# A Psychometric Framework for Evaluating Fairness in Algorithmic Decision Making: Differential Algorithmic Functioning

Youmi Suk [*1] and Kyung T. Han [†2]

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

April 8, 2023

**Abstract**

As algorithmic decision making is increasingly deployed in every walk of life, many researchers have raised concerns about fairness-related bias from such algorithms. But there is little research on harnessing psychometric methods to uncover potential discriminatory bias inside decision-making algorithms. The main goal of this paper is to propose a new framework for algorithmic fairness based on *differential item functioning* (DIF), which has been commonly used to measure item fairness in psychometrics. Our fairness notion, which we call *differential algorithmic functioning* (DAF), is defined based on three pieces of information: a decision variable, a "fair" variable, and a protected variable such as race or gender. Under the DAF framework, an algorithm can exhibit uniform DAF, nonuniform DAF, or neither (i.e., non-DAF). For detecting DAF, we provide modifications of well-established DIF methods: Mantel-Haenszel test, logistic regression, and residual-based DIF. We demonstrate our framework through a real dataset concerning decision-making algorithms for grade retention in K-12 education in the United States.

## 1 Introduction

As algorithmic decision making is increasingly used in every walk of life (e.g., hiring, lending, online advertising, online learning, criminal justice), many researchers have raised fairness-related concerns with such algorithms (Corbett-Davies & Goel, 2018). A significant concern surrounding automated decision-making algorithms is that they may produce unconscious bias in decision making against vulnerable subgroups (Mehrabi et al., 2021). That is, an algorithm may give an unfair advantage to one subgroup, such as whites, over another subgroup, such as blacks. Detecting and resolving such fairness-related biases in machine learning algorithms have motivated an area of research now called *algorithmic fairness* (Mitchell et al., 2021). Broadly speaking, this rapidly emerging literature provides formal, quantifiable measures of fairness, such as statistical parity and separation, which can be used either to diagnose existing algorithms or to inform designs of new algorithms (Barocas et al., 2019; Corbett-Davies et al., 2017; Dwork

[*]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.

et al., 2012; Feldman et al., 2015; Mitchell et al., 2021; Pessach & Shmueli, 2022); see details in Section 2. The overarching goal of this paper is to propose a new theoretical and practical framework for evaluating algorithmic fairness based on *differential item functioning* (DIF), which has been commonly used to measure item bias in test development and psychometrics.

Although discussions on algorithmic fairness are recent, the concept of *test fairness* was formulated in the 1960s and has evolved over the past six decades. Psychometricians have developed DIF frameworks and relevant methods to measure the fairness and validity of tests at the item level (Angoff & Sharon, 1974; Cleary & Hilton, 1968; Crocker & Algina, 1986; Pine, 1977). Briefly, an item is considered to exhibit DIF if the item behaves or functions differently across groups of examinees (often, focal versus reference groups) after accounting for examinees' ability; see equation (5) for a formal definition. Group categories considered in DIF often include gender, race, or ethnicity in the social equality context, and common methods to detect the presence of DIF include the Mantel-Haenszel test (Shealy & Stout, 1993), logistic regression (Swaminathan & Rogers, 1990), and item-response-theory-based methods, e.g., area measures (Kim & Cohen, 1991; Raju, 1988) and residual-based DIF (Lim et al., 2022). When an item is detected to have DIF, the item is typically reviewed by content experts for revision or removal from the item pool. DIF analysis is an essential component of standardized tests developed in the United States, such as the SAT, ACT, Graduate Record Examinations (GRE), and Graduate Management Admissions Test (GMAT). But, to the best of our knowledge, there is no research on harnessing the concept of DIF and its detection methods to uncover potential fairness-related harms in modern, automated algorithms.

In this paper, we propose a DIF-based approach to assess algorithmic fairness in modern, machine learning algorithms. In a nutshell, our approach, which we call *differential algorithmic functioning* (DAF), expands existing DIF to encompass algorithmic fairness based on three pieces of information typically available from a modern algorithm: (i) a decision variable, (ii) a "fair" variable, and (iii) a protected variable such as race, ethnicity, or gender. With these pieces of information, an algorithm can exhibit what we call uniform DAF, nonuniform DAF, or neither (i.e., non-DAF, fair). We also modify existing DIF detection methods, notably the Mantel-Haenszel test, logistic regression, and residual-based DIF, to assess the presence of DAF in algorithms; see Section 4 for details of the proposed DAF framework.

Throughout the manuscript, we use an example concerning student grade retention where an automated algorithm assists teachers' decisions on whether a student is retained or promoted. Typically, grade retention is recommended if students make inadequate progress in academic achievement or show developmental immaturity (Greene & Winters, 2006; Jackson, 1975). We measure the fairness of such decision-making algorithms using DAF and compare DAF to other notions of algorithmic fairness, notably statistical parity.

The remainder of this paper is organized as follows. Sections 2 and 3 briefly review algorithmic fairness and DIF, respectively. Section 4 discusses our DAF framework. Section 5 shows the empirical results about designing a new, fair algorithm that assists teachers' decisions to retain a student or not. Discussion and conclusions are in Section 6.

# 2 Review: Algorithmic Fairness

## 2.1 Notation

Suppose we have a classification algorithm that is trained on data from $N$ study units, indexed by $i = 1, 2, \ldots, N$. Each study unit's data consists of features/covariates $V_i \in \mathcal{V}$ and a binary outcome $Y_i$. For evaluating algorithmic fairness, the covariates are partitioned into protected variables $G_i$ and unprotected variables $X_i$, i.e., $V_i = (G_i, X_i)$. We define a decision rule $\delta : \mathcal{V} \to \{0, 1\}$ which takes on two possible actions based on $V_i$, i.e., $D_i = \delta(V_i)$. If a correct decision is made, $Y_i = D_i$. The goal for a classification algorithm is to find a decision rule that makes

correct decisions.

For example, in the case of grade retention, $Y$ would be a student's retention status where 0 indicates that he/she/they were promoted and 1 indicates that he/she/they were retained. $V$ would be a student's characteristics in the kindergarten year, and $D$ would be the algorithm's decision to retain or promote a student based on $V$ where 0 corresponds to promoting him/her/them to the next grade and 1 corresponds to retaining him/her/them in the same grade.

## 2.2 Notions of Algorithmic Fairness

We review four common measures of algorithmic fairness: (i) statistical parity, (ii) conditional statistical parity, (iii) separation, and (iv) sufficiency (Barocas et al., 2019; Corbett-Davies et al., 2017; Feldman et al., 2015; Mitchell et al., 2021). *Statistical parity* requires that an algorithm's decision be independent of protected group membership, and in the case of binary classification, it is defined as:

$$Pr(D = 1|G = g) = Pr(D = 1|G = g'). \tag{1}$$

In our retention example, statistical parity means that retention decision rates are equal across sub-populations such as gender or racial groups. Statistical parity is often referred to as demographic parity, disparate impact, or independence (Barocas et al., 2019; Corbett-Davies et al., 2017; Mitchell et al., 2021). Statistical parity pursues equality of outcomes/results and does not account for intrinsic characteristics of each individual, which may ultimately decrease the overall prediction performance (e.g., accuracy, recall) of the algorithm for all groups (Xu et al., 2022). For example, among racial groups, if black students were more likely to be retained than white students during kindergarten for some reasons, it would be reasonable to consider the actual racial differences for the retention predictions in the kindergarten year. But statistical parity may prevent an algorithm from reflecting this intrinsic difference.

