Evaluating Intersectional Fairness in Algorithmic Decision Making Using Intersectional Differential Algorithmic Functioning

Youmi Suk [*1] and Kyung T. Han [†2]

[1]Teachers College, Columbia University
[2]Graduate Management Admission Council

June 11, 2024

Abstract

Ensuring fairness is crucial in developing modern algorithms and tests. To address potential biases and discrimination in algorithmic decision making, researchers have drawn insights from the test fairness literature, notably the work on *differential algorithmic functioning* (DAF) by Suk and Han (2024). Nevertheless, the exploration of intersectionality in fairness investigations, within both test fairness and algorithmic fairness fields, is still relatively new. In this paper, we propose an extension of the DAF framework to include the concept of intersectionality. Similar to DAF, the proposed notion for intersectionality, which we term "interactive DAF," leverages ideas from test fairness and algorithmic fairness. We also provide methods based on the generalized Mantel-Haenszel test, generalized logistic regression, and regularized group regression to detect DAF, interactive DAF, or other subtypes of DAF. Specifically, we employ regularized group regression with three different penalties and examine their performance via a simulation study. Finally, we demonstrate our intersectional DAF framework in real-world applications on grade retention and conditional cash transfer programs in education.

Keywords: Fairness, Intersectionality, Discrimination, Algorithms, Machine learning, Decision analysis, Differential item functioning, and Regularized regression

1 Motivation: Test Fairness and Intersectional Fairness

Over the past six decades, numerous methods for *differential item functioning* (DIF) have been developed and widely used to assess the test fairness and validity at the item level of test (e.g., Hanson, 1998; Holland & Wainer, 1993; Lim et al., 2022; Magis et al., 2011; Pine, 1977). Briefly, a test item is often a single question on an assessment test, and an item is considered to function differently (or DIF for short) if a subgroup of test-takers, such as the male group, perform differently on the item than another subgroup, such as the female group. For example, if the male group is more likely to answer a test item correctly than the female group given the same latent trait level (e.g., ability), the item is considered a DIF item and is typically suggested to

[*]ysuk@tc.columbia.edu
[†]truetheta@gmail.com
This article has been accepted for publication in *Journal of Educational and Behavioral Statistics*, published by SAGE Publishing.

be removed from the assessment. There is a long and rich literature on DIF in test development and psychometrics, including characterizing subtypes of DIF and developing powerful tests to detect DIF. Recently, Suk and Han (2024) extended DIF to create an alternative framework for assessing algorithmic fairness, which they call *differential algorithmic functioning* (DAF). The DAF framework uses a decision variable, a set of "fair" variables, and a protected variable such as race and gender, and addresses whether a decision functions differently across different subgroups of the protected variable. Suk and Han (2024) also characterize different subtypes of DAF and provide statistical tests to detect these subtypes of DAF; see the section on "Review: Differential Algorithmic Functioning" for more details. Unfortunately, a major limitation of the current DAF framework is that it does not take into account intersectionality. It is possible to test whether an algorithm is fair with respect to each protected attribute separately, but it is not feasible to test concurrent and compound injustices that arise across intersectional subgroups. The overarching goal of this paper is to extend the DAF framework to include the concept of intersectionality.

While discussions on test fairness have continued over the past six decades, the concept of algorithmic fairness recently emerged in the 2010s to identify and rectify unfair biases in machine learning and artificial intelligence systems (Barocas et al., 2019; Mitchell et al., 2021; Pessach & Shmueli, 2022). Definitions of modern-day algorithmic fairness are often similar or identical to earlier definitions of test fairness (Hutchinson & Mitchell, 2019). However, the exploration of intersectionality in fairness investigations within both fields is relatively recent. Russell and his colleagues have recently addressed intersectionality in the DIF literature (Russell & Kaplan, 2021; Russell et al., 2021; Russell et al., 2022). Additionally, a handful of researchers have examined intersectional fairness in the field of algorithmic fairness (Foulds et al., 2020; Hébert-Johnson et al., 2018; Kearns et al., 2018; M. P. Kim et al., 2019; Yang et al., 2021). The concept of intersectionality originally arose from the feminist movement that highlights the importance of considering sexism and racism simultaneously rather than separately (Crenshaw, 1989), and it is now understood as an analytical lens for investigating societal unfairness along overlapping dimensions including gender, race, class, and disability (Foulds et al., 2020).

As a tangible example of intersectionality issues, suppose we are interested in designing an algorithm to assist teachers' decisions to retain a student or not. We consider two protected variables, gender and race. In a working retention algorithm, it is possible to be fair for gender and race separately, whereas it can be unfair among intersectional subgroups determined by two protected variables (e.g., the black female group). Specifically, the working algorithm may produce a systematic disadvantage to black female students in ways that are more than the sum of those by being female and being black. Such discriminatory bias will be ignored unless intersectionality is taken into account in fairness investigations.

In this paper, we propose an extension of the DAF framework to include the concept of intersectionality. To address intersectionality, we introduce the notion of *interactive DAF*, which assesses whether an algorithm functions differently among subgroups of one protected variable after accounting for another protected variable. We also expand the capabilities of the existing three DIF methods with a multi-categorical protected variable so that they can work properly under the intersectional DAF framework. Specifically, we use the generalized Mantel-Haenszel test and the generalized logistic regression method, using a decision variable as the dependent variable. As the final method, we propose a regularized group regression approach to detect different types of DAF, including interactive DAF. We evaluate the performance of DAF detection methods through a simulation study. We also demonstrate our proposed framework in real-world applications about grade retention and conditional cash transfer programs. To the best of our knowledge, this paper is the first attempt to integrate algorithmic fairness and test fairness in the context of intersectionality.

The remainder of the paper is organized as follows. We first provide a brief review of the DAF framework and prior works on intersectionality in test fairness and algorithmic fairness. Next,

we discuss an extension of the DAF framework to include intersectionality and provide DAF detection methods. We then offer the designs and results of our simulation study to examine the performance of DAF methods. Subsequently, we demonstrate our approach for intersectional DAF in two empirical data in education. Finally, we provide our discussion and conclusions in the last section.

2 Review: Differential Algorithmic Functioning

Consider a classification algorithm that is trained using data with N study units, indexed by $i = 1, 2, \ldots, N$. Data of study unit i is composed of covariates $V_i \in \mathcal{V}$ and a binary outcome Y_i. The covariates V_i are divided into finite-dimensional protected variables G_i and finite-dimensional unprotected variables X_i, i.e., $V_i = (G_i, X_i)$. A decision variable $D_i \in \mathcal{D} = \{0, 1\}$ is determined according to a decision rule based on V_i, i.e., $\delta : \mathcal{V} \to \mathcal{D}$. The most common goal of a classification algorithm is to find a decision rule that makes correct decisions, $Y_i = D_i$.

We review the DAF framework (Suk & Han, 2024) for evaluating fairness in algorithmic decision making. DAF uses three pieces of information: a decision variable D, a set of "fair" variables W, and a protected variable G. Suk and Han (2024) define DAF as conditional dependence of algorithmic decision D and protected variable G given fair attribute W. The fair attribute W means a set of justifiable variables that are important and valid in decision-making processes. It is a function of unprotected variables X, i.e., $W = h(X)$, $h : \mathbb{R}^{p_x} \to \mathbb{R}^{p_w}$, and is determined based on domain knowledge. Formally, DAF is defined with one protected variable G as:

Definition 1 $Pr(D = 1|W, G = g) \neq Pr(D = 1|W, G = g'),\ \exists g \neq g'.$

In words, an algorithm has DAF if the probability of receiving the treatment decision differs across subgroups of a protected attribute (e.g., male vs. female) after accounting for the fair attribute; otherwise, the algorithm is non-DAF. One possible measure of the statistical relationship between Y and G given W is odds ratios, and the DAF, based on the odds ratio, can be expressed as:

$$\Delta(W) = \frac{Pr(D = 1|W, G = g)/\{1 - Pr(D = 1|W, G = g)\}}{Pr(D = 1|W, G = g')/\{1 - Pr(D = 1|W, G = g')\}} \neq 1,\ \exists g \neq g'. \tag{1}$$

Equation (1) is an alternative way of expressing that DAF exists, i.e., Y and G are conditionally dependent given W. Typically, the focal group ($G = g$) represents the subgroup expected to have a systemic disadvantage by the algorithm, whereas the reference group ($G = g'$) represents the subgroup expected to have an advantage. Importantly, DAF emphasizes procedural fairness rather than outcome fairness by incorporating the fair attribute and treating individuals with the identical fair attribute similarly.

Moreover, the DAF framework provides a detailed description of disparity patterns by defining different subtypes of DAF as:

Definition 2 *In the presence of DAF, uniform DAF exists if the statistical relationship (e.g., odds ratios) between D and G is constant for all levels of W; otherwise, DAF is classified as nonuniform DAF.*

For example, when a statistical relationship of interest between D and G is odds ratios, uniform DAF is defined to be present if $\Delta(W) = c \neq 1$ for all values of W where c is a constant but not 1. When an algorithm is uniform DAF, the algorithm consistently (dis)favors the focal group over the reference group across all the values of the fair attribute. An algorithm with uniform DAF shows *static disparity* with respect to decision allocations. On the other hand, nonuniform

DAF is a type of DAF that is not uniform DAF. For example, it occurs when the advantage a group receives from an algorithm varies depending on the fair attribute. This can lead an algorithm to provide more favorable decisions to the focal group within certain ranges of the fair attribute, while favoring the reference group within other ranges. Nonuniform DAF displays *dynamic disparity* in decision allocations.[1]

For detecting DAF, three DAF methods are available, which are modifications of three well-established DIF methods, namely Mantel-Haenszel test (Holland & Thayer, 1986), logistic regression (Swaminathan & Rogers, 1990), and residual-based DIF (Lim et al., 2022); see Suk and Han (2024) for more details on DAF detection methods with a binary protected variable.

