Constructing Categories: Moving Beyond Protected Classes in Algorithmic Fairness

Clara Belitz[1], Jaclyn Ocumpaugh[2], Steven Ritter[3], Ryan S. Baker[2], Stephen E. Fancsali[3], and Nigel Bosch[1]

[1]School of Information Sciences, University of Illinois Urbana–Champaign

[2]Graduate School of Education, University of Pennsylvania

[3]Carnegie Learning, Inc.

Author Note

This research was supported by NSF grant no. 2000638. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to Clara Belitz, School of Information Sciences, 501 E. Daniel St., Champaign, IL 61820. Email: cbelitz2@illinos.edu

Abstract

Automated, data-driven decision-making is increasingly common in a variety of application domains. In educational software, for example, machine learning has been applied to tasks like selecting the next exercise for students to complete. Machine learning methods, however, are not always equally effective for all groups of students. Current approaches to designing fair algorithms tend to focus on statistical measures concerning a small subset of legally protected categories like race or gender. Focusing solely on legally protected categories, however, can limit our understanding of bias and unfairness by ignoring the complexities of identity. We propose an alternative approach to categorization, grounded in sociological techniques of measuring identity. By soliciting survey data and interviews from the population being studied, we can build context-specific categories from the bottom up. The emergent categories can then be combined with extant algorithmic fairness strategies to discover which identity groups are not well-served, and thus where algorithms should be improved or avoided altogether. We focus on educational applications but present arguments that this approach should be adopted more broadly for issues of algorithmic fairness across a variety of applications.

Keywords: algorithmic fairness, categorization, machine learning, identity, education

Constructing Categories: Algorithmic Fairness and Grounded Identity

Automated, data-driven decision-making is increasingly being applied in a variety of domains. The algorithms used to make these decisions consume massive amounts of information and affect outcomes from information search and retrieval to parole decisions (Barocas & Selbst, 2016). One such system used in middle and high school math, MATHia, uses machine learning models to predict content mastery, engagement-related behaviors and affective states, and self-regulated learning behaviors using information extracted from records of students' interactions with the software (Ritter et al., 2007; Ritter & Fancsali, 2016). MATHia and similar software like ASSISTments (Heffernan & Heffernan, 2014) are utilized by large, diverse populations of students. Such systems therefore need to cater to a wide variety of learning needs, though in practice these systems may not be equally effective for all groups (Baker & Hawn, 2021).

As algorithmic decisions become more prevalent, so do questions of algorithmic bias and fairness (Hutchinson & Mitchell, 2019). Algorithmic fairness can be difficult to define universally, given the ethical and pragmatic factors which impact perceptions of that fairness (Saxena et al., 2019; Woodruff et al., 2018). One potential operational definition is given for decision making by Mehrabi et al. (2021), who define algorithmic fairness as "the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics." Much as we expect decisions made by a human to be fair, we might expect the same of automated decisions. Neither humans nor algorithms, however, are always fair. And bias, of course, already exists in school environments (Aronson & Laughter, 2016). The added danger of algorithmic (un)fairness is the scale at which algorithmic decisions can proliferate biases (Benjamin, 2019). Reducing algorithmic bias, therefore, has the potential for broad impacts. In the context of adaptive learning software, research uncovering exactly which groups are not well served by such software has only just begun (Kizilcec & Lee, 2020; Paquette et al., 2020). In this paper we focus on issues of how individuals affected by algorithms are grouped together and how this affects approaches to evaluating and improving fairness. We examine adaptive learning software as an example, but these ideas are relevant in any context in which information about individuals is used to make algorithmic decisions.

Many statistical definitions of fairness for categorization have been proposed in recent years (Berk et al., 2017). These definitions are generally understood as the absence of discrimination and tend to define a *protected class*, which is frequently related to demographic categories. While the protected class is often a legally protected category like race or gender, it can also be a proxy for legally protected classes (e.g., ZIP code as a proxy for race) or any other potentially disadvantaged status. These definitions are then used to produce a mathematical formulation of equality. Formulations can focus on individual or group outcomes, but generally account for only one protected class and measure fairness at the time of decision making. Some formulations might try to incorporate more than one protected class, long term impacts, or other types of parameters that require more than simply maximizing accuracy (Belitz et al., 2021; Liu et al., 2019; Mouzannar et al., 2019). Overall, however, research on algorithmic fairness has historically favored mathematical rigor over investigating complicated social implications. Such approaches are important first steps but have limitations in breadth as well as effectiveness (Mitchell et al., 2021).