*Conditional statistical parity* requires that an algorithm's decision be independent of protected group membership after controlling for a set of "legitimate" risk factors $L = l(X)$ (Corbett-Davies et al., 2017). Formally, it is defined as:

$$Pr(D = 1|L, G = g) = Pr(D = 1|L, G = g') \tag{2}$$

This notion aims to treat people who are similar in their legitimate risk factors similarly regardless of group membership. For example, among students who had the same developmental immaturity, black and white students are retained at equal rates. It should be noted that achieving conditional statistical parity does not always guarantee statistical parity (and vice versa), in particular if legitimate risk factors are correlated with protected variables.

*Separation* requires that an algorithm's decision should be independent of protected group membership conditional on the outcome. Formally, it is defined as:

$$Pr(D = 1|Y = y, G = g) = Pr(D = 1|Y = y, G = g'), \quad y \in \{0, 1\}. \tag{3}$$

Here, $Pr(D = 1|Y = 1, G = g)$ represents the true positive rate among group $g$ and $Pr(D = 1|Y = 0, G = g)$ represents the false positive rate among group $g$. Separation is also called error rate balance or equalized odds (Chouldechova, 2017; Hardt et al., 2016), and there are relaxed versions of separation, say satisfying equation (3) only with $y = 1$ (or $y = 0$) (Barocas et al., 2019). Separation would treat people with the same outcomes similarly regardless of group membership. In our retention example, separation is satisfied when black students and white students have the same false positive rates.

*Sufficiency* requires that the outcome be independent of the group conditional on the algorithmic decision:

$$Pr(Y = 1|D = d, G = g) = Pr(Y = 1|D = d, G = g'), \quad d \in \{0, 1\} \tag{4}$$

Here, $Pr(Y = 1|D = 1, G = g)$ represents the positive predictive value among group $g$ and $Pr(Y = 1|D = 0, G = g)$ represents the false discovery rate among group $g$. For example, a retention algorithm will satisfy sufficiency when black and white students who are recommended retention are actually retained at the same rate.

We also make some general remarks about existing fairness criteria in the literature. First, it is impossible to satisfy all the fairness notions simultaneously because some are inherently in conflict; see Chouldechova (2017), Berk et al. (2021), and Kleinberg et al. (2017). Second, there is no consensus as to what notion of fairness should be used in each context. Instead, researchers need to select fairness notion(s) that are the most appropriate in their own context (Xu et al., 2022).

# 3 Review: Differential Item Functioning

DIF has been widely used to detect items that exhibit discriminatory bias in assessments (Lim et al., 2022). DIF refers to different functioning of items across different groups of examinees (Holland & Wainer, 1993), and it typically means the difference in the probabilities of endorsing an item between groups conditional on ability (Magis et al., 2010; Pine, 1977), i.e.,

$$Pr(Y = 1|\theta, G = g) \neq Pr(Y = 1|\theta, G = g'), \tag{5}$$

Here, $Y$ represents whether an examinee's response to the test item is correct (i.e., $Y = 1$) or incorrect (i.e., $Y = 0$); $\theta$ represents ability scores, and $G$ represents group membership that consists of a focal group (i.e., $G = g$) and a reference group (i.e., $G = g'$). In the DIF literature, the focal group represents the particular group of interest who is expected to be disadvantaged by the test, whereas the reference group represents the group who is expected to have an advantage (Holland & Wainer, 1993). Typically, test developers investigate the presence of DIF to create a test where the performance of examinees is only affected by their abilities and not by other factors like examinees' demographics (Ackerman, 1992). They assume that if there exists DIF, the item discriminates the examinees mainly (or partially) based on their group membership (Holland & Wainer, 1993).

There are two common types of DIF: uniform DIF and nonuniform DIF (Mellenbergh, 1982; Swaminathan & Rogers, 1990). A test item exhibits uniform DIF when the item is always more advantageous to one group (e.g., whites) than another group (e.g., blacks), showing a higher probability of correctly answering the item at any ability level. In contrast, a test item exhibits nonuniform DIF when the advantage in the item depends on ability level, and it often results in an interaction between ability and group membership.

A wide array of statistical methods have been developed to evaluate the presence and impact of DIF. They can be categorized into two streams depending on whether the methods rely on item response theory (IRT) or not (Magis et al., 2010). Non-IRT-based methods match examinees based on their test scores, and some of the most popular methods include the Mantel-Haenszel test (Holland & Thayer, 1986), logistic regression (Swaminathan & Rogers, 1990), and simultaneous item bias test (Shealy & Stout, 1993). In contrast, IRT-based methods estimate examinees' latent ability, and some of the most popular methods include Lord's $\chi^2$ (Lord, 1980), area measures (Kim & Cohen, 1991; Raju, 1988), and most recently, residual-based DIF method (Lim et al., 2022).

To better illustrate our procedures in later sections, we review three DIF detection methods: the Mantel-Haenszel test, logistic regression, and residual-based DIF. First,the Mantel-Haenszel

test is based on a contingency table where the rows of the table correspond to group membership (focal group $G = g$ versus reference group $G = g'$) and the columns correspond to correct ($Y = 1$) or incorrect ($Y = 0$) responses. After discretizating ability scores into $K$ non-overlapping strata ($k = 1, 2, \ldots, K$), the Mantel-Haenszel test computes the differences in the responses between the two groups at each $k$-th stratum of ability. Under the null hypothesis that the item is non-DIF, the Mantel-Haenszel test has a chi-squared null distribution with one degree of freedom, and if the test statistic exceeds a critical value based on the null distribution, an item exhibits DIF (Holland & Thayer, 1986; Magis et al., 2010).

Second, the detection method based on logistic regression regresses the item response $Y_i$ on ability/test scores $\theta_i$, group membership $G_i$, and their interaction (i.e., $\theta_i G_i$) (Swaminathan & Rogers, 1990) as:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1\theta_i + \beta_2 G_i + \beta_3\theta_i G_i \tag{6}$$

where $\pi_i$ is the probability of getting the studied item correct. The term $\beta_2$ represents the main effect coefficient of $G_i$, and the term $\beta_3$ represents the coefficient of the interaction effect between $G_i$ and $\theta_i$. The null hypothesis that the item is non-DIF (i.e., $\beta_2 = 0$ and $\beta_3 = 0$) is rejected if either $\beta_2$ or $\beta_3$ is significant through a likelihood ratio test. Unlike the Mantel-Haenszel test, the logistic regression can differentiate uniform DIF and nonuniform DIF by individually testing $\beta_2$ and $\beta_3$ via Wald or likelihood ratio tests. The item is uniform DIF if $\beta_2 \neq 0$ and $\beta_3 = 0$ and is nonuniform DIF if $\beta_3 \neq 0$ regardless of the value of $\beta_2$.