3 Review: Intersectionality in Test Fairness and Algorithmic Fairness

An intersectional approach to DIF analyses in the field of test fairness was initially addressed in Russell and Kaplan (2021) and further discussed in Russell et al. (2021) and Russell et al. (2022). These works emphasize that intersectionality acknowledges the existence of multiple identities that shape individuals' lived experiences, and thus, forms intersectional groups based on multiple protected variables. Specifically, three variables of interest in their works are gender (male and female), racial stratification (white, black, Hispanic, and Asian), and economic status (economically advantaged and economically disadvantaged). One historically advantaged group, such as the "male-white-advantaged" group, is used as the reference group, while others are considered focal groups. Pairwise comparisons are then conducted between the reference group and each of the focal groups, using four DIF detection methods for a binary protected variable, namely the standardized D statistic (Dorans & Kulick, 1986), Mantel-Haenszel test, logistic regression, and the simultaneous item bias test (Shealy & Stout, 1993). In essence, their framework views intersectionality as the redefinition of multiple protected attributes into a single product category. This aligns with the issue of a multi-categorical protected variable in the DIF literature.

In the field of algorithmic fairness, intersectionality issues have gained increasing attention and are often discussed with formal and quantifiable definitions (Foulds et al., 2020; Hébert-Johnson et al., 2018; Kearns et al., 2018; M. P. Kim et al., 2019; Yang et al., 2021). Among these, we highlight two approaches that extend beyond merely creating intersectional groups: one related to privacy (Foulds et al., 2020) and another to causality (Yang et al., 2021). Specifically, Foulds et al. (2020) extend the concepts of 80% rule[2] or differential privacy (Dwork & Roth, 2013) to define intersectional fairness. Their proposed definition of *differential fairness* (DF) measures the probability ratios across different subgroups determined by all the protected variables, and it aims to make the ratios similar to achieve intersectional fairness. They also introduce two more notions based on DF, namely *DF bias amplification* and *differential fairness with confounders* (DFC). DF bias amplification measures a difference in DF between the original dataset and an algorithm based on the same dataset, quantifying the bias induced by the algorithm. DFC evaluates DF with confounders affecting outcome distributions; for more details, see Foulds et al. (2020). Moreover, Foulds et al. (2020) outline five important criteria for formally defining intersectional fairness of fairness in AI. We have applied these criteria to assess the validity of our proposed framework for intersectional fairness in the "Discussion and Conclusions" section.

[1]We remark that researchers have the flexibility to select a statistical relationship for the subtypes of DAF. They may choose a measure that is widely accepted in their field or one that best fits their data and hypotheses. It is also possible to utilize multiple relationships that are equally important in the DAF assessment and report the results for each one, if applicable.

[2]The 80% rule indicates legal evidence of adverse impact when the probability ratio of a favorable outcome between disadvantaged and advantaged groups is below 0.8.

In addition, Yang et al. (2021) introduce a causal modeling approach to intersectional fairness in ranking tasks, considering the intersecting dimensions of multiple protected variables. Their framework builds on *counterfactual fairness* (Kusner et al., 2017), a fairness framework based on causal inference. This framework assumes that a decision is fair towards an individual if it is the same between the actual world and in a counterfactual world where the protected attribute were different. Yang et al. (2021) adapt the framework to address intersectionality and ranking settings, where a ranking is counterfactually fair if it remains the same when comparing using actual protected variables to one counterfactual, intersectional, reference subgroup. Yang et al. (2021) further categorize the roles of unprotected, mediator variables into non-resolving, resolving, and semi-resolving types in causal models, and outline a procedure to compute counterfactually fair rankings considering these mediator types. While theoretically compelling, there are difficulties in implementing the framework, as it requires no unmeasured confounding and a correctly specified causal model.

Overall, these previous studies have explored intersectionality from various angles. Yet, they have not formally distinguished how the effects of intersectionality manifest, whether in additive or non-additive ways.

4 Our Proposal: Intersectional Differential Algorithmic Functioning

In this section, we have extended the DAF framework to include the concept of intersectionality across protected variables. Our intersectional DAF analysis involves more than two protected variables, creating intersectional subgroups determined by these protected variables. The effect of intersectionality across the combination of protected variables can often be additive or become more complex. To detect interactive effects among intersectional subgroups, we define interactive DAF, which measures whether an algorithm functions differently among subgroups of one protected variable after accounting for another protected variable. With at least two protected variables (e.g., G_1, G_2), interactive DAF is written as:

Definition 3 *Interactive DAF exists if the statistical relationship (e.g., relative risks) between a decision variable (D) and one protected variable varies depending on the levels of other protected variable(s), after accounting for a set of fair variables (W).*

For example, a statistical relationship of interest can be expressed using relative risks. In this case, interactive DAF is defined to exist if $\frac{Pr[D=1|W,G_1=g_1,G_2]}{Pr[D=1|W,G_1=g_1',G_2]} \neq \frac{Pr[D=1|W,G_1=g_1]}{Pr[D=1|W,G_1=g_1']}$. Another statistical relationship of interest can be odds ratios, and interactive DAF is detected if $\frac{Odds(W,G_1=g_1,G_2)}{Odds(W,G_1=g_1',G_2)} \neq \frac{Odds(W,G_1=g_1)}{Odds(W,G_1=g_1')}$, where $Odds(W,\cdot) \coloneqq \frac{Pr(D=1|W,\cdot)}{1-Pr(D=1|W,\cdot)}$. Moreover, if a statistical relationship of interest is on the difference scale, interactive DAF is present if $Pr[D=1|W,G_1=g_1,G_2] - Pr[D=1|W,G_1=g_1',G_2] \neq Pr[D=1|W,G_1=g_1] - Pr[D=1|W,G_1=g_1']$. Regardless, interactive DAF aims to detect distinctive effects by intersecting protected attributes, and it will be a useful concept to examine non-additive systematic (dis)advantages in algorithms across intersecting dimensions. In the absence of interactive DAF, additive effects across multiple protected attributes can be captured by examining at existing subtypes of DAF, i.e., uniform DAF and nonuniform DAF.

To characterize different subtypes of DAF in the intersectional framework, we extend the types of DAF discussed in Suk and Han (2024), which distinguishes between uniform DAF and nonuniform DAF. Specifically, we add another dimension to include interactive DAF. Table 1 provides a description of the allocation patterns associated with four types of DAF: (1) uniform (and simple) DAF, (2) nonuniform (and simple) DAF, (3) uniform and interactive DAF, and (4) nonuniform and interactive DAF. Simple DAF refers to unfair decision disparities associated with individual protected variables at a time or those not falling under the category of interactive

DAF. These classifications facilitate a more informative discussion of intersectionality across protected variables.

Table 1: Types of differential algorithmic functioning (DAF) and allocation patterns in the framework of intersectionality.

DAF Type	Simple DAF	Interactive DAF
Uniform DAF	Static disparity	Static, interactive disparity
Nonuniform DAF	Dynamic disparity	Dynamic, interactive disparity

Figure 1 provides illustrations of DAF types with an intersectional variable between gender and race. An intersectional protected variable represents a combination of protected variables, say $G^* = G_1 \times G_2$. For example, if gender (male vs. female) and race (white vs. black) are used as protected variables, intersectional variable G^* contains four levels: white male, white female, black male, and black female. In plot A of Figure 1, the algorithm shows uniform DAF because the odds ratios of the group advantages in an algorithm are constant across the fair attribute, and it shows *static disparity* in decision allocations. In contrast, nonuniform DAF exists in plot B because the group advantages vary depending on the fair attribute, showing *dynamic disparity*. Moreover, when an interactive effect is of interest, we can further inspect the presence of interactive DAF. For example, we observe nonuniform DAF in plot C since group advantages depend on the fair attribute, and in particular, only people who are both black and female are disadvantaged systematically with increasing W. That is, we observe both nonuniform DAF and interactive DAF in plot C, showing *dynamic, interactive disparity* in decision allocations.

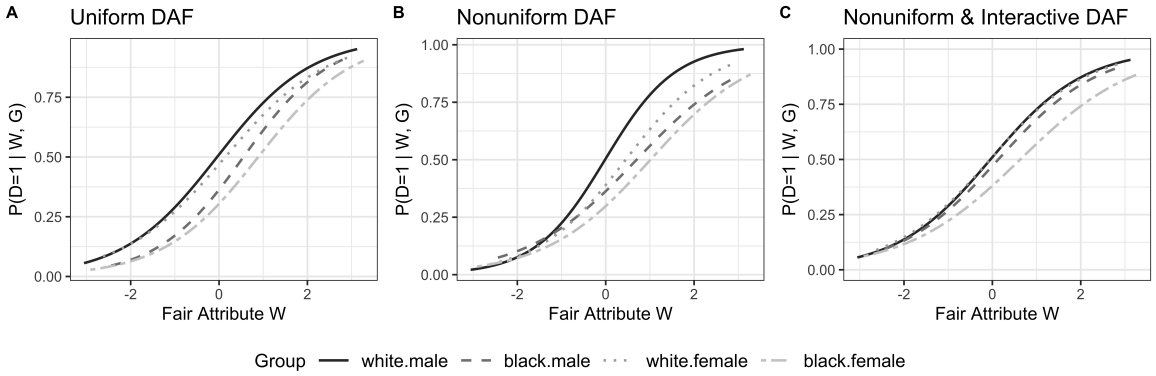

Figure 1: Illustrations of decision characteristic curves for intersectional groups by race (white vs. black) and gender (male vs. female) with different types of differential algorithmic functioning (DAF).

To detect the presence of DAF with intersectionality in algorithms, we utilize existing DIF methods that can handle multi-categorical protected variables and modify them properly. Such methods include the generalized Mantel-Haenszel test (GMH; Penfield, 2001; Somes, 1986), generalized logistic regression (GLR; Magis et al., 2011), generalized Lord's χ^2 test (S.-H. Kim et al., 1995), and most recently, regularized regression methods (Tutz & Schauberger, 2015; Wang et al., 2022). Although these DIF methods show their effectiveness in detecting DIF in test items for one protected variable at a time, they are not designed for discovering DAF coupled with intersectionality across protected variables, as defined in Definition 3. Therefore, we expand the capabilities of existing DIF methods so that they can work within the DAF framework under the context of intersectionality. In particular, we harness GMH, GLR, and regularized group regression methods because of their direct applicability. For simplicity, we

assume two protected variables $G = (G_1, G_2)$ and a one-dimensional fair attribute W in the observed data; see details for each method below.

4.1 Generalized Mantel-Haenszel

The GMH statistic is a multivariate generalization of the Mantel-Haenszel chi-square statistic that is used to analyze a multi-categorical protected variable (Holland & Thayer, 1986). We use the GMH statistic to test whether an algorithm has DAF or not across all intersectional subgroups simultaneously. Here, intersectional protected variable $G^* \in \{1, 2, \ldots, J\}$, which is either categorical or categorized, is used, and fair attribute W is discretized into K strata ($k = 1, 2, \ldots, K$). It is important to note that the GMH test does not provide separate test statistics to detect different subtypes of DAF (e.g., uniform DAF, interactive DAF), and it can only detect the composite effect of DAF.