A promising trend in recent research has sought to identify and counteract underlying sources of algorithmic bias. For example, in natural language processing (NLP), the hierarchies on which language rest can (re)create self-perpetuating cycles of social-category stereotypes in NLP systems. The way these social categories are defined and subsequently used for labeling can thus play an important role in perpetuating or breaking stereotypes in language applications (Beukeboom & Burgers, 2019). As such, a framework has been proposed which describes specific countermeasures, informed by social science and desired outcomes, to common sources of predictive bias (Shah et al., 2020). Similarly, Blodgett et al. (2020) have developed concrete questions to interrogate the way "social hierarchies and language ideologies" impact NLP systems as well as the way these systems can reproduce and enforce these hierarchies and ideologies.

Across machine learning domains, the highly mathematical nature of these definitions of bias means that assumptions must be met for the definition to hold (Cooper & Abrams, 2021). The assumptions made around protected groups are both statistical and cultural. Statistical assumptions include the idea that the sample accurately represents the population from which it was drawn. Even representative data may produce skewed results, however. For example, if conducting data-driven research in a school that is 90% white and 10% Black, perfectly representative data could still produce inaccurate predictions for Black students. Data can also encode societal biases. Consider, for example, that schools in lower socioeconomic ZIP codes generally score lower on standardized tests; this correlation most likely reflects a lack of educational resources, not the intrinsic testing ability of the students (Geiser, 2015). Additionally, race is socially constructed and does not neatly fit into a few pre-defined boxes (Hanna et al., 2020). Narrow demographic definitions of protected categories tend to ignore these social factors. Therefore, singular dimensions of protected categories are not necessarily the appropriate frame with which to measure fairness.

We argue that simple, unidimensional notions of demographics as the only appropriate categories for studying bias miss the other ways in which bias can manifest, as these categories are often incompletely representative of identity. We develop a holistic, bottom-up, mixed methods approach. By utilizing surveys and interviews, we can identify categories that reflect the population and context in which we are automating decisions, make conscious and intentional choices about the groups for whom we measure algorithmic fairness, and return agency to algorithm-affected individuals by allowing them to define their own identities. We propose that there needs to be a shift in how groups are defined to allow algorithmic bias to be explored with respect to the sociopolitical nature of individual identity, power structures, and individual as well as group outcomes. Researchers must be flexible in our measured categories and understand that traditional demographic variables may not be sufficient for us to achieve our desired results of fairness. For example, if we were to measure against binary definitions of sex alone in adaptive learning software, or even to include a non-binary category, we may think that we have achieved equitable outcomes in learning when in fact we are ignoring other important variables like race, family responsibilities, or interests. And, a non-binary category is still likely not reflective of student gender, since non-binary identity can manifest along multiple dimensions. Fair decision-making would ideally include similar accuracy for all relevant identities—including non-demographic identities—and ensure equitable learning outcomes for students.

Throughout the rest of this paper, we discuss theoretical foundations and practical ideas for a shift in the analysis of algorithmic fairness. We intend to bring a concern for human values and a more nuanced use of identity in technical systems to the study of algorithmic bias.

Parallels to Other Fields and Related Work

Categorization

Categorization and organization are vital to data science and have long been topics of interest in information science (Bowker & Star, 1999). Categorization, by definition, makes choices about what information is deemed important and relevant. Making choices is not a bad thing in and of itself, and categorizing information is vital to being able to use and retrieve it. But, because categories require value judgement and have ethical implications, it is imperative that our categorizations are made consciously (Bowker & Star, 1999). As typically defined, common categories like race or sex are rigid and unidimensional, leaving little room for fluid boundaries (Cunningham, 1997). While collapsing the differences between individual experiences and racial identities can be problematic for all racial categories (e.g., being white in rural Idaho might be a very different experience than being white in New York City), the growing category of "multiracial" Americans, who find themselves grouped into one label of "more than one race" or "other," highlights how these categories may obscure even group-level identities. Likewise, in both data science and other areas of research; the question of where to classify transgender identities, for example, has caused cataloging practices to enforce normative boundaries on queer identities (Roberto, 2011).