Finally, Lim et al. (2022) propose a residual-based DIF (RDIF) procedure to detect the presence of DIF by using an IRT model. Specifically, first, the residual-based DIF procedure fits an IRT model.[1] Second, it obtains examinee $i$'s residuals from the estimated IRT model $r_i = Y_i - \widehat{Y}_i$ where $\widehat{Y}_i$ is the prediction from the fitted IRT model. Third, the residuals are used to compute three different statistics: $RDIF_R$, $RDIF_S$, and $RDIF_{RS}$. The first DIF statistic $RDIF_R$ is the difference of mean raw residuals between the focal group and the reference group, and it follows asymptotically a normal distribution. $RDIF_R$ has (statistical) power to detect uniform DIF as Lim et al. (2022) showed via the simulations. The second DIF statistic $RDIF_S$ is the difference of mean squared residuals between the two groups and also follows asymptotically a normal distribution. $RDIF_S$ has power to detect nonuniform DIF. The third DIF statistic $RDIF_{RS}$ is a weighted combination of the two test statistics $RDIF_R$ and $RDIF_S$, and it follows asymptotically a chi-squared distribution with two degrees of freedom. $RDIF_{RS}$ is designed to detect any type of DIF (Lim et al., 2022); see Lim et al. (2022) for more details.

# 4 Our Proposal: Differential Algorithmic Functioning

## 4.1 Definitions

We propose a DAF framework to assess the fairness of algorithmic decision making by modifying the notion of DIF and its detection methods. Under the DAF framework, a fair algorithm should not make discriminatory decisions based on protected variables (e.g., gender, race/ethnicity) after accounting for *fair* attribute $W$, where $W$ is some function of $X$, i.e. $W = h(X)$, $h : \mathbb{R}^{p_x} \to \mathbb{R}^{p_w}$. The fair attribute $W$ means a set of justifiable variables that are important and valid in decision making processes, and it can be continuous or discrete.

We define DAF as conditional dependence of algorithmic decision $D$ and group membership $G$ given fair attribute $W$. Following the DIF literature, we refer to the focal group ($G = g$) as

[1] An IRT model for the residual-based DIF procedure is a three-parameter model and formalized as: $P(Y_i = 1; \theta_i) = c + \frac{1-c}{1+\exp(-a(\theta_i - b))}$. The parameter $\theta_i$ represents examinee $i$'s ability parameter. The item parameters $a$, $b$, and $c$ represent the item discrimination, difficulty/location, and pseudo guessing parameters, respectively.

the group anticipated to be disadvantaged by the algorithm and the reference group ($G = g'$) as the group who is anticipated to have an advantage, though the designation does not affect a DAF analysis. Formally, DAF and non-DAF are written as:

$$\text{DAF} : Pr(D = 1|W, G = g) \neq Pr(D = 1|W, G = g'), \tag{7}$$

$$\text{Non-DAF} : Pr(D = 1|W, G = g) = Pr(D = 1|W, G = g'). \tag{8}$$

In words, an algorithm exhibits DAF if the probability of receiving the treatment decision is different across groups after accounting for the fair attribute; otherwise, the algorithm is non-DAF. Our DAF notion does not pursue fairness of outcomes/results (i.e., the objective of statistical parity). Rather, it aims to highlight the fairness of the process with respect to decision allocations by treating individuals with the same fair attribute similarly. Also, the DAF notion is related to the negation of an existing fairness notion known as conditional statistical parity (Corbett-Davies et al., 2017). Specifically, DAF and conditional statistical parity are identical if legitimate risk factors for conditional statistical parity are the same as the fair attributes identified in DAF analysis. However, DAF provides a more detailed description of disparity patterns by defining different types of DAF; see Table 1.

Table 1: Types of Differential Algorithmic Functioning (DAF)

| Type | Definition | Allocation Pattern |
|---|---|---|
| Uniform DAF | The statistical relationship (e.g., odd ratios) between $D$ and $G$ is constant for all levels of $W$. | Static disparity |
| Nonuniform DAF | if it is not uniform DAF, but is still DAF | Dynamic disparity |

Borrowing from the DIF literature (e.g., Hanson, 1998), we define two types of DAF, uniform DAF and nonuniform DAF. Uniform DAF exists when the statistical relationship between $D$ and $G$ is constant for all levels of $W$. For example, a statistical relationship between $D$ and $G$ can be expressed using odd ratios: $\Delta(W) := \frac{Pr(D=1|W,G=g)(1-Pr(D=1|W,G=g'))}{(1-Pr(D=1|W,G=g))Pr(D=1|W,G=g')}$. In such a case, uniform DAF is defined to exist if $\Delta(W) = c \neq 1$ for every value of $W$ where $c$ is a constant but not equal to one; note that $\Delta(W) = 1$ achieves the conditional independence of $D$ and $G$ given $W$, i.e., non-DAF. In contrast, nonuniform DAF is DAF that is not uniform DAF; see Figure 1 for illustrations based on simulated data. When an algorithm is uniform DAF, the algorithm is consistently more advantageous to one group than the other group by recommending more

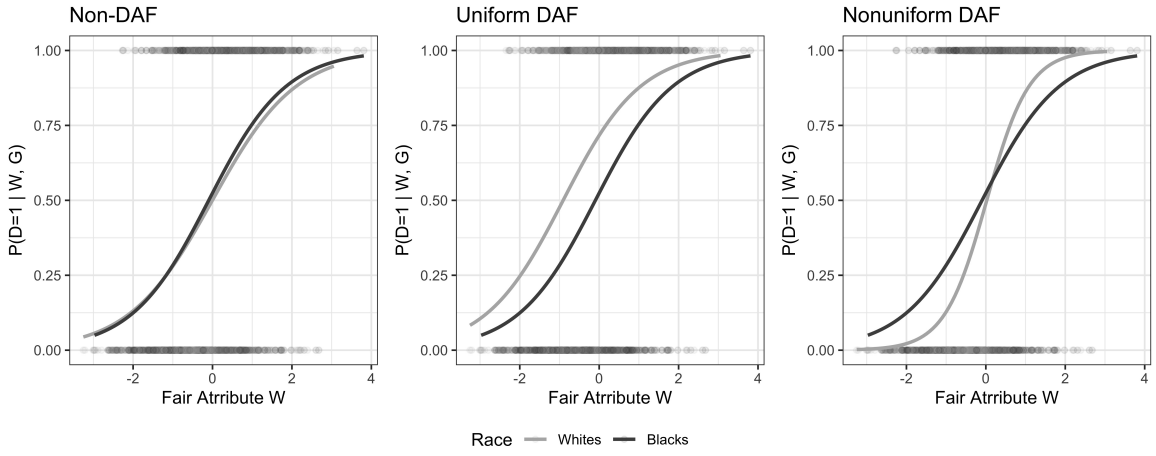


Figure 1: Illustrations of decision characteristic curves for the reference group (whites) and focal group (blacks) with different types of differential algorithmic functioning (DAF)

favorable decisions to one group across the entire range of the fair attribute; that is, it shows *static disparity* with respect to decision allocations. In contrast, an algorithm is nonuniform DAF when the advantage for one group in the algorithm depends on the fair attribute. That is, a decision may favor one group within a certain range of the fair attribute, but may favor the other group within another range of the attribute. As shown in Figure 1, nonuniform DAF would have different steepness between the decision curves of the two groups and result in *dynamic disparity* in decision allocations.[2]