Consider the data shown in Table 2 that contains treatment and control decisions ($D = 1$ and $D = 0$) from an algorithm for J categorical, intersectional groups. Let $A_k = (n_{11k}, n_{21k}, \ldots, n_{(J-1)1k})^\intercal$ be a $(J-1) \times 1$ vector containing the sample size of any $(J-1)$ groups who receive $D = 1$. $E(A_k)$ and $V(A_k)$ represent the expectation and variance-covariance matrix of A_k, respectively, and can be written as:

$$E(A_k) = \frac{n_{\cdot 1k} n_k}{n_{\cdot\cdot k}}, \quad V(A_k) = n_{\cdot 1k} n_{\cdot 0k} \frac{n_{\cdot\cdot k}\text{diag}(n_k) - n_k n_k^\intercal}{(n_{\cdot\cdot k} - 1) n_{\cdot\cdot k}^2},$$

where $n_k = (n_{1\cdot k}, n_{2\cdot k}, \ldots, n_{(J-1)\cdot k})$, and $\text{diag}(n_k)$ is a $(J-1) \times (J-1)$ diagonal matrix with elements n_k. The GMH statistic is given by

$$\chi^2_{GMH} = (A - E(A))^\intercal V(A)^{-1} (A - E(A)), \tag{2}$$

where $A = \sum_{k=1}^K A_k$, $E(A) = \sum_{k=1}^K E(A_k)$, $V(A) = \sum_{k=1}^K V(A_k)$. This statistic asymptotically follows a chi-squared null distribution with $(J-1)$ degree of freedom, under the null hypothesis that an algorithm is non-DAF. If the GMH statistic is significant, it supports that the algorithm has DAF.

Table 2: A contingency table by intersectional protected variable G^* and decision variable D within the k-th stratum of fair attribute W.

	Treatment Decision ($D = 1$)	Control Decision ($D = 0$)	Total
Group 1	n_{11k}	n_{10k}	$n_{1\cdot k}$
Group 2	n_{21k}	n_{20k}	$n_{2\cdot k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group J	n_{J1k}	n_{J0k}	$n_{J\cdot k}$
Total	$n_{\cdot 1k}$	$n_{\cdot 0k}$	$n_{\cdot\cdot k}$

4.2 Generalized Logistic Regression

The GLR method is based on Magis et al. (2011) and Swaminathan and Rogers (1990) and it requires fitting a logistic regression model that regresses D on fair attribute W and protected variable G. Let us denote π_i as the probability that individual i from intersectional group $G_i^* = g$ receives $D = 1$, where $g = 1, 2, ..., J-1$ for the first, second, ..., $J-1$ focal group, respectively, and $g = J$ for the reference group. Let $\widetilde{G}^* \in \mathbb{R}^{n \times (J-1)}$ be an $n \times (J-1)$ dummy

matrix indicating individuals' intersectional protected membership in one of the $(J-1)$ focal groups and consisting of $(J-1)$ dimensional binary vector variables where the g-th element is 1 and the rest are 0. A logistic model for the GLR method with W_i and $\widetilde{G}_i^*$ can be written as:

$$\text{logit}(\pi_i) = [1, W_i]\beta_0 + \widetilde{G}_i^*\beta_1 + W_i\widetilde{G}_i^*\beta_2, \tag{3}$$

where $\beta_0 = (\beta_{00}, \beta_{01})^\intercal$ represents the intercept and slope parameters of the reference group $g = J$; $\beta_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1,(J-1)})^\intercal$ is a coefficient vector that denotes differences in the intercept between each focal group $g = 1, \ldots, J-1$ and the reference group; $\beta_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2,(J-1)})^\intercal$ is a coefficient vector that indicates differences in the slope between each focal group and the reference group. Model (3) serves as the full model, and using (3), we can detect DAF, uniform DAF, and nonuniform DAF. The null and alternative hypotheses for each case are provided in Table 3.

Table 3: Null and alternative hypotheses in generalized logistic regression

	H_0		H_A	
DAF	$\beta_1 = 0$	and $\beta_2 = 0$	$\beta_1 \neq 0$	or $\beta_2 \neq 0$
Uniform DAF	$\beta_1 = 0$	under $\beta_2 = 0$	$\beta_1 \neq 0$	under $\beta_2 = 0$
Nonuniform DAF	$\beta_2 = 0$		$\beta_2 \neq 0$	

Note: H_0 and H_A represent null and alternative hypotheses, respectively. 0 in the null and alternative hypotheses represents coefficient vectors of zeros.

We use likelihood-ratio test statistics between the respective null models and respective alternative models to detect DAF, uniform DAF, and nonuniform DAF. We denote the three test statistics as GLR, GLR_{UNI}, and GLR_{NUNI}, respectively.

Additionally, let $\widetilde{G}_1$ be an $n \times (J_1 - 1)$ dummy matrix indicating individuals' protected group membership in one focal level of the first protected variable G_1; likewise, $\widetilde{G}_2$ be an $n \times (J_2 - 1)$ dummy matrix indicating individuals' protected group membership in one focal level of the second protected variable G_2. To analyze interactive DAF, we fit the following model:

$$\text{logit}(\pi_i) = [1, W_i]\beta_0 + [\widetilde{G}_{1i}, \widetilde{G}_{2i}]\beta_1' + W_i[\widetilde{G}_{1i}, \widetilde{G}_{2i}]\beta_2' + [\widetilde{I}_i, W_i\widetilde{I}_i]\beta_3. \tag{4}$$

Here, $\widetilde{I}_i = \widetilde{G}_{1i} \otimes \widetilde{G}_{2i}$ is the resulting matrix that captures the interaction terms between $\widetilde{G}_{1i}$ and $\widetilde{G}_{2i}$ where $\otimes$ represents the Kronecker product; $\beta_1' = (\beta_{11}', \beta_{12}' \ldots, \beta_{1,(J_1+J_2-2)}')^\intercal$ is a coefficient vector that indicates intercept differences for each focal level of one protected variable when another protected variable is fixed at its reference level; $\beta_2' = (\beta_{21}', \beta_{22}', \ldots, \beta_{2,(J_1+J_2-2)}')^\intercal$ is a coefficient vector that indicates slope differences for each focal level of one protected variable when another protected variable is fixed at its reference level; $\beta_3 = (\beta_{31}, \beta_{32}, \ldots, \beta_{3,2(J_1-1)(J_2-1)})^\intercal$ is a coefficient vector that represents interactive effects across different protected variables. Based on Model (4), we can detect the presence of interactive DAF by testing the null hypothesis that an algorithm has no interactive DAF, i.e., $H_0 : \beta_3 = 0$. We use a likelihood-ratio test to detect interactive DAF, and the test statistic is denoted as GLR_{INT}. If GLR_{INT} is significant, it supports the existence of interactive DAF in the algorithm. By using GLR_{INT}, we can examine whether the DAF effect among intersectional subgroups is interactive or not, and it provides another dimension to describe the DAF effect. We remark that alternatively, it is possible to use a contrast matrix based on Model (3) to detect interactive DAF. However, fitting Model (4) allows us to directly test interactive DAF as defined in Definition 3.

4.3 Regularized Group Regression

We propose an approach based on regularized group regression to detect the composite effect of DAF and specific types of DAF, including uniform DAF, nonuniform DAF, and interactive DAF.

This approach was inspired by the work of Tutz and Schauberger (2015) on group lasso penalty and Wang et al. (2022) on (ungrouped) lasso and adaptive lasso. We refer to our proposed approach as regularized group differential algorithmic functioning (rgDAF). Specifically, the rgDAF method groups regression coefficients responsible for DAF and employs three different penalties: lasso, the smoothly clipped absolute deviation (SCAD; Fan & Li, 2001), and the minimax concave penalty (MCP; Zhang, 2010). An important distinction between an ungrouped penalty and a group penalty is that a group penalty selects coefficient groups, not individual coefficient parameters within the group; that is, within a group, coefficients will either all be zero or all be non-zero.

Let $\beta_{\mathcal{M}}$ represent the set of coefficient parameters in a fitting model $\mathcal{M}$, e.g., $\beta_{\text{Model 3}} = (\beta_0, \beta_1, \beta_2)$ in Model (3) and $\beta_{\text{Model 4}} = (\beta_0, \beta_1', \beta_2', \beta_3)$ in Model (4). The objective function for regularized group regression optimization is written as:

$$Q(\beta_{\mathcal{M}}; \lambda) = L(\beta_{\mathcal{M}}) + \sum_{s=1}^{S} p(\beta_{(s)}; \lambda), \tag{5}$$

where the loss function $L(\beta_{\mathcal{M}})$ is the negative log-likelihood of a binomial distribution, i.e., $L(\beta_{\mathcal{M}}) = -\frac{1}{n}\sum_i log\mathbb{P}(y_i|\eta_i)$, with η_i being unit i's linear prediction in model $\mathcal{M}$. The penalty term $\sum_s p(\beta_{(s)}; \lambda)$ contains only the parameters responsible for detecting specific types of DAF, forming S different coefficient groups ($s = 1, \ldots, S$). Specifically, the group lasso penalty term is given by:

$$p(\beta_{(s)}; \lambda) = \lambda\sqrt{d_s}\|\beta_{(s)}\|. \tag{6}$$

Here, the tuning parameter $\lambda \geq 0$ controls the penalty size. The presence of $\sqrt{d_s}$ in the penalty term accounts for different sizes of coefficient groups, and normalizes the penalty impact across groups of different sizes; we denote $\lambda\sqrt{d_s} \coloneqq \lambda_s$. The group lasso penalty specifically applies a lasso penalty to the Euclidean (L_2) norm of each coefficient group (denoted as $\|\beta_{(s)}\|$), and thus, encourages sparsity and variable selection at the group level. That is, the solution of the penalized parameters has the property that if coefficient group s is selected, then all individual coefficients within coefficient group s are included; otherwise, $\beta_{(s)} = 0$ for all s. Additionally, the rgDAF method requires choosing the tuning parameter λ. We considered several criteria for selecting the tuning parameter, such as cross-validation, the Akaike information criterion (AIC), and the Bayesian Information Criterion (BIC). Following Tutz and Schauberger (2015), Belzak and Bauer (2020), and our investigations via simulations, we use BIC to choose the tuning parameter.