Emergent categories and structures are vital to creating information domains and thus to the creation of knowledge (Bates, 2005). Ignoring the constructed reality of labels like race or gender is thus a loss of information. As such, some data science applications are beginning to be scrutinized regarding the value, limitations, and implications of highly standardized race and gender identity labels. Databases used for training facial recognition algorithms contain assumptions about the static and apolitical nature of these categories, but rarely explain how their categories were constructed (Scheuerman et al., 2020). These databases also often use labels generated by outsourced human labor like Amazon Mechanical Turk, which rely on human assumptions drawn from visual images alone, and only designate a small number of racial categories (Khan & Fu, 2021). The policies used when creating datasets therefore have lasting impacts on the categories to which we have access, the unfairness we can measure, and the policy changes we recommend based on model outcomes (Kasy & Abebe, 2021).

Identity and Intersectionality

When discussing identity, it is vital to recognize that unidimensional categorizations cannot capture all forms of (dis)advantage—a phenomenon that has been discussed at length in a growing body of literature on intersectionality but that data science has been slow to adopt (Hoffmann, 2019). Intersectionality moves beyond a simple additive approach with regards to categories like race, gender, and class, acknowledging that individuals at the intersection of one or more marginalized groups may experience discrimination in a way that is not experienced by members of only one of those groups; intersectional approaches also seek to make axes of privilege visible and explicit (Crenshaw, 1989). Because identity and oppression are not always neatly categorized, fields outside of data science have argued that mixed methods are the best way to approach intersectional identity issues (Trahan, 2011).

Intersectionality specifically deals with those identities that confer privilege (or oppression). Not all identities, however, are inherently privileged, but even those that are not may correspond with societal categories that typically confer power. Consider a student's identity as a "math person." Math identity is a non-protected identity but can still influence learning experiences and outcomes. Of course, social norms and privileges may play into who identifies as a "math person." For example, students from higher socioeconomic classes and with more enthusiastic teachers typically have higher rates of math interest (Frenzel et al., 2010). Math is also still frequently considered a "male" domain (Frenzel et al., 2010). Math identity is developed by both self-belief and perception of how others see the student and has a positive feedback loop with interest (Cribbs et al., 2015). It is a confluence of both personal and societal factors. Though "non-math people" are not a legally protected group, understanding students' relationships to their math identities can lead to better learning outcomes for all students, challenge the idea of "math people," and help us design for the multiple underlying identities that contribute to math performance.

Proposed Approach

The question of how to best build categories for measurements of algorithmic fairness is broad and likely differs to some degree between applications. Firm rules and definitions can never fully encompass the complications of human reality, but we must strive to ensure that our algorithmic categorizations are contextually sensitive with an eye to desired outcomes. We are currently investigating research gaps in categorization for algorithmic fairness using the context of adaptive learning. Adaptive learning systems aim to create a flexible environment that supports learning for students. In theory, many adaptations could support students with a range of (dis)abilities, interests, and backgrounds, but the challenge of accurately providing these adaptations means many systems focus on adapting for skills, mastery, and prior knowledge only (Shute & Zapata-Rivera, 2012). Adaptive learning software is an appropriate investigative venue due to its current widespread usage in classrooms; diverse student populations provide opportunities to explore, challenge, and improve our approaches to categorization. Additionally, learning outcomes and goals provide measurable outcomes for the population being studied.

To uncover identity categories from the bottom up, as mentioned in the Introduction, we will employ a variety of qualitative methods. Our current work in progress uses the Twenty Statements Test (TST), which uses a free-form answer format to elicit self-concept by having individuals answer the question “Who am I?” up to 20 different ways in a short amount of time (Kuhn & McPartland, 1954). Responses can then be coded into categories that either match those that are well established in literature, such as Gordon & Gergen's schema (1968), or that arise from the data via thematic analysis. Pilot studies have allowed us to add specific categories that arose, such as economic status, and anticipate likely new categories, such as sexual orientation. The categories are hierarchical, allowing for the integration or separation of groups of responses as appropriate. For example, technological activities, athletic activities, and intellectual concerns all fall into the parent category of “Interests and Activities.” Because there are twenty questions on the survey, students may also respond with multiple answers that fall into the same category, such as being both American and Cape Verdean (i.e., both nationalities). Second, we will use qualitative interviews to elaborate on TST responses to better understand students and their experiences with adaptive educational software. This approach will allow space for the fluid reality of gender and racial categories, avoid missing multi-racial, genderfluid, or otherwise “complicated” demographic identities, and capture non-demographic identities. In the pilot data, for example, a composite profile included someone who describes themselves as a nice, intelligent, lesbian, African American hobby artist.