## 4.2 Methods

An important advantage of DAF is that it can be easily tested by borrowing existing test statistics in the DIF literature. To detect the presence of DAF in algorithms, we adopt the aforementioned methods for DIF: the Mantel-Haenszel test, logistic regression, and residual-based DIF. First, we use the Mantel-Haenszel test by discretizing the fair attribute into $K$ strata ($k = 1, 2, \ldots, K$) and creating a contingency table where the rows of the table correspond to group membership and the columns correspond to treatment decision ($D = 1$) or control decision ($D = 0$) for each $k$-th stratum of the fair attribute:

Table 2: A Contingency Table by Group Membership $G$ and Algorithmic Decision $D$ Within the $k$-th Stratum of Fair Attribute $W$

| | Treatment Decision ($D = 1$) | Control Decision ($D = 0$) |
|---|---|---|
| Focal ($G = g$) | $N_{g1k}$ | $N_{g0k}$ |
| Reference ($G = g'$) | $N_{g'1k}$ | $N_{g'0k}$ |

Here, $N_{(\cdot\cdot)k}$ denotes observed cell frequencies within the $k$-th stratum of the fair attribute. For example, $N_{g'1k}$ denotes the number of study units who have $D = 1$ in the reference group at the $k$-th stratum. Then, the Mantel-Haenszel test computes the differences in the decisions between the focal and reference groups at each $k$-th stratum of the fair attribute. Under the null hypothesis that the algorithm is non-DAF, the Mantel-Haenszel test for detecting DAF is:

$$\chi^2_{MH} = \frac{\left\{ \left| \sum_{k=1}^{K} N_{g'1k} - \sum_{k=1}^{K} E(N_{g'1k}) \right| - .5 \right\}^2}{\sum_{k=1}^{K} V(N_{g'1k})}. \tag{9}$$

Here, $E(N_{g'1k})$ and $V(N_{g'1k})$ represents the expectation and variance of $N_{g'1k}$.[3] If the test statistic exceeds a critical value based on the chi-squared null distribution (e.g., 3.84 for $\alpha = 0.05$ and one degree of freedom), the algorithm has DAF.

Second, the logistic regression procedure requires fitting the following model:

$$\text{logit}(e_i) = \alpha_0 + \alpha_1 W_i + \alpha_2 G_i + \alpha_3 W_i G_i, \tag{10}$$

where $e_i$ is the conditional probability of unit $i$'s receiving the treatment decision given $W$ and $G$. The term $\alpha_2$ represents the effect of the group membership on the decision and $\alpha_3$ represents the interaction effect between the fair attribute and group membership. We can detect the presence of DAF in a decision-making algorithm by testing the null hypothesis that the algorithm is non-DAF (i.e., $\alpha_2 = \alpha_3 = 0$) via a likelihood ratio test. We can also detect uniform DAF by testing the null hypothesis $\alpha_2 = 0, \alpha_3 = 0$ versus the alternative $\alpha_2 \neq 0, \alpha_3 = 0$, with a Wald

[2] We note that the slight differences observed in the far left plot of non-DAF are due to sampling variability.

[3] $E(N_{g'1k}) = \frac{(N_{g'1k}+N_{g'0k})(N_{g'1k}+N_{g1k})}{N_k}$; $V(N_{g'1k}) = \frac{(N_{g'1k}+N_{g'0k})(N_{g'1k}+N_{g1k})(N_{g1k}+N_{g0k})(N_{g'0k}+N_{g0k})}{N_k^2(N_k-1)}$.

test or a likelihood ratio test. Furthermore, we can detect nonuniform DAF by testing the null hypothesis $\alpha_3 = 0$ versus the alternative $\alpha_3 \neq 0$.

Third, we revise the existing residual-based DIF method (Lim et al., 2022) to detect DAF. We replace the first step in residual-based DIF based on an IRT model with a more flexible, ensemble learning algorithm from machine learning and use the residuals from the ensemble learning algorithm in the subsequent steps. Note that an IRT model is not suitable in our setting because our outcomes of interest are not item responses. Algorithm 1 summarizes the steps of the residual-based DAF method. A bit more formally, in the first step, we estimate $E[D|W]$ via machine learning and in particular, the *SuperLearner* algorithm (van der Laan et al., 2007) that combines predictions from different supervised learning models. We use a super learning algorithm because an ensemble estimator of functionals like $E[D|W]$ will perform at least as well as the best individual estimator in terms of the cross-validated error, thereby increasing the prediction performance (Porter et al., 2011; Suk & Kang, 2022; van der Laan et al., 2007). Then, the proposed residual-based DAF method computes three statistics, $RDAF_R$, $RDAF_S$, and $RDAF_{RS}$, which, similar to their RDIF counterparts in Lim et al. (2022), are able to detect different types of DAF. The three test statistics are written as:

$$RDAF_R = \frac{\sum_{i=1}^{N} r_i I(G_i = g)}{\sum_{i=1}^{N} I(G_i = g)} - \frac{\sum_{i=1}^{N} r_i I(G_i = g')}{\sum_{i=1}^{N} I(G_i = g')} \tag{11}$$

$$RDAF_S = \frac{\sum_{i=1}^{N} r_i^2 I(G_i = g)}{\sum_{i=1}^{N} I(G_i = g)} - \frac{\sum_{i=1}^{N} r_i^2 I(G_i = g')}{\sum_{i=1}^{N} I(G_i = g')} \tag{12}$$

$$RDAF_{RS} = Q'\Sigma^{-1}Q \tag{13}$$

Here, $r_i$ represents an individual $i$'s residual, i.e., $r_i = D_i - \widehat{D}_i$. $Q$ represents a $2 \times 1$ matrix that contains differences between the first two test statistics and their respective population means, i.e., $Q = \begin{bmatrix} RDAF_R - \mu_R \\ RDAF_S - \mu_S \end{bmatrix}$, and $\Sigma$ represents a $2 \times 2$ covariance matrix of $RDAF_R$ and $RDAF_S$, i.e., $\Sigma = \begin{bmatrix} \sigma_R^2 & \sigma_{R,S} \\ \sigma_{R,S} & \sigma_S^2 \end{bmatrix}$. Likewise, the first test statistic $RDAF_R$ follows asymptotically a normal distribution and is designed to detect uniform DAF. The second test statistic $RDAF_S$ follows asymptotically a normal distribution and is designed to detect nonuniform DAF. The third test statistic $RDAF_{RS}$ follows asymptotically a chi-squared distribution with two degrees of freedom and is designed to detect any type of DAF. The residual-based DAF method using software R (R Core Team, 2021) is available in the first author's GitHub repository.[4]

**Algorithm 1** Residual-based differential algorithmic functioning (RDAF)

**Input:** Decision $D_i$, fair attribute $W_i$, and group membership $G_i$

1: Fit a super learning algorithm[5] that regresses decision $D_i$ on fair attribute $W_i$, and compute its prediction $\widehat{D}_i$ (i.e., $\widehat{D}_i = \widehat{E}[D_i|W_i]$)
2: Compute the residuals, $r_i = D_i - \widehat{D}_i$.
3: Compute the three test statistics $RDAF_R$, $RDAF_S$, and $RDAF_{RS}$ and their p-values.