Table 4 describes the specific groups of penalized coefficients $\beta_{\mathcal{M}}^{P}$ used in our rgDAF method to detect DAF and its specific types. We group the first-order terms together as one coefficient group, while each second-order or higher term constitutes its own respective coefficient group. Specifically, for detecting DAF, we have J coefficient groups, where the first coefficient group $\beta_{(1)} = \beta_1$ with $d_1 = J - 1$, and the last coefficient group $\beta_{(J)} = \beta_{2,(J-1)}$ with $d_{J-1} = 1$. To detect uniform DAF, only one coefficient group of β_1 exists. In contrast, the number of coefficient groups for detecting nonuniform DAF is $J - 1$, where the first coefficient group $\beta_{(1)} = \beta_{21}$ and the last coefficient group $\beta_{(J-1)} = \beta_{2,(J-1)}$, with $d_s = 1$ for all coefficient groups. To detect interactive DAF, we create $2(J_1 - 1)(J_2 - 1)$ coefficient groups, and each coefficient element in β_3 forms a separate group, as for nonuniform DAF. We note that one could use a different specification for grouping regression coefficients if it is chosen reasonably.[3]

[3]Regarding reasonable choices for grouping, we need to consider two key factors. First, only the parameters that are responsible for DAF should be penalized and grouped. Second, if they are grouped, all coefficients associated with a single protected variable should be grouped together within the same degree of polynomials.

Table 4: Group specification of coefficients responsible for differential algorithmic functioning (DAF)

DAF Type	Number of Groups, S	Group Specification in $\beta^P_{\mathcal{M}}$	Model $\mathcal{M}$
DAF	J	$(\beta_1, \beta_{21}, \beta_{22}, \ldots, \beta_{2,(J-1)})$	(3)
Uniform DAF	1	(β_1)	$\beta_2 = 0$ in (3)
Nonuniform DAF	$J-1$	$(\beta_{21}, \beta_{22}, \ldots, \beta_{2,(J-1)})$	(3)
Interactive DAF	$2(J_1-1)(J_2-1)$	$(\beta_{31}, \beta_{32}, \ldots, \beta_{3,2(J_1-1)(J_2-1)})$	(4)

Note: The comma inside the parenthesis for the group lasso penalty serves as a separator for different groups.

For estimating $\beta_{(s)}$, we use the multivariate soft-thresholding operator within the group coordinate descent algorithm (Breheny & Huang, 2013). The soft-thresholding operator (Donoho & Johnstone, 1994) is a mathematical function that operates on a vector z by reducing its magnitude towards 0 by an amount λ while preserving its direction. Formally, it is written as $F(z, \lambda) = \left(1 - \frac{\lambda}{||z||}\right)_+ z$ where $\frac{z}{||z||}$ is the unit vector in the direction of z (Yuan & Lin, 2005). Algorithm 1 summarizes the rgDAF method for detecting the composite effect of DAF based on the coordinate descent algorithm for group lasso logistic regression. In Algorithm 1, Q represents the design matrix in a fitting model, and $Q_{(s)}$ is the portion of the design matrix associated with $\beta_{(s)}$. Briefly, the group descent algorithm optimizes the target function with respect to a single coefficient group at a time and iterates through the coefficient groups until convergence. Then, we assess whether the coefficients responsible for DAF are zero or not based on the estimated coefficients. If any of the grouped, penalized coefficients are non-zero, it provides evidence for DAF. Likewise, to detect subtypes of DAF, we use an appropriate set of penalized coefficients and adjust a fitting model in Algorithm 1 as summarized in Table 4. For example, in the case of interactive DAF, if any coefficient in $(\beta_{31}, \beta_{32}, \ldots, \beta_{3,2(J_1-1)(J_2-1)})$ does not shrink to 0, it supports the presence of interactive DAF.

Algorithm 1 Regularized group differential algorithmic functioning (rgDAF) for detecting DAF based on the group coordinate descent algorithm

Input: Decision D_i, fair attribute W_i, and protected variable G_i

1: Set $m = 0$. Initialize vector of residuals $r^{(m)} = (y - \pi)/v$ for all individuals, where $\pi_i = \frac{exp(\eta_i)}{1+exp(\eta_i)}$, $\eta_i = Q_i\beta^{(m)}$, $Q_i = [1, W_i, \widetilde{G}^*_i, W_i\widetilde{G}^*_i]$, $v = \max_i \sup_\eta\{\nabla^2 L_i(\eta)\} \leq 1/4$.

2: For $s = 1, 2, \ldots, S$, carry out the following calculations:

(i) Calculate $z_{(s)} = Q^\intercal_{(s)} r^{(m)} + \beta^{(m)}_{(s)}$

(ii) Update $\beta^{(m+1)}_{(s)} \leftarrow F_{gLasso}(vz_{(s)}, \lambda_s)/v$

(iii) Update $r^{(m+1)} \leftarrow r^{(m)} - Q^\intercal_{(s)}(\beta^{(m+1)}_{(s)} - \beta^{(m)}_{(s)})$

3: Update $m \leftarrow m + 1$.

4: Repeat Steps 2 and 3 until convergence to obtain the final estimate, $\hat{\beta}_{(s)}$.

5: Detect the presence of DAF; if any of $\hat{\beta}_{(s)}$ is not zero, DAF exists.

Output: $\hat{\beta}_{(s)}$ and the presence of DAF

Additionally, we incorporate two additional penalties, SCAD and MCP in (5). These two penalties aim to alleviate lasso bias by reducing the strength of the penalty on large estimates in absolute value (Hastie et al., 2017). Between SCAD and MCP, MCP immediately relaxes the penalization rate, while with SCAD, the rate remains constant for a while before decreasing. To obtain the group SCAD and group MCP solutions using the group descent algorithm, we replace $F_{gLasso}(vz_{(s)}, \lambda_s)/v$ in Algorithm 1 with corresponding variations for SCAD and MCP;

see Supplemental Appendix A for more details on these penalty functions and solutions.

4.4 Summary of DAF Detection Methods

We provide a summary of test statistics from three methods—GMH, GLR, and rgDAF—to detect different types of DAF in Table 5. Based on our simulation study below and the section on "Our Proposal: Intersectional Differential Algorithmic Functioning," we outline the strengths and limitations of DAF detection methods. The GMH test is a non-parametric test that does not require statistical modeling, which ensures valid Type-1 error control even when the relationship between D, G, and W is complex. As we will demonstrate in simulations, however, the GMH test has low power to detect DAF when nonuniform DAF is present, and it is not capable of differentiating different types of DAF: uniform DAF, nonuniform DAF, and interactive DAF. On the other hand, the tests from the GLR method have the power to detect different types of DAF. The asymptotic properties of these GLR tests are only valid if the fitting models are correctly specified, and if misspecified, it can inflate Type-1 error rates. Similar to GLR, the rgDAF method can detect subtypes of DAF but might suffer from model misspecification. Between the GLR and rgDAF methods, the rgDAF method becomes more advantageous when dealing with a larger number of protected attributes, especially if many of the focal groups within these protected attributes exhibit no differences from their respective references. This is because lasso regression is generally more advantageous when dealing with a larger number of predictors, especially when many of them are irrelevant or redundant.

Table 5: Test statistics from three methods that detect differential algorithmic functioning (DAF) with intersectionality.

DAF Type	GMH	GLR	rgDAF
DAF	GMH	GLR	$rgDAF$
Uniform DAF		GLR_{UNI}	$rgDAF_{\text{UNI}}$
Nonuniform DAF		GLR_{NUNI}	$rgDAF_{\text{NUNI}}$
Interactive DAF		GLR_{INT}	$rgDAF_{\text{INT}}$

Note. GMH = generalized Mantel-Haenszel; GLR = generalized logistic regression; rgDAF = regularized group differential algorithmic functioning

Lastly, we emphasize that while technical solutions provide valuable insights into the overall presence of DAF, visual inspection is highly recommended. We specify our strategies using graphical tools in Supplemental Appendix B and present specific examples in the section on "Conditional Cash Transfer Programs."

5 Simulation Study

5.1 Designs and Evaluation

We conduct a simulation study to assess the performance of GMH, GLR, and rgDAF methods with nine test statistics: one from the GMH method (GMH), four from the GLR method (GLR, GLR_{UNI}, GLR_{NUNI}, and GLR_{INT}), and four from the rgDAF method ($rgDAF$, $rgDAF_{\text{UNI}}$, $rgDAF_{\text{NUNI}}$, and $rgDAF_{\text{INT}}$). As a comparison, we include the Pearson chi-square test statistic (denoted as $Pearson$) for the statistical parity criterion. Note that statistical/demographic parity requires that an algorithm's decision be independent of group membership, i.e., $Pr(D = 1|G = g) = Pr(D = 1)$ (Feldman et al., 2015), and it aims to achieve equality of outcome rather than procedural fairness.

Our simulation study is categorized into five designs; see Figure 2 for illustrations of our simulation designs. Design 1 assumes non-DAF (e.g., $\beta_0 \neq 0$, $\beta_1' = 0$, $\beta_2' = 0$, and $\beta_3 = 0$ in model (4)), and Design 2 assumes uniform DAF (e.g., $\beta_0 \neq 0$, $\beta_1' \neq 0$, $\beta_2' = 0$, and $\beta_3 = 0$ in model (4)). Design 3 assumes "balanced" nonuniform DAF where the advantages of intersectional groups are balanced across the levels of the fair attribute (e.g., $\beta_0 \neq 0$, $\beta_1' = 0$, $\beta_2' \neq 0$, and $\beta_3 = 0$ in model (4)). Design 4 assumes "unbalanced" nonuniform DAF where the group advantages are not balanced across the fair attribute (e.g., $\beta_0 \neq 0$, $\beta_1' \neq 0$, $\beta_2' \neq 0$, and $\beta_3 = 0$ in model (4)). Our last design, Design 5, is based on Design 4, but additionally assumes interactive DAF (e.g., $\beta_0 \neq 0$ and $\beta_3 \neq 0$ in model (4)). More specifically, our data-generating model was based on logistic regression model (4) with one fair attribute and two binary protected variables, where non-zero coefficients responsible for specific DAF effects were set to -0.45 on a logit scale. We also varied the sample size, where intersectional subgroup size n_g was set to either 500 or 1000.