We claim that the best approach will combine this bottom-up strategy with categories informed by a top-down approach from years of sociological research; this combination can be adapted and reapplied in other domains. The generation of these categories will be done using a coding scheme developed using the TST and adapted to fit modern responses (e.g., gaming in addition to artistic and athletic activities) (Gordon & Gergen, 1968). These holistic identity categories can then be used to identify algorithmic fairness via existing measures that compare accuracy across groups or propensity to make a particular algorithmic decision for one group (Hutchinson & Mitchell, 2019). While we cannot hope to completely eliminate bias from all algorithmic systems, we can strive for an approach that better serves the need of a diverse group of users. Measures that assess algorithmic biases inherently bring their own value judgments, but concretely thinking about the meanings of bias and fairness in a specific domain and context allow the choice of measurements that help achieve fairness goals.

In the context of adaptive learning, we propose to use these holistic identity categories to address bias in the measurement of academic success. We plan to focus on three examples of algorithmic measurements that are common in adaptive learning: predictions of student knowledge, student engagement, and self-regulated learning behaviors. For example, a machine learning algorithm may mis-predict the initial knowledge of a group of students who differ from the majority of students in terms of parental education level, which could affect students' exposure to concepts before they are introduced in class. Similarly, students for whom English is not a first language may be more likely to make mistakes due to language barriers, which could in turn be interpreted by the system as careless errors or lack of engagement. By developing a widely-applicable approach to constructing algorithmic categories, we can be cognizant of individual lived realities and responsive to the desired outcomes. In the example of adaptive learning software, the goal is to ensure all students can benefit from the individualized learning experience such software provides.

Directions for Research

Algorithmic decisions that correlate with demographics do not necessarily directly reflect the underlying causes. For example, if underlying rates of failure with an automated math tutoring system depend on help seeking behavior, and culturally-conditioned, gendered practices affect this behavior, gender is not itself the cause. One important research goal for our proposed approach is to recognize where algorithmic biases based on demographic information might also be explained by other correlated, non-demographic factors, like personality or social identity. Similarly, existing power structures, like variable resource distribution to schools, may impact algorithmic biases. For example, comparisons of emergent categories across school districts may help us understand how gender and socioeconomic class interact to cause adaptive educational software to be less effective for historically under-resourced groups. This information can help us to develop more equitable and effective technology.

Data collected to date, in adaptive learning and well beyond, often lack critical demographic and identity information. Work like ours can lead to evidence of this lack of information as well as demonstrate the harms caused by current categorization practices. As such, we are calling for a shift in data collection techniques, beyond just a single project. We should go beyond simply asking about a small set of variables. We should go beyond technologies assuming individual identity based on demographics, web searches, or otherwise public data. Users should have the ability to describe themselves. When the categories we define come from those with power, or are otherwise over-simplified, we cannot fully understand the way that specific identities interact with algorithms to produce unfair outcomes. Our proposed approach can shed light on algorithmic bias as a whole and help tease out the impacts of poor data representation, existing social bias, and previously unexamined underlying factors. Our goal is not to develop universally applicable answers, but rather to develop a framework for pursuing this work going forward.

Conclusion

Though the proposed approach moves beyond traditional demographic categories, we can still aim to develop general strategies and methodologies to inform research that uses these categories going forward. Categorization has implications for what we are able to know and study. As such, it is important to acknowledge that the choices we make when creating categories for machine learning have human impacts. We believe that our specific research can act as a blueprint for developing grounded, interactive, human-centered data science. We recognize that this blueprint can be expanded and adapted. Data science should see categorization in algorithmic fairness as an opportunity to challenge, rather than replicate, extant discrimination.