**Output:** $\widehat{RDAF}_R$, $\widehat{RDAF}_S$, $\widehat{RDAF}_{RS}$, and p-values.

We summarize strengths and limitations of three methods to detect DAF; see also the results of our simulation study that investigated the Type-1 error and power rates of each method in

[4] https://github.com/youmisuk/DAF

[5] We include generalized linear models, random forests, and neural network, as default individual estimators for our residual-based DAF approach. If we only select a generalized linear model as an individual estimator insides the super learning method, the predictions are made based solely on the parametric model.

Appendix A. Overall, the Mantel-Haenszel test is a non-parametric test that does not depend on a model and hence, has valid Type-1 error control irrespective of the potentially complex relationship between $D$, $W$, and $G$ (Sireci & Rios, 2013); note that the Type-1 error control does not depend on how the strata are defined so long as they are non-overlapping. But the use of the Mantel-Haenszel test requires discretization of the fair attribute, and it has low power to detect nonuniform DAF. In contrast, the tests based on logistic regression or residual-based DAF have power to detect different types of DAF (i.e., uniform and nonuniform DAF). However, for the tests based on logistic regression, the asymptotic distributions of these tests rely on the correctness of the logistic regression model, which if mis-specified, can lead to Type-1 error inflation. The residual-based DAF method can alleviate concerns for model mis-specification by using ensemble, super learning algorithms. But it is certainly not as simple as the tests based on logistic regression.

Lastly, we make a few remarks about DAF analysis. First, researchers must use subject matter knowledge to determine which variables are fair attributes. If, however, there is limited subject matter knowledge, researchers may resort to more data-driven measures to choose fair attributes, say those based on changes in R-squared, Gini index, or classification accuracy. Second, DAF, by definition, allows multi-dimensional fair attributes, and the DAF methods above can easily accommodate multiple fair attributes. For example, the Mantel-Haenszel test just needs to create strata ($k = 1, 2, ..., K$) that are non-overlapping based on the multiple fair attributes. The logistic regression and residual-based DAF approaches need to add multiple fair attributes as predictors in the models. Third, researchers can reduce the dimension of fair attributes using dimensionality reduction tools (e.g., factor analysis, principle component analysis) in particular when multiple fair attributes are highly correlated. As we will see below in Section 5, we use kindergarten year's test scores from the Early Childhood Longitudinal Study-Kindergarten cohort (ECLS-K) as fair attributes. Specifically, tests about math, reading, and general knowledge were conducted in the fall and spring of their kindergarten year, thus producing six test scores during the kindergarten year. To account for multicollinearity among the observed fair attributes, we used factor analysis (Gorsuch, 1983; Lawley & Maxwell, 1962) in our empirical example; see the next section. Fourth, our DAF detection methods can be applied to detect other notions of group fairness, such as separation (i.e., equalized odds). Specifically, by replacing the use of $W$ as a conditioning variable with $Y$, the DAF methods can be used to evaluate whether an algorithm achieves separation or not.

# 5 Empirical Example: Retention in ECLS-K

## 5.1 Data and Methods

Decisions to retain students in grade have historically been based on teachers' assessment or test-based assessment (Huddleston, 2014). But recent algorithmic decision making can be used to assist teachers' decision making processes. Grade retention is typically considered a last-resort option and recommended to students who make inadequate progress in academic achievement or show developmental immaturity (Cannon & Lipscomb, 2011; Greene & Winters, 2006; Jackson, 1975). Prior research found that there are disparities in grade retention where retention has been skewed towards male, ethnic minority, or low-income students (Huddleston, 2014; Xia & Kirby, 2009). Given these existing disparities in retention, it is of paramount importance to consider a fairness constraint in a new algorithm for grade retention. We consider DAF as our fairness notion of interest to ensure fairness of the process in algorithmic decision making.

Specifically, we used the ECLS-K data for a retention decision-making algorithm. ECLS-K, sponsored by the National Center for Education Statistics, is a national longitudinal study to examine the school achievement and student experiences from kindergarten to middle school. ECLS-K selected a nationally representative sample of kindergarteners in the fall of 1998 and

followed them until the spring of 2007 (Walston & McCarroll, 2010). For data analysis, we used the data collected in the fall and the spring of the kindergarten year (i.e., the fall in 1998 and the spring in 1999) to obtain covariates. We also used the data collected in the spring of 2000 when most of the students were in the first grade to find whether a student was actually retained or not, which is our outcome of interest. Our analytic sample included 11,532 students that allowed for kindergarten retention.

We selected 60 covariates (i.e., $V$) that are expected to affect whether a student is retained or promoted to design a decision-making algorithm for retention based on prior works (Cannon & Lipscomb, 2011; Greene & Winters, 2006; Hong & Raudenbush, 2006; Jackson, 1975); see Appendix B for a list of covariates used in our data analysis. Among them, protected variables included gender (GENDER), race (RACE), ethnicity (WKHISP), and poverty (W1POVRTY), and fair attributes included prior achievement scores in math, reading, and general knowledge, all collected in the kindergarten year (C1RSCALE, C1MSCALE, C1GSCALE, C2RSCALE, C2MSCALE, and C2GSCALE). We summarized the fairness attributes into one variable by using factor analysis to account for multicollinearity and make it easier to demonstrate our DAF framework. This was also done as grade retention should be based on a student's general ability, rather than their performance in a specific subject. Specifically, we conducted the factor analysis using maximum likelihood estimation. The factor scores were used in their original form for the logistic regression and residual-based DAF methods, whereas for the Mantel-Haenszel test, they were categorized into 20 non-overlapping categories based on quantiles. Note that we imputed any missing values in the covariates with predictive mean matching (White et al., 2011).