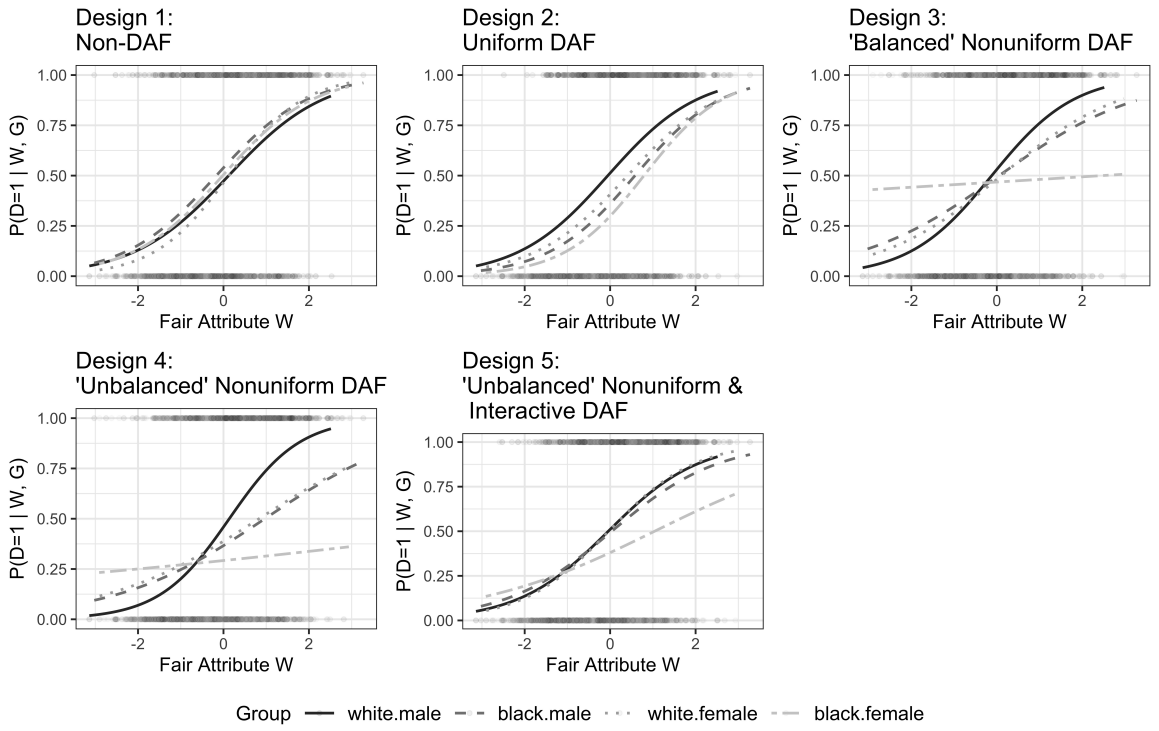

Figure 2: Simulation Designs

For all the designs, we evaluated the performance of nine DAF test statistics (as well as the Pearson statistic for statistical parity) by measuring Type-1 error and power rates. GMH and GLR statistics were assessed at the 5% alpha level to determine whether they provide evidence for DAF. The average detection rates represent the power rates when a test statistic is specifically designed for detecting a particular type of DAF; otherwise, they represent the Type-1 error rates. We repeated the simulation with five designs 1,000 times. For the rgDAF method implementation, we used the R package *grpreg* (Breheny & Huang, 2013).

5.2 Results

Table 6 summarizes the results of the simulation study. In Design 1 that assumes non-DAF with a sample size of $n_g = 500$, the DAF statistics from GMH and GLR methods, as well as the Pearson chi-square statistic ($Pearson$), have well-controlled Type-1 error rates of ≤ 0.06. Test statistics from rgDAF methods exhibit very small Type-1 errors across different penalty terms,

which is desirable. In Design 2, with uniform DAF and the sample size of $n_g = 500$, all the DAF methods perform as expected. The DAF statistics that specialize in detecting uniform DAF (i.e., GLR_{UNI} and $rgDAF_{\text{UNI}}$) or any type of DAF (i.e., GMH, GLR, and $rgDAF$) show high power rates (> 0.9). The DAF statistic for nonuniform DAF from the GLR method (GLR_{NUNI}) has a well-controlled Type-1 error rate of < 0.05, but the one for interactive DAF (GLR_{INT}) shows a slightly inflated Type-1 error rate of 0.07. Additionally, rgDAF statistics for nonuniform DAF and interactive DAF exhibit very small Type-1 error rates of < 0.02 across different penalty terms.

For Design 3, with balanced nonuniform DAF and $n_g = 500$, the GLR and rgDAF methods perform well; the test statistics for detecting nonuniform DAF (GLR_{NUNI} and $rgDAF_{\text{NUNI}}$) or any type of DAF (GLR and $rgDAF$) show high power rates (> 0.9). In contrast, the GMH method fails to detect the presence of DAF because it is not able to accurately detect DAF when the group advantages are canceled out across the levels of the fair attribute. The Pearson statistic for statistical parity also shows low retention rates because it is anticipated to have no marginal difference under the balanced nonuniform design. In Design 4 which assumes unbalanced nonuniform DAF and $n_g = 500$, all test statistics perform as anticipated. That is, test statistics except for GLR_{INT} and $rgDAF_{\text{INT}}$ show high detection rates (i.e., > 0.9 power rates), and GLR_{INT} and $rgDAF_{\text{INT}}$ with three different penalties show low detection rates (i.e., < 0.05 Type-1 error rates).

In Design 5 which assumes both unbalanced nonuniform DAF and interactive DAF with $n_g = 500$, all the average detection rates from GMH, GLR, and Pearson statistics are high, ranging from 79.7% to 99.7%. Importantly, GLR_{INT} performs well in detecting the presence of interactive DAF. Regarding rgDAF methods, unlike other designs, we observe noticeable performance differences across different penalty terms. The rgDAF statistic with group lasso has detection rates lower than about 5% compared to group SCAD and MCP. This may be because the (group) lasso tends to overshrink large coefficients in absolute value compared to the other two (Fan & Li, 2001; Huang et al., 2012), and it could lead to lower detection rates for $rgDAF_{\text{INT}}$ with lasso penalty. Also, we find that all $rgDAF_{\text{INT}}$ results have much lower detection rates than GLR_{INT} under the sample size condition of 500. This lower power of rgDAF method may be because sparsity only applies to a small set of coefficients responsible for interactive DAF (i.e., β_3), while all the other coefficients remain un-penalized in the current data generating models. This specification could lead the regularized method to readily have the penalized coefficients shrunk to zero, in order to reduce model variance at the expense of a small increase in bias. Additionally, the lower detection rates of $rgDAF_{\text{INT}}$ may be explained in part by our simple data-generating model with only one fair attribute and two binary-protected variables.

When the sample size of $n_g = 1000$ increases, similar patterns are found in most statistics; desirable Type-1 errors and power rates are observed in Designs 1, 2, 3, 4, and 5. A noticeable difference is the increasing power rates of $rgDAF_{\text{INT}}$ compared to the sample size of $n_g = 500$, but the detection rates of group MCP or group SCAD for interactive DAF are still higher than that from group lasso. Given these findings, we recommend implementing the rgDAF method with group MCP or group SCAD rather than group lasso.

6 Real Data Analysis

In this section, we illustrate our DAF framework with two real datasets: one about grade retention and the other about conditional cash transfer programs. The former dataset is used to test the presence of DAF in the development of a new algorithm with the aforementioned test statistics, while the latter is used to diagnose an existing algorithm and highlight the importance of visual inspection for detecting DAF.

Table 6: Average detection rates under Designs 1, 2, 3, 4, and 5

	Penalty	Design 1	Design 2	Design 3	Design 4	Design 5
$n_g = 500$						
$Pearson$		0.057	1.000	0.036	1.000	0.943
GMH		0.054	**1.000**	**0.045**	**1.000**	**0.960**
GLR		0.060	**1.000**	**1.000**	**1.000**	**0.997**
GLR_{UNI}		0.060	**1.000**	0.047	**1.000**	**0.965**
GLR_{NUNI}		0.053	0.047	**1.000**	**1.000**	**0.859**
GLR_{INT}		0.049	0.070	0.041	0.035	**0.797**
$rgDAF$	Group lasso	0.028	**1.000**	**0.998**	**1.000**	**0.975**
$rgDAF_{\text{UNI}}$	Group lasso	0.004	**1.000**	0.004	**1.000**	**0.800**
$rgDAF_{\text{NUNI}}$	Group lasso	0.027	0.017	**0.999**	**0.992**	**0.829**
$rgDAF_{\text{INT}}$	Group lasso	0.007	0.011	0.007	0.005	**0.437**
$rgDAF$	Group SCAD	0.030	**1.000**	**0.999**	**1.000**	**0.975**
$rgDAF_{\text{UNI}}$	Group SCAD	0.004	**1.000**	0.004	**1.000**	**0.800**
$rgDAF_{\text{NUNI}}$	Group SCAD	0.027	0.017	**0.999**	**0.992**	**0.829**
$rgDAF_{\text{INT}}$	Group SCAD	0.007	0.012	0.010	0.004	**0.484**
$rgDAF$	Group MCP	0.031	**0.999**	**0.999**	**1.000**	**0.974**
$rgDAF_{\text{UNI}}$	Group MCP	0.003	**0.999**	0.003	**1.000**	**0.785**
$rgDAF_{\text{NUNI}}$	Group MCP	0.027	0.017	**0.999**	**0.992**	**0.838**
$rgDAF_{\text{INT}}$	Group MCP	0.007	0.012	0.010	0.004	**0.484**
$n_g = 1000$						
$Pearson$		0.045	1.000	0.043	1.000	0.998
GMH		0.039	**1.000**	**0.052**	**1.000**	**0.999**
GLR		0.039	**1.000**	**1.000**	**1.000**	**1.000**
GLR_{UNI}		0.037	**1.000**	0.049	**1.000**	**0.999**
GLR_{NUNI}		0.052	0.049	**1.000**	**1.000**	**0.996**
GLR_{INT}		0.049	0.041	0.046	0.043	**0.978**
$rgDAF$	Group lasso	0.019	**1.000**	**1.000**	**1.000**	**1.000**
$rgDAF_{\text{UNI}}$	Group lasso	0.000	**1.000**	0.003	**1.000**	**0.993**
$rgDAF_{\text{NUNI}}$	Group lasso	0.021	0.017	**1.000**	**1.000**	**0.990**
$rgDAF_{\text{INT}}$	Group lasso	0.009	0.007	0.009	0.006	**0.782**
$rgDAF$	Group SCAD	0.020	**1.000**	**1.000**	**1.000**	**1.000**
$rgDAF_{\text{UNI}}$	Group SCAD	0.000	**1.000**	0.003	**1.000**	**0.993**
$rgDAF_{\text{NUNI}}$	Group SCAD	0.022	0.017	**1.000**	**1.000**	**0.990**
$rgDAF_{\text{INT}}$	Group SCAD	0.010	0.006	0.011	0.007	**0.829**
$rgDAF$	Group MCP	0.020	**1.000**	**1.000**	**1.000**	**1.000**
$rgDAF_{\text{UNI}}$	Group MCP	0.000	**1.000**	0.003	**1.000**	**0.992**
$rgDAF_{\text{NUNI}}$	Group MCP	0.021	0.018	**1.000**	**1.000**	**0.990**
$rgDAF_{\text{INT}}$	Group MCP	0.010	0.006	0.010	0.007	**0.828**