References

Aronson, B., & Laughter, J. (2016). The theory and practice of culturally relevant education: A synthesis of research across content areas. *Review of Educational Research*, *86*(1), 163–206. https://doi.org/10.3102/0034654315582066

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00285-9

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(671).

Bates, M. J. (2005). Information and knowledge: An evolutionary framework for information science. *Information and Knowledge*, 30.

Belitz, C., Jiang, L., & Bosch, N. (2021). Automating procedurally fair feature selection in machine learning. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 11. https://doi.org/10.1145/3461702.3462585

Benjamin, R. (2019). *Race After Technology*. Polity Press.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. *ArXiv:1703.09207*.

Beukeboom, C. J., & Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, *7*, 1–37. https://doi.org/10.12840/issn.2255-4165.017

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Bowker, G. C., & Star, S. L. (1999). *Sorting Things Out: Classification and its consequences*. MIT Press.

Cooper, A. F., & Abrams, E. (2021). Emergent unfairness in algorithmic fairness-accuracy trade-off research. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21, New York, NY. https://doi.org/10.1145/3461702.3462519

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist Politics. *University of Chicago Legal Forum*, *1989*(1), 31.

Cribbs, J. D., Hazari, Z., Sonnert, G., & Sadler, P. M. (2015). Establishing an explanatory model for mathematics identity. *Child Development*, *86*(4), 1048–1062. https://doi.org/10.1111/cdev.12363

Cunningham, E. C. (1997). The rise of identity politics I: The myth of the protected class in Title VII disparate treatment cases. *Connecticut Law Review*, *30*(2), 441–502.

Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence*, *20*(2), 507–537. https://doi.org/10.1111/j.1532-7795.2010.00645.x

Geiser, S. (2015). The growing correlation between race and SAT scores: New findings from California. *UC Berkeley: Center for Studies in Higher Education*. https://escholarship.org/uc/item/9gs5v3pv

Gordon, C., & Gergen, K. J. (1968). Self-conceptions: Configurations of content. In *The Self in Social Interaction*. John Wiley & Sons, Inc.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 501–512. https://doi.org/10.1145/3351095.3372826

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58. https://doi.org/10.1145/3287560.3287600

Kasy, M., & Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–586. https://doi.org/10.1145/3442188.3445919

Khan, Z., & Fu, Y. (2021). One label, one billion faces: Usage and consistency of racial categories in computer vision. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 587–597. https://doi.org/10.1145/3442188.3445920

Kizilcec, R. F., & Lee, H. (Forthcoming). Algorithmic fairness in education. *Ethics in Artificial Intelligence in Education*. http://arxiv.org/abs/2007.05443

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2019). Delayed impact of fair machine learning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 6196–6200. https://doi.org/10.24963/ijcai.2019/862

M Kuhn, & McPartland, T. S. (1954). An empirical investigation of self-attitudes. *American Sociological Review*, *19*, 68–76.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Mouzannar, H., Ohannessian, M. I., & Srebro, N. (2019). From fair decision making To social equality. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 359–368. https://doi.org/10.1145/3287560.3287599

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. https://doi.org/10.1111/bjet.12156

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's learning? Using demographics in EDM research. *Journal of Educational Data Mining*, *12*(3), 1–30. https://doi.org/10.5281/zenodo.4143612

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249–255. https://doi.org/10.3758/BF03194060

Ritter, S., & Fancsali, S. E. (2016). MATHia X: The next generation cognitive tutor. *Proceedings of the EDM 2016 Workshops and Tutorials*, 2.

Roberto, K. R. (2011). Inflexible bodies: Metadata for transgender identities. *Journal of Information Ethics*, *20*(2), 56–64. https://doi.org/10.3172/JIE.20.2.56

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 99–106. https://doi.org/10.1145/3306618.3314248

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–35. https://doi.org/10.1145/3392866

Shah, D., Schwartz, H. A., & Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264. https://doi.org/10.18653/v1/2020.acl-main.468

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. J. Durlach & A. M. Lesgold (Eds.), *Adaptive Technologies for Training and Education* (pp. 7–27). Cambridge University Press. https://doi.org/10.1017/CBO9781139049580.004

Trahan, A. (2011). Qualitative research and intersectionality. *Critical Criminology*, *19*(1), 1–14. https://doi.org/10.1007/s10612-010-9101-0

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174230