlled for, with the exception of the variables that the hypothesis is testing (1934).

In existing research, most controlled empirical experiments looking to show causality have followed one of two designs: between-subjects design, where participants in a study are separated into groups and each group is exposed to a separate condition and differences are measured between groups, and within-subjects design, where participants in a study are exposed to all of the conditions and differences are measured within the entire participant space (Charness, Gneezy, & Kuhn, 2012). Both between-subjects and within-subjects designs have been used in existing experiments on bias in

education. Here are two examples of controlled experiments that study first-name bias in the education system – one uses between-subjects design, and the other uses within-subjects design:

- **Between-Subjects Design:** In this study, researchers measured first-name bias in the online classroom by having the same instructor pose as both male and female for different sections of their class (MacNell, Driscoll, & Hunt, 2015). MacNell et al.'s study is an example of between-subjects design: each student was exposed to either the male or female presentation of the instructor (2015). This experiment was performed with multiple instructors of all genders to verify that the results were statistically significant (MacNell, Driscoll, & Hunt, 2015).
- **Within-Subjects Design**: In this study, researchers created posts in an online classroom using fabricated student names, and measured the subsequent replies and engagement towards the post based on the perceived race and ethnicity from the student's name (Baker, Dee, Evans, & John, 2018). Baker et al.'s study used within-subjects design: each class in the study was exposed to posts by each fabricated student name, and the other dependent variables (contents of the post, date/time the post was created) were rotated within the classes to avoid false correlations (2018).

Controlled experiments can be designed to show strong links of causality in their results, but designing an experiment also has no shortage of challenges, with the biggest challenge being that the researchers might project their own biases onto their experiment's design. In the two examples above, the authors specifically discussed measures that they took to validate their experiments and results. Similarly, there are several internal biases that any researcher should minimize or avoid in their experiments to validate their results: selection bias may occur if there are significant differences between populations selected for the control/experimental research and the general population, historical and maturation bias may occur if factors such as time affect the experiment and are not accounted for, and the absence of control data might render experimental data completely invalid (Kirk, 2007). While checking a controlled experiment for validity is not a purely quantitative task, it is as important to producing reliable results as any other quantitative measure would be.

In addition to controlled experiments, there are several other experimental methods designed to quantify bias, and the Implicit Association Test (IAT) has been one of the most popular methods for decades. In the IAT, participants are shown pictures that fall into one of two categories (such as dogs and cats), they are then instructed to label pictures within a category as either pleasant or unpleasant, and bias is measured as quicker association (whether positive or negative) towards one category of pictures (Greenwald, McGhee, & Schwartz, 1998). The IAT acts as a starting point to identify potential biases in a person, but it is limited in scope because it can't tell researchers anything about how one's implicit biases manifest themselves in a classroom environment. Therefore, causality (at least in the context of bias in education) cannot be demonstrated using the IAT alone.

Another method of demonstrating the presence of human bias in education is through a novel statistical analysis of pre-existing data. When working with existing datasets, several statistical techniques can be used to validate findings of bias. The proper statistical technique to use may depend on several factors within the dataset, including the number of variables/features, the number of entries, and whether the dataset is obtained from a single study or a meta-analysis of multiple studies.

One of the simplest statistical methods in use is analysis of variance (ANOVA). ANOVA involves measuring the total variance, or difference between sum of squares, between a control group and one or more experimental groups (Kirk, 2007). ANOVA is useful in certain contexts within bias detection: for example, ANOVA was used in MacNell et al.'s between-subjects study to verify that student behaviors such as engagement were not statistically significantly different across groups (2015). However, ANOVA has its downsides: when used with data that has multiple dependent variables, ANOVA is weak at showing causality (Northcott, 2008). To make up for the weaknesses of ANOVA in multivariate data, two other statistical methods may be used: ANCOVA and MANOVA. ANCOVA adds regression analysis to

ANOVA to control for multiple variables (Kirk, 2007), while MANOVA performs multiple iterations of ANOVA with slight changes to independent variables such as time (O'Brien & Kaiser, 1985). Together, these three statistical methods can be used in an experimental design to quantitatively demonstrate causality for research topics such as bias in education.

Statistical methods often rely on meta-analysis or similar data-munging techniques to gather enough statistical data to perform a reliable statistical analysis. In the cases where multiple datasets are used in a novel analysis, there are several algorithms to test the quality of the data being concatenated. One such example of this is the fail-safe analysis, which calculates the number of studies that would need to refute the data contained within the meta-analysis in order to provide a significant rebuttal to the meta-analysis (Persaud, 1996). By verifying that the fail-safe analysis is sufficiently high, a researcher is demonstrating that their meta-analysis uses a diverse set of data rather than just cherry-picking the data that supports their research hypotheses the best. Another technique specific to meta-analysis is the concept of statistical mediation analysis, which is used in Forscher et al.'s meta-analysis concerning the causality between implicit bias and behavior (2016):

> "Of course, correlations between variables can be produced by many relationships besides ones that are causal. To get closer to questions of causality, we looked at whether changes in implicit measures correspond with and mediate changes in behavior in our sample of randomized experiments (p. 41)."