To develop a retention algorithm, we fitted random forests (Breiman, 2001) with 60 covariates as predictors and actual retention status $Y_i$ as the outcome of interest. Then, we used unit $i$'s prediction $P_i$ to make a decision, i.e., whether to give retention or not. To account for the small number of retained students reported in prior works (e.g., Hong & Raudenbush, 2006; Zill et al., 1997) and in our sample, we considered a wide range of threshold values below 0.5 in our analysis to assess which threshold value would exhibit DAF in algorithmic decision making. In particular, we used a set of threshold values, ranging from 0.25 to 0.50 with an increment of 0.05. For example, based on a threshold value of 0.25, our decision is made as: $D_i = I(P_i \geq 0.25)$. Then, we assessed the presence of DAF in each working algorithm using three DAF detection methods (i.e., Mantel-Haenszel test, logistic regression, and residual-based DAF) regarding the four protected variables. Our three DAF methods have total seven test statistics: one from the Mantel-Haenszel test (denoted as MH), three from logistic regression, and three from residual-based DAF. The logistic regression procedure in (10) include the following three statistics: Wald statistic for $\alpha_2$ (denoted as $LR_{\alpha_2}$), Wald statistic for $\alpha_3$ (denoted as $LR_{\alpha_3}$), and likelihood ratio test statistic for $\alpha_2$ and $\alpha_3$ (denoted as $LR_{\alpha_2,\alpha_3}$). The residual-based DAF method has three $RDAF$ statistics: $RDAF_R$, $RDAF_S$, and $RDAF_{RS}$. As a comparison, we considered the statistical parity notion to assess a marginal difference in decision proportions where a two proportion $Z$-test is conducted (denoted as $Z$-test).

As for software, we used the R package *randomForest* (Liaw & Wiener, 2002) for developing a random-forests-based algorithm, the R package *psych* (Revelle, 2021) for factor analysis, and the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2011) for predictive mean matching.

## 5.2 Results

We include the results of DAF analysis with four protected variables: gender, race, ethnicity, and poverty. Figure 2 summarizes the p-values of three detection methods for DAF and one detection method (i.e., a two proportion $Z$-test) for statistical parity, with different threshold values. Each row of Figure 2 represents protected variables of interest, and the x-axis within each subplot varies the threshold values used in algorithmic decision making. The red dashed

line indicates the p-value of 0.05. If the p-value is below 0.05, there is sufficient evidence of DAF for DAF detection methods (or the marginal difference in decision proportions for statistical parity).

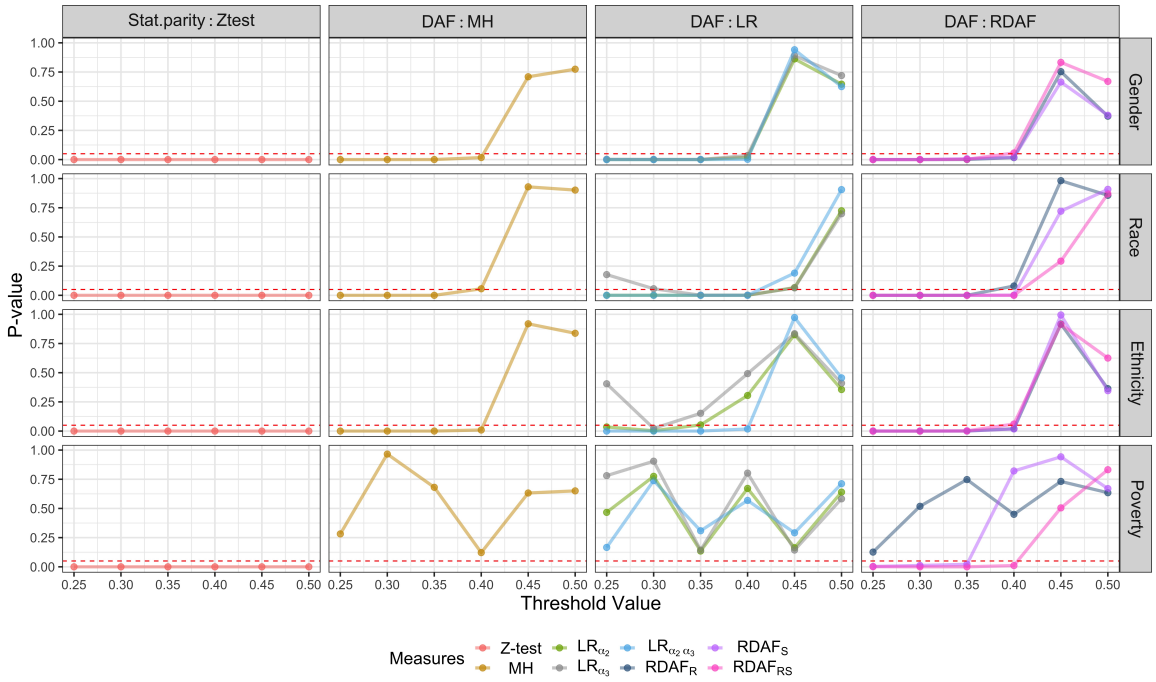


Figure 2: P-values of fairness measures about differential algorithmic functioning (DAF) and statistical parity with four protected variables: gender, race, ethnicity, and poverty. For DAF detection methods, p-values below 0.05 provide sufficient evidence of DAF, whereas p-values of 0.05 or above do not provide sufficient evidence.

As seen from Figure 2, the statistical parity notion that seeks fairness of overall outcomes/results is not satisfied for all combinations of the threshold values and different protected variables. This means that there is the marginal difference in the proportions of retention decisions between the focal group and the reference group within each protected variable. But satisfying statistical parity is not of much interest in designing this retention algorithm, and as mentioned before, we aim to make the working algorithm DAF-free.

DAF results depend on the threshold values, protected variables, and DAF detection methods. All the DAF detection methods detect DAF if the threshold value is less than or equal to 0.40 except for the Mantel-Haenszel test and logistic regression with the poverty variable. More specifically, logistic regression and residual-based DAF methods detect the presence of nonuniform DAF with respect to gender, race, and ethnicity if the threshold value is below or at 0.4; see the results of $LR_{\alpha_3}$ and $RDAF_{RS}$. Regarding the poverty level, the residual-based DAF detects the presence of nonuniform DAF if the threshold value is below 0.4, whereas logistic regression does not detect any type of DAF across different threshold values. This difference may be partly due to different power rates of detecting nonuniform DAF between two methods.

Furthermore, we investigated the decision characteristic curves from logistic regression to better understand the presence of DAF between the threshold value of 0.25 (Decision$_{0.25}$) and 0.45 (Decision$_{0.45}$). Figure 3 visualizes the decision characteristic curves from logistic regression. Note that omitting the fair attribute of above 2.3 permitted clearer comparison of the logistic regression curves. Regarding gender, race, and ethnicity, the line for the reference group (in gray) does not agree with the line for the focal group (in black) at the threshold value of 0.25. That is, DAF exists, and in particular, nonuniform DAF is clearly shown for the gender variable.