Note: Design 1 assumes non-DAF; Design 2 assumes uniform DAF; Design 3 assumes balanced nonuniform DAF; Design 4 assumes unbalanced nonuniform DAF; Design 5 assumes unbalanced nonuniform DAF and interactive DAF. Bold fonts indicate the power for DAF, and regular fonts indicate the Type-1 error.

6.1 Grade Retention

We used data from the Early Childhood Longitudinal Study-Kindergarten cohort (ECLS-K) (Walston & McCarroll, 2010) to design a retention decision-making algorithm. The ECLS-K is a national longitudinal study that examines school achievement and student experiences from

kindergarten to middle school, and it is sponsored by the National Center for Education Statistics. We used students' retention in grade 1 as the outcome of interest Y, where $Y = 0$ means that they moved onto grade 1 and $Y = 1$ means that they were retained (i.e., they repeated kindergarten). We selected 60 students' covariates in the kindergarten year (i.e., V) that are expected to affect whether a student is retained or not to design a decision-making algorithm for retention based on prior works (e.g., Cannon & Lipscomb, 2011; Hong & Raudenbush, 2006). We used intersectional groups determined by gender (male vs. female) and race (white vs. black) as protected variables of interest. The sample sizes for each intersectional subgroup are as follows: 3,627 white male students, 670 black male students, 3,391 white female students, and 663 black female students. Fair variables contained prior achievement scores in math, reading, and general knowledge, all measured during the kindergarten year; we chose them as fair variables because grade retention is often determined based on test-based scores according to prior work (e.g., Penfield, 2010). We constructed one summary variable using factor analysis with the prior achievement scores to alleviate multicollinearity and effectively capture the latent construct of ability.

To make algorithm-based decisions, we fitted random forests (Breiman, 2001) using 60 covariates as predictors, and we made a student's predictions P from the model. Then, we used a set of threshold values that ranged from 0.20 to 0.30 with an increment of 0.02 to decide whether to retain or promote a student; we specifically used a range of threshold values below the average estimated retention probability (i.e., 0.039) in our analysis to account for the small number of retained students reported in prior works (e.g., Hong & Raudenbush, 2006; Suk & Han, 2024; Zill et al., 1997) and observed in our sample. For instance, we made our decision based on a threshold value of 0.20 as: $D = I(P \geq 0.20)$. Here, $D = 0$ indicates promoting them to grade 1, and $D = 1$ indicates retaining them in kindergarten. Using the aforementioned three DAF methods (i.e., GMH, GLR, and rgDAF with group MCP), we assessed the presence of DAF among intersectional groups in the working retention algorithm.

Figure 3 summarizes the results of the test statistics, where satisfying a fairness criterion is denoted as 0, and unsatisfying is denoted as 1. Our intersectional DAF analysis reveals that the evidence of DAF depends on the threshold values and DAF test statistics. The $Pearson$ statistic detects marginal unfairness across all the threshold values. The GMH, GLR, and $rgDAF$ statistics detect DAF for all the threshold values. When looking at statistics for subtypes of DAF from GLR and rgDAF methods, GLR_{NUNI} detects nonuniform DAF when the threshold value is larger than or equal to 0.26, whereas $rgDAF_{\text{NUNI}}$ detects nonuniform DAF when the threshold value is at 0.20 or larger than 0.26; otherwise, uniform DAF is detected via GLR_{UNI} and $rgDAF_{\text{UNI}}$ statistics. Note that if there is evidence of nonuniform DAF, the DAF subtype is considered nonuniform DAF, regardless of the presence of uniform DAF. Regarding interactive DAF, we observe no interactive DAF from the results of GLR_{INT} across all the threshold values, while interactive DAF is detected at some of the threshold values from $rgDAF_{\text{INT}}$. Overall, our analysis suggests that the working algorithm may potentially lead to discriminatory bias among intersectional groups while also presenting some conflicting evidence for interactive DAF. Therefore, the algorithm should be revised by adjusting statistical models, altering the algorithm procedure, or modifying the variables used, to ensure it is DAF-free.

6.2 Conditional Cash Transfer Programs

We demonstrate the application of our intersectional DAF framework on an existing recommendation algorithm developed by Suk and Park (2023). The authors developed data-driven, optimal recommendations for individual students regarding conditional cash transfer programs using data from Colombia (Barrera-Osorio et al., 2019; Barrera-Osorio et al., 2011), but they did not incorporate fairness-related considerations in their model development. To detect the presence of DAF in Suk and Park (2023)'s model, we used the program recommendation status

from one of the best-fitting models as the decision variable of interest (D); $D = 1$ represents a recommendation for the program and $D = 0$ represents no recommendation. We also used Colombia's poverty index, known as the SISBEN, where a lower value indicates lower socioeconomic status for individuals. We selected the SISBEN index as the fair attribute because it is widely recognized and utilized in Colombia as a measure for identifying beneficiaries of social programs (Castañeda & Fernandez, 2005). Two protected variables of interest are (i) a student's gender and (ii) the age gap between a student's age and the typical age for their grade. The gender variable is binary (male vs. female), and the age-gap variable is measured on two different scales: binary and continuous. The binary age-gap variable indicates whether a student is older than the typical age in their grade (older vs. not older), and the continuous age-gap variable indicates the number of years older the student is for their grade (min = -14, max = 52, mean=0.26).

We first analyze intersectionality with gender and a binary age-gap variable, and in this case, the intersectional subgroups contain 1443 not-older male students, 437 older male students, 1603 not-older female students, and 389 older female students. Before conducting a DAF analysis, we checked the marginal differences in decisions among the four subgroups. We observe that 59.4% of the students in the older female group are recommended to receive the program, but in other groups, about 75% of those are recommended; specifically, 73.9% are recommended among the not-older female group, 75.3% among the older male group, and 74.7% among the not-older male group.

Figure 4 provides decision characteristic curves among the four groups determined by gender and the binary age-gap variable, similar to Figure 1. In Figure 4, we observe that the decision probabilities for female students who are older than the typical age in their grade are similar to those among other groups around the minimum of the fair attribute, but they rapidly decrease as the fair attribute of SISBEN increases. In contrast, the decision probabilities for the other

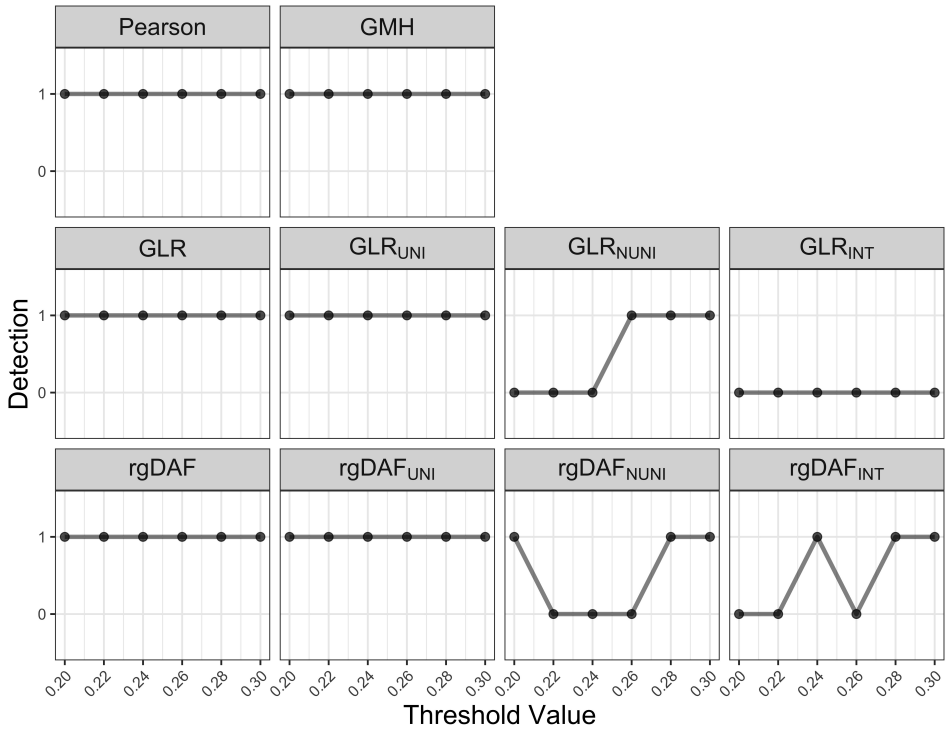

Figure 3: Results of test statistics from fairness metrics about differential algorithmic functioning (DAF) and statistical parity with an intersectional group variable.