However, Forscher et al. outlines many weaknesses in the mediation model, including the fact that not enough studies may exist for the technique to be accurate, and even the studies that provide input to the mediation model may vary in quality (2016). While meta-analysis is an interesting quantitative technique for biases that have been heavily studied, I believe that its applicability towards bias in online education is limited until more data is available.

In my own research, I plan to perform a novel analysis for a couple of reasons: much of the data I want to experiment with already exists and is easily accessible, and I want to maximize my findings within the time constraints of a semester-long course. While gathering enough evidence through controlled experiments might be outside of the scope of a single semester of work, many of the methodologies used in controlled experiments can still be applied to a novel analysis, if there is enough high-quality data. In fact, if the dataset being used for novel analysis is large enough such that certain subsets of the data control for all variables except for the hypotheses being tested, the scientific method can be used to demonstrate causality. This practice is also known as secondary data collection, and must be performed with several precautions, including verifying that the data is properly cleaned and that the quantitative measures are internally and externally valid (Hox & Boeije, 2005). External validity can be verified using ANOVA/ANCOVA/MANOVA, similarly to the demonstrations shown earlier in this paper. Internal validity could be manually demonstrated by explaining the procedural steps taken to avoid any biases that may be injected into the design of the experiment and analysis of the data.

Quantitatively Measuring Algorithmic Bias

Many of the quantitative methods for measuring human bias are also somewhat applicable to measuring algorithmic bias. However, to truly understand the depth and severity of algorithmic bias, researchers are tasked with an additional set of complexities, such as determining the source of the bias and proving algorithmic fairness.

Once statistical analysis is complete, it is fairly trivial in human-based bias studies to determine the individuals who hold biases within a larger group of people, whereas it is significantly more difficult to single out a layer in a neural network or a node in a decision tree that is biased within a multifaceted algorithm. In most cases, a decision-making algorithm isn't even inherently biased until biased data is used to train the algorithm (Zarsky, 2016). Some researchers have found

success in detecting and removing the subset of biased data from the larger dataset (Amorim, Cancado, & Veloso, 2018), but not every dataset will have enough data to train AI models after biased data is removed.

An added roadblock to adopting algorithmic decision-making processes in the education system is the inherent need for educators, students, and even parents to perceive the algorithm as fair. Humans inherently don't trust algorithms as much as other humans to make decisions, because they believe that algorithmic decision-making processes do not have the capabilities to explain their decisions like a human might, and they may not be held to the same accountability standards as humans are (Binns et al., 2018). It is clear that quantitatively detecting zero bias in an algorithm would not necessarily be enough for people to be confident in its fairness.

While OMSCS still relies heavily on human instructors in the educational process, it could be interesting to use quantitative methods to detect bias in existing auto-grading algorithms across a dataset of existing papers. In the future, perhaps the analysis of variance between auto-grading algorithms and human grades could even be used as a method of detecting human bias, once the algorithms themselves can be proven to be fair and unbiased themselves. Many challenges still remain to using algorithms as a quantitative method of demonstrating bias (or lack of it) manifested by humans, but it is a promising area of research which deserves to be investigated.

References

1. Alavi, S. M., & Bordbar, S. (2018). Differential Item Functioning Analysis of High-Stakes Test in Terms of Gender: A Rasch Model Approach. *MOJES: Malaysian Online Journal of Educational Sciences*, *5*(1), 10-24.
2. Amorim, E., Cançado, M., & Veloso, A. (2018). Automated Essay Scoring in the Presence of Biased Ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*(Vol. 1, pp. 229-237).
3. Baker, R., Dee, T., Evans, B., & John, J. (2018). Bias in online classes: Evidence from a field experiment. In *SREE Spring 2015 Conference, Learning Curves: Creating and Sustaining Gains from Early Childhood through Adulthood* (pp. 5-7).
4. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 377). ACM.
5. Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1-8.
6. Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method*. Oxford, England: Harcourt, Brace.
7. Forscher, P. S., Lai, C., Axt, J., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2016). A meta-analysis of change in implicit bias.
8. Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.
9. Hox, J. J., & Boeije, H. R. (2005). Data collection, primary versus secondary. *Encyclopedia of Social Measurement*, 1(1), 593-599.
10. Jacoby-Senghor, D. S., Sinclair, S., & Shelton, J. N. (2016). A lesson in bias: The relationship between implicit racial bias and performance in pedagogical contexts. *Journal of Experimental Social Psychology*, *63*, 50-55.
11. Kirk, R. E. (2007). Experimental design. *The Blackwell Encyclopedia of Sociology*.
12. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076).
13. MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4), 291-303.
14. Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, *60*(3), 245-256.
15. Northcott, R. (2008). Can ANOVA measure causal strength?. *The Quarterly review of biology*, *83*(1), 47-55.
16. O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: an extensive primer. *Psychological bulletin*, *97*(2), 316.
17. Persaud, R. (1996). Misleading meta-analysis. *BMJ: British Medical Journal*, *312*(7023), 125.
18. Saul, J. (2013). Scepticism and implicit bias. *Disputatio*, *5*(37), 243-263.
19. Unterhalter, E. (2017). Negative capability? Measuring the unmeasurable in education. *Comparative Education*, *53*(1), 1-16.
20. Xu, D., & Jaggars, S. S. (2014). Performance gaps between online and face-to-face courses: Differences across types of students and academic subject areas. *The Journal of Higher Education*, *85*(5), 633-659.
21. Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.