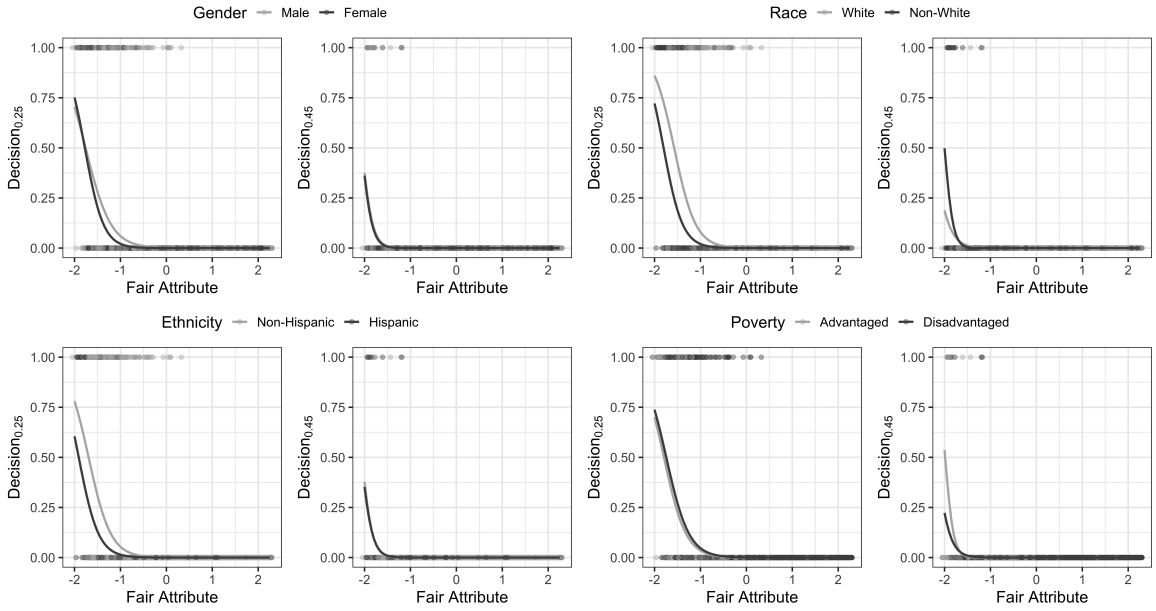


Figure 3: Decision characteristic curves from logistic regression between the threshold value of 0.25 ($\text{Decision}_{0.25}$) and of 0.45 ($\text{Decision}_{0.45}$) with four protected variables: gender, race, ethnicity, and poverty.

But in Figure 3 there seems no DAF in the poverty level at the threshold value of 0.25, which is confirmed from Figure 2 with the logistic regression procedure. At the threshold value of 0.45, the two curves in each subplot generally agree with each other (i.e., show non-DAF), though we observe a somewhat departure at the low extreme for race and poverty. Based on all our findings, we conclude that the working algorithm exhibits nonuniform DAF at some threshold values, and using a threshold of above 0.4 can make it DAF-free.

# 6 Discussion and Conclusions

This paper presents a novel framework for evaluating fairness in algorithmic decision making, referred to as DAF. The framework is based on DIF and places emphasis on the fairness of the decision-making process rather than the fairness of the outcomes or results by considering a fair attribute. We define DAF as conditional dependence of algorithmic decision $D$ and group membership $G$ given fair attribute $W$. Compared to other fairness notions, one of the key innovations of this framework is the ability to distinguish between two subtypes of DAF: uniform DAF and nonuniform DAF. This distinction is made by examining disparities in decision allocations, where uniform DAF exhibits static disparity and nonuniform DAF shows dynamic disparity. This differentiation is crucial as it helps in understanding the underlying disparity mechanisms of algorithms over fair attributes, the impact of unfair bias, and the connection to other fairness concepts such as statistical parity. Moreover, to detect the presence of DAF, we provide three different DAF detection methods: Mantel-Haenszel test, logistic regression, and residual-based DAF. Unlike the Mantel-Haenszel test, logistic regression and residual-based DAF are capable of distinguishing between uniform DAF and nonuniform DAF. The effectiveness of the DAF framework is demonstrated through an application to assess the presence of DAF in an algorithm for grade retention.

Tradeoffs between different fairness notions may be inherent because they fulfill different objectives and are often in conflict. As mentioned above, when comparing statistical parity with DAF, the statistical parity notion pursues the long-term goal of fairness in outcomes or results

from algorithms, whereas the DAF notion underscores the fairness in the process of algorithmic decision making by incorporating the fair attribute. Obviously, satisfying the outcome equity via the statistical parity notion does not guarantee satisfying process fairness via the DAF notion (as can be inferred from our simulation study in Appendix A). Therefore, researchers should prioritize a fairness notion that is of most importance depending on their contexts.

While satisfying a particular notion of fairness restricts a set of decision rules in algorithmic decision making, multiple rules may satisfy the given fairness notion. Thus, researchers have to determine which rule is optimal among those satisfying the fairness constraint. In general, one seeks to maximize a certain notion of the prediction performance (e.g., accuracy, area under curve, F1 score) or utility (i.e., a function of the benefit and cost) in designing an algorithm and thus, they can choose an optimal decision rule by accounting for metrics on both fairness and prediction performance (or utility). Also, researchers should avoid designing a fair but ineffective algorithm which is of little use in practice. A working algorithm, despite achieving the fairness, can still give poor results when evaluated against prediction performance or utility metrics.

In choosing a decision rule, a single threshold value might not be a solution to ensure that an algorithm is fair. In this case, researchers may consider setting different thresholds for different groups as in prior works (e.g., Corbett-Davies et al., 2017; Lee & Kizilcec, 2020). However, under the DAF framework, using group-specific thresholds may not be justifiable if using group-specific thresholds decreases the perceived fairness of the decision-making process. That is, if individuals who receive decisions view the algorithmic decision-making process unfavorably due to the use of group-specific thresholds, it would go against the objective of DAF notion. Also, it should be noted that using group-specific threshold values is at odds with other fairness criteria such as anti-classification (Corbett-Davies & Goel, 2018) because the decision rule depends on protected group membership.

Based on all the findings of this paper, we provide some suggestions for future research concerning our proposed DAF framework. First, we did not consider intersectionality, i.e., *systematic disadvantages along intersecting dimensions*, which contain not only gender, but also race, ethnicity, or disability status (Foulds et al., 2020). Further research will investigate how to consider intersectionality in DAF analysis. Second, while we briefly discussed an optimal decision rule above, we did not formalize how to choose the optimal decision. Future research will examine how to choose an optimal decision rule by considering tradeoffs between DAF and prediction performance or utility metrics. Third, among many other DIF methods, we utilize three DIF methods for DAF analysis. Future research would examine how to modify other DIF methods like simultaneous item bias test (Shealy & Stout, 1993) for DAF analysis and evaluate their Type-1 error and power. Fourth, we only used the conditional independence tests within the DIF literature and did not consider other statistical methods proposed elsewhere (e.g., Azadkia & Chatterjee, 2021; Neykov et al., 2021). Future work could consider these alternative methods. Fifth, DIF is a necessary but not sufficient condition for bias and fairness (Angoff, 1993). Likewise, algorithms flagged as DAF only have the potential to be unfair. A holistic approach spanning technical and non-technical solutions would be required to scrutinize fairness-related biases inside algorithms, such as Mulligan et al. (2019)'s fairness analytic and Madaio et al. (2020)'s co-designed checklist for AI fairness. Sixth, we applied an imputation technique to handle missing values in the covariates in the nationally representative ECLS-K data. This imputation may have introduced unwanted bias if the imputed values reinforced any unfair dependencies between the protected attributes, fair attributes, and decisions. Thus, future studies would examine solutions to handle missing data when evaluating fairness in algorithmic decision making. Lastly, in the literature from industrial and organizational (I/O) psychology, ethical decision making is one of the most important research areas (e.g., Jarrahi, 2018; Jones, 1991; Lefkowitz, 2017). Future research would enrich our findings on algorithmic fairness by identifying biases in decision making from the I/O psychology literature and

borrowing associated methodology.