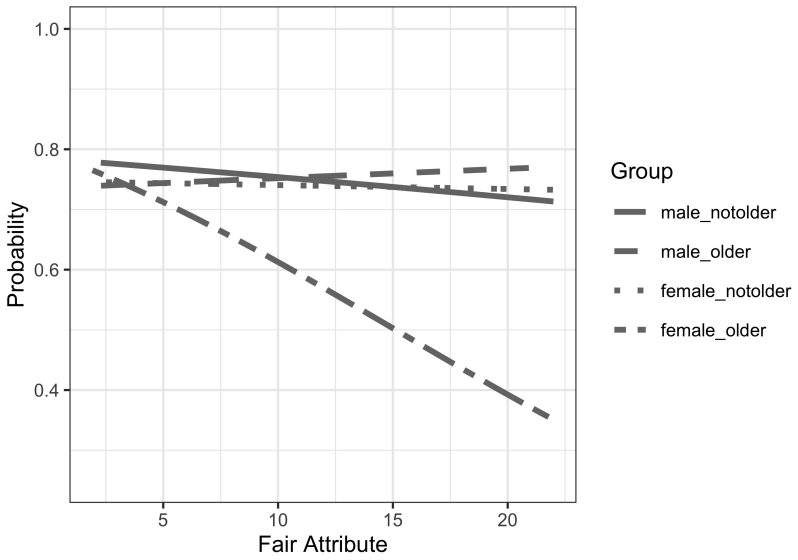

Figure 4: Decision characteristic curves for intersectional groups by gender (male vs. female) and age gap (older vs. not older).

three groups are relatively stable over the fair attribute. This indicates that both nonuniform DAF and interactive DAF are present in the existing algorithm. Using the GLR and rgDAF methods, we also find evidence of both nonuniform DAF and interactive DAF, based on the test statistics that specialize in detecting nonuniform DAF (i.e., GLR_{NUNI} and $rgDAF_{\text{NUNI}}$) and interactive DAF (i.e., GLR_{INT} and $rgDAF_{\text{INT}}$).

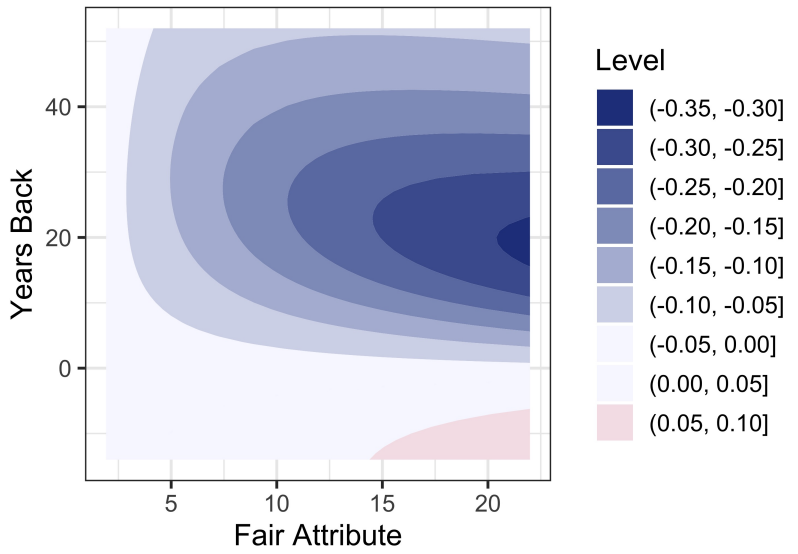

Figure 5: Differences in decision probabilities between female and male students.

To examine the intersectionality between gender and age-gap variables in more detail, we used the continuous age-gap variable that represents the number of years older the child is for their grade (denoted as Years Back). Using the continuous protected variable, we can inspect detailed patterns of intersectionality, compared to its binary version used in Figure 4. Figure 5 describes how differences in decision probabilities between female and male students change over

the fair attribute and Years-Back variable through a contour plot. Positive values indicate that female students have higher probabilities than male students, but negative values indicate that male students have higher probabilities than female students. We see that there are no or very small differences (highlighted in white) around the minimum of the fair attribute and around the zero value of Years-Back over the fair attribute. In contrast, when the fair attribute goes to the maximum, differences in decision probabilities diverge depending on the values of Years-Back. When we look at fair-attribute values of 20 or above, differences in the probabilities are the most positive near the minimum of Years-Back (highlighted in light pink), whereas differences are the most negative when Years-Back amounts to about 20 (highlighted in dark blue). These results imply that the algorithm discriminates against students impacted by age and gender in ways that are more than additive.

7 Discussion and Conclusions

In this paper, we have extended the DAF framework for algorithmic fairness to incorporate the concept of intersectionality across protected variables. We introduced the notion of interactive DAF to identify how protected variables interact with each other, highlighting the importance of considering their intersecting dimensions in algorithmic decision making. To test the presence of DAF within the intersectional DAF framework, we proposed three DAF detection tools: GMH, GLR, and rgDAF methods. Among them, the GLR and rgDAF methods can identify different types of DAF, including uniform DAF, nonuniform DAF, and interactive DAF. Our simulation study revealed that the performance of the rgDAF method for interactive DAF is underwhelming when the sample size of one intersectional subgroup is 500. However, with an increasing sample size of 1000, its performance improves. When comparing the performance of the rgDAF method with different penalties, we found that group MCP and group SCAD are preferred over group lasso. We believe that the rgDAF method is particularly useful when dealing with a larger number of protected attributes, especially when many of the focal groups within these attributes do not differ significantly from their respective references. Additionally, we emphasized the significance of visual inspection in understanding disparity patterns in decision allocations, and we demonstrated our framework using two real datasets on grade retention and conditional cash transfer programs; the former dataset was evaluated in the development of a new algorithm, while the latter was used to diagnose an existing algorithm.

Importantly, our framework of intersectional DAF satisfies the intersectional fairness criteria proposed by Foulds et al. (2020), which include (i) considering multiple protected variables, (ii) defining and protecting the intersecting values of the protected variables, (iii) protecting individual protected variables, (iv) protecting minority groups, and (v) considering structural oppression among protected groups. Our framework achieves criteria (i) and (ii) by considering intersectional groups across different protected variables. Criterion (iii) is met because any subtypes of DAF should not be present to ensure a DAF-free algorithm, and this guarantees the protection of individual protected variables. We do not introduce any penalty against minority groups, such as the weights of the group sample size, thus satisfying criterion (iv). Additionally, criterion (v) is satisfied because the use of fair attributes aims to consider the important and valid decision-making process and can account for structural oppression among protected groups.

Moreover, the intersectional DAF framework depends on the choices of fair attributes as discussed in Suk and Han (2024). Unlike DIF analysis, where the latent ability score is used as a fair attribute, DAF analysis allows fair attributes to be either manifest or latent. When selecting fair attributes in DAF analysis, researchers should leverage subject matter knowledge about factors behind the decision-making process to identify which variables qualify as fair attributes. For example, in our real-data application of the grade retention algorithm, there are two approaches to determining grade retention: test-based retention and teacher-based retention (Huddleston, 2014). In this demonstration, we adopt the test-based approach, where prior achievement scores

are considered key and valid variables in the decision-making process for grade retention. Thus, we use prior achievement scores as a fair attribute, given our knowledge about the test-based retention decision-making process and the role of the prior achievement scores. However, in cases where subject matter knowledge is limited, researchers might consider data-driven measures, such as those based on changes in R-squared, Gini index, or classification accuracy, as potential candidates for the fair attributes among unprotected variables. Nonetheless, the validity and reliability of fair attributes should be critically evaluated based on subject matter expertise, involving various stakeholders in the design, implementation, and use of algorithms. Failure to do so may result in the selection of inappropriate fair attributes and/or the use of variables with substantive measurement errors, potentially drawing misleading conclusions in the subsequent DAF analysis.

Based on the findings of this paper, we provide some suggestions for future research regarding our proposed intersectional DAF framework. First, while we focused on parametric approaches to detect subtypes of DAF, future work would investigate nonparametric techniques for detecting them, particularly interactive DAF. Second, we employed regularized group regression, where penalized coefficients are grouped in specific ways to detect DAF or its subtypes. Future research would examine exploring different specifications for grouping regression coefficients in the method to compare their performance. Third, we did not focus on the impact of measurement bias in variables (e.g., fair attributes or outcomes) on DAF analysis. While this source of bias potentially influences the results of our DAF analysis, more systematic research would be required to investigate its impact in the DAF assessment. Fourth, it is essential to recognize that algorithms flagged as DAF may only have the potential to be unfair. Thus, addressing fairness-related biases in algorithms requires a holistic approach that combines both technical and non-technical solutions. Lastly, our framework of intersectional DAF has the potential to be adapted to DIF settings in test development, and this enables the detection of "interactive DIF" in addition to the existing subtypes of DIF. Future research would explore how to incorporate interactive DIF within the context of test fairness.

While no universal definition is suitable for all systems and contexts, our intersectional DAF framework focuses on achieving the intersectional fairness of decision allocations and sheds light on different patterns of decision disparities across multiple protected variables. We believe that our intersectional DAF framework will serve as a valuable tool for assessing fairness in algorithmic decision-making and can be viewed as a meaningful integration between the concepts of test fairness and algorithmic fairness.

Acknowledgements

The authors are grateful to Dubravka Svetina Valdivia and Montserrat Valdivia Medinaceli for their comments on an earlier version of the manuscript, shared during the 2024 National Council for Measurement in Education (NCME) conference. This research was partly funded by the National Science Foundation under Grant No. 2225321. The opinions, findings, conclusions, or recommendations expressed in this work are solely those of the authors and do not necessarily reflect the views of the National Science Foundation. Also, the original collector of the data, ICPSR, and the relevant funding agency bear no responsibility for use of the data or for interpretations or inferences based upon such uses.

References

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org. http://www.fairmlbook.org

Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2019). *Replication data for: Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia* (tech. rep.). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E113783V1

Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2011). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in colombia. *American Economic Journal: Applied Economics*, *3*(2), 167–195. https://doi.org/10.1257/app.3.2.167

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673–690. https://doi.org/10.1037/met0000253

Breheny, P., & Huang, J. (2013). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*(2), 173–187. https://doi.org/10.1007/s11222-013-9424-2

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Cannon, J. S., & Lipscomb, S. (2011). *Early grade retention and student success: Evidence from Los Angeles*. Public Policy Institute of California.

Castañeda, T., & Fernandez, L. (2005). Targeting social spending to the poor with proxy-means testing: Colombia's sisben system. *World bank human Development Network social protection unit discussion paper*, *529*.

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*, 139–167.

Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*(3), 425–455. https://doi.org/10.1093/biomet/81.3.425

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Dwork, C., & Roth, A. (2013). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3-4), 211–407. https://doi.org/10.1561/0400000042

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268. https://doi.org/10.1145/2783258.2783311

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1918–1921. https://doi.org/10.1109/icde48307.2020.00203

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics*, *23*(3), 244–253. https://doi.org/10.2307/1165247

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. https://doi.org/https://doi.org/10.48550/arXiv.1707.08692

Hébert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. *2018 International Conference on Machine Learning (ICML)*, 1939–1948.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203357811

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, *1986*(2), i–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910. https://doi.org/10.1198/016214506000000447

Huang, J., Breheny, P., & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, *27*(4). https://doi.org/10.1214/12-sts392

Huddleston, A. P. (2014). Achievement at whose expense? a literature review of test-bas ed grade retention policies in us schools. *Education Policy Analysis Archives*, *22*(18), 1–31. https://doi.org/http://dx.doi.org/10.14507/epaa.v22n18.2014

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

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *2018 International conference on machine learning (ICML)*, 2564–2572.

Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. *2019 AAAI/ACM Conference on AI, Ethics, and Society*, 247–254.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*(3), 261–276. https://doi.org/10.1111/j.1745-3984.1995.tb00466.x

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *2017 Conference on Neural Information Processing Systems (NIPS)*. https://doi.org/30

Lim, H., Choe, E. M., & Han, K. T. (2022). A residual-based differential item functioning detection framework in item response theory. *Journal of Educational Measurement*, *59*(1), 80–104. https://doi.org/10.1111/jedm.12313

Magis, D., Raıche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, *11*(4), 365–386. https://doi.org/10.1080/15305058.2011.602810

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*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*(3), 235–259. https://doi.org/10.1207/s15324818ame1403_3

Penfield, R. D. (2010). Test-based grade retention: Does it stand up to professional standards for fair and appropriate test use? *Educational Researcher*, *39*(2), 110–119. https://doi.org/https://doi.org/10.3102/0013189X10363007

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1–44. https://doi.org/10.1145/3494672

Pine, S. M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of military testing association*

(pp. 37–43). University of Minnesota, Department of Psychology, Psychometric Methods Program.

Russell, M., & Kaplan, L. (2021). An intersectional approach to differential item functioning: Reflecting configurations of inequality. *Practical Assessment, Research, and Evaluation*, *26*(1), 21. https://doi.org/https://doi.org/10.7275/20614854

Russell, M., Szendey, O., & Kaplan, L. (2021). An intersectional approach to DIF: Do initial findings hold across tests? *Educational Assessment*, *26*(4), 284–298. https://doi.org/10.1080/10627197.2021.1965473

Russell, M., Szendey, O., & Li, Z. (2022). An intersectional approach to DIF: Comparing outcomes across methods. *Educational Assessment*, *27*(2), 115–135. https://doi.org/10.1080/10627197.2022.2094757

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/bf02294572

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, *40*(2), 106–108. https://doi.org/10.2307/2684866

Suk, Y., & Han, K. T. (2024). A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning. *Journal of Educational and Behavioral Statistics*, (2), 107699862311717. https://doi.org/10.3102/10769986231171711

Suk, Y., & Park, C. (2023). Designing optimal, data-driven policies from multisite randomized trials. *Psychometrika*, *88*(4), 1171–1196. https://doi.org/10.1007/s11336-023-09937-2

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika*, *80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

Walston, J., & McCarroll, J. C. (2010). Eighth-grade algebra: Findings from the eighth-grade round of the early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K). statistics in brief. NCES 2010-016. *National Center for Education Statistics*.

Wang, C., Zhu, R., & Xu, G. (2022). Using lasso and adaptive lasso to identify DIF in multidimensional 2PL models. *Multivariate Behavioral Research*, *58*(2), 387–407. https://doi.org/10.1080/00273171.2021.1985950

Yang, K., Loftus, J. R., & Stoyanovich, J. (2021). Causal intersectionality and fair ranking. *2021 Symposium on Foundations of Responsible Computing*. https://doi.org/10.4230/LIPIcs.FORC.2021.7

Yuan, M., & Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2). https://doi.org/10.1214/09-aos729

Zill, N., Loomis, L. S., & West, J. *The elementary school perfor mance and adjustment of children who enter kindergarten late or repeat kindergarten: Findings from national surveys (statistical analysis report NCES 98-097)*. 1997. http://www.rand.org/pubs/technical_reports/TR678/

A Penalty Functions and Solutions for group SCAD and group MCP

Among various nonconvex penalties, the smoothly clipped absolute deviations (SCAD; Fan & Li, 2001) penalty, denoted as p^{SCAD}, is one of the earliest and most important. The SCAD penalty and its derivative are written as:

$$p^{SCAD}(\theta;\lambda,\gamma)=\begin{cases}\lambda\theta, & \text{if } |\theta|\le\lambda\\ \frac{\gamma\lambda|\theta|-0.5(\theta^2+\lambda^2)}{\gamma-1}, & \text{if } \lambda<|\theta|\le\gamma\lambda\\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}, & \text{if } |\theta|>\gamma\lambda\end{cases}, \quad \dot{p}^{SCAD}(\theta;\lambda,\gamma)=\begin{cases}\lambda, & \text{if } |\theta|\le\lambda\\ \frac{\gamma\lambda-|\theta|}{\gamma-1}, & \text{if } \lambda<|\theta|\le\gamma\lambda\\ 0, & \text{if } |\theta|>\gamma\lambda\end{cases} \tag{7}$$

Here, tuning parameter λ represents the penalty size, and tuning parameter γ controls the concavity of the penalty, which determines how rapidly the penalty tapers off; for SCAD, $\gamma>2$. SCAD behaves like the lasso until $|\theta|=\lambda$, and then it smoothly transitions to a quadratic function until $|\theta|=\gamma\lambda$. Beyond this point, it remains constant for all $|\theta|>\gamma\lambda$. The SCAD penalty maintains the penalization rate of the lasso for small coefficients but continuously relaxes the rate of penalization as the absolute value of the coefficient increases.

The minimax concave penalty (MCP; Zhang, 2010) follows a similar principle. The MCP penalty and its derivative are written as:

$$p^{MCP}(\theta;\lambda,\gamma)=\begin{cases}\lambda|\theta|-\frac{\theta^2}{2\gamma}, & \text{if } |\theta|\le\gamma\lambda\\ \frac{1}{2}\gamma\lambda^2, & \text{if } |\theta|>\gamma\lambda\end{cases}, \quad \dot{p}^{MCP}(\theta;\lambda,\gamma)=\begin{cases}(\lambda-\frac{|\theta|}{\gamma})\text{sign}(\theta), & \text{if } |\theta|\le\gamma\lambda\\ 0, & \text{if } |\theta|>\gamma\lambda\end{cases} \tag{8}$$

where $\gamma>1$. Similar to SCAD, MCP starts with the same rate of penalization as the lasso and then gradually reduces the penalization rate to zero as the absolute value of the coefficient increases. However, unlike SCAD, MCP immediately relaxes the penalization rate, whereas SCAD maintains it for a while before decreasing.

The close-form solutions for updating coefficients under group SCAD and MCP are as follows:

$$F_{gSCAD}(vz_{(s)},\lambda_s,\gamma)/v=\begin{cases}\frac{1}{v}S(vz_{(s)},\lambda_s), & \text{if } \|z_{(s)}\|\le\lambda_s\\ \frac{\gamma-1}{\gamma-2}\cdot\frac{1}{v}S(vz_{(s)};\frac{\gamma\lambda}{\gamma-1}), & \text{if } \lambda_s<\|z_{(s)}\|\le\gamma\lambda_s\\ z_{(s)} & \text{if } \|z_{(s)}\|>\gamma\lambda_s\end{cases}, \tag{9}$$

$$F_{gMCP}(vz_{(s)},\lambda_s,\gamma)/v=\begin{cases}\frac{\gamma}{\gamma-1}\cdot\frac{1}{v}S(vz_{(s)};\lambda_s), & \text{if } \|z_{(s)}\|\le\gamma\lambda_s\\ z_{(s)}, & \text{if } \|z_{(s)}\|>\gamma\lambda_s.\end{cases} \tag{10}$$

While Algorithm 1 is presented for the rgDAF method with group lasso, it can be readily adapted to fit group MCP and group SCAD models. Specifically, one can replace $F_{gLasso}(vz_{(s)},\lambda_s,\gamma)/v$ with $F_{gSCAD}(vz_{(s)},\lambda_s,\gamma)/v$ for SCAD and $F_{gMCP}(vz_{(s)},\lambda_s,\gamma)/v$ for MCP, respectively, to implement the rgDAF method with group SCAD and group MCP.

B Visual Inspection

Although tests of statistical significance in the section on "Our Proposal: Intersectional Differential Algorithmic Functioning" offer a useful indicator of the overall presence of DAF, conducting visual inspection is still strongly advised. Visual inspection is the process of examining DAF patterns using graphs/plots for data visualization, without relying on significance tests. This can provide an effective means to extract additional information and critical insights about the nature of DAF that might have been completely undetected or overlooked in the statistical significance testing. To find disparity patterns in algorithms, as seen from Figure 1, we can draw *decision characteristic curves* in a two-dimensional plot, in particular when intersectional subgroups are categorical or categorized. Decision characteristic curves assist researchers in understanding how the decision probabilities (or transformation thereof) change over the fair attribute among intersectional subgroups.

We can also use heatmaps or contour plots. In DAF analysis, heatmaps or contour plots display the 3-dimensional relationship in two dimensions, where the fair attribute and one protected variable (G_1) are plotted on the x- and y-scales and decision probabilities (or transformation thereof) are represented by colors or contours within groups of another protected variable (G_2). A contour line is a curve that joins points of equal values in decision probabilities. Or, one can create heatmaps or contour plots that represent differences in decision probabilities (or transformation thereof) between each focal group and the reference group of G_2 by colors or contours.

Similar to heatmaps and contour plots, surface plots display the 3-dimensional relationship, but they are in three dimensions with decision probabilities represented by a smooth surface on the z-scale. Instead of decision probabilities, one can also put differences in decision probabilities between groups on the z-scale in surface plots. Using these visualizations helps researchers discover discriminatory bias from algorithms and identify disparity patterns without relying solely on significance tests. Even when the significance tests do not indicate any presence of DAF, visual inspection can sometimes reveal localized DAFs (DAF exhibiting only in isolated spaces of fair attributes), which can still be consequential for the corresponding group of people and call for fairness-related attention. Therefore, visual inspection should be routinely used in DAF analysis.