While no one-size-fits-all definition suits all systems and contexts, our DAF framework highlights the fairness of decision allocations and provides insights on different patterns of decision disparities from the lens of psychometric testing. We believe that our DAF framework will serve as a useful tool to assess fairness in algorithmic decision making and can be a meaningful starting point that connects the concept of test fairness and algorithmic fairness.

## Acknowledgements

The authors thank Dan Bolt for his valuable feedback on the manuscript. 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.

## References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Routledge.

Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement*, *34*(4), 807–816. https://doi.org/10.1177/001316447403400408

Azadkia, M., & Chatterjee, S. (2021). A simple measure of conditional dependence. *The Annals of Statistics*, *49*(6), 3070–3102. https://doi.org/10.1214/21-aos2073

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

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

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. http://www.ppic.org/main/publication.asp?i=910

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement*, *28*(1), 61–75. https://doi.org/10.1177/001316446802800106

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv*. https://doi.org/10.48550/ARXIV.1808.00023

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 797–806. https://doi.org/10.1145/3097983.3098095

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. https://doi.org/10.1145/2090236.2090255

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

Gorsuch, R. L. (1983). *Factor analysis*. Lawrence Erlbaum Associates.

Greene, J. P., & Winters, M. A. (2006). Getting ahead by staying behind: An evaluation of florida's program to end social promotion. *Education Next*, *6*(2), 65–70.

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

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

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

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

Jackson, G. B. (1975). The research evidence on the effects of grade retention. *Review of Educational Research*, *45*(4), 613–635. https://doi.org/10.3102/00346543045004613

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review*, *16*(2), 366–395. https://doi.org/10.5465/amr.1991.4278958

Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, *15*(3), 269–278. https://doi.org/10.1177/014662169101500307

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th innovations in theoretical computer science conference (ITCS 2017)* (43:1–43:23). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *12*(3), 209–229. https://doi.org/10.2307/2986915

Lee, H., & Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. https://doi.org/10.48550/ARXIV.2007.00088

Lefkowitz, J. (2017). *Ethics and values in industrial-organizational psychology*. Routledge. https://doi.org/10.4324/9781315628721

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. https://CRAN.R-project.org/doc/Rnews/

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. http://www.jennwv.com/papers/checklists.pdf

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an r package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/brm.42.3.847

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), 1–35. https://doi.org/10.1145/3457607

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105–118. https://doi.org/10.2307/1164960

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

Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction, 3(CSCW)*, 1–36. https://doi.org/10.1145/3359221

Neykov, M., Balakrishnan, S., & Wasserman, L. (2021). Minimax optimal conditional independence testing. *The Annals of Statistics*, *49*(4), 2151–2177. https://doi.org/10.1214/20-aos2030

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.

Porter, K. E., Gruber, S., Van Der Laan, M. J., & Sekhon, J. S. (2011). The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, *7*(1), 31. https://doi.org/10.2202/1557-4679.1308

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/bf02294403

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.1.9]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

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

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170–187. https://doi.org/10.1080/13803611.2013.767621

Suk, Y., & Kang, H. (2022). Robust machine learning for treatment effects in multilevel observational studies under cluster-level unmeasured confounding. *Psychometrika*, *87*(1), 310–343. https://doi.org/10.1007/s11336-021-09805-x

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1). https://doi.org/10.2202/1544-6115.1309

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*.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Xia, N., & Kirby, S. N. *Retaining students in grade: A literature review of the effects of retention on students' academic and nonacademic outcomes. (technical report no. 678)*. 2009. http://www.rand.org/pubs/technical_reports/TR678/

Xu, J., Xiao, Y., Wang, W. H., Ning, Y., Shenkman, E. A., Bian, J., & Wang, F. (2022). Algorithmic fairness in computational medicine. *medRxiv*. https://doi.org/10.1101/2022.01.16.21267299

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 Simulation Study

## A.1 Designs and Evaluation

We conduct simulation studies to assess the performance of three DAF methods with total seven test statistics: one from the Mantel-Haenszel test (denoted as MH), three from logistic regression, and three from residual-based DAF. For the logistic regression procedure in (10), we use the following three statistics: Wald statistic for $\alpha_2$ (denoted as $LR_{\alpha_2}$), Wald statistic for $\alpha_3$ (denoted as $LR_{\alpha_3}$), and likelihood ratio test statistic for $\alpha_2$ and $\alpha_3$ (denoted as $LR_{\alpha_2,\alpha_3}$). For the residual-based DAF method, we use three $RDAF$ statistics: $RDAF_R$, $RDAF_S$, and $RDAF_{RS}$. As a comparison, we include the statistical parity metric where a test statistic is based on a two proportion $Z$-test that compares two independent population proportions (denoted as $Z$-test).

Our simulation study is categorized into four designs; see Figure 4 for illustrations of our simulation designs. Design 1 assumes a non-DAF case where $\alpha_2 = 0$ and $\alpha_3 = 0$ in equation (10). Design 2 assumes a uniform DAF case where $\alpha_2 \neq 0$ and $\alpha_3 = 0$. Design 3 assumes a "balanced" nonuniform DAF case where the group advantage is balanced across the fair attribute, i.e., $\alpha_2 = 0$ and $\alpha_3 \neq 0$. Design 4 assumes a "unbalanced" nonuniform DAF case where the group advantage is not balanced across the fair attribute i.e., $\alpha_2 \neq 0$ and $\alpha_3 \neq 0$. Specifically, we set $\alpha_2$ and $\alpha_3$ to be 0.4 if they are not equal to zero. For each design, we varied the measurement scale of the fair attribute with three different levels: a continuous scale, a discrete scale with 15 unique values, and a discrete scale with 5 unique values. We did this to consider that a fair attribute may not be measured on the continuous scale. Also, for the three scale levels, we formed 20, 9, and 5 intervals in the Mantel-Haenszel procedure, respectively. As for the sample size, we used the fixed value of 2,000 that consists of 1,000 in the focal group and 1,000 in the reference group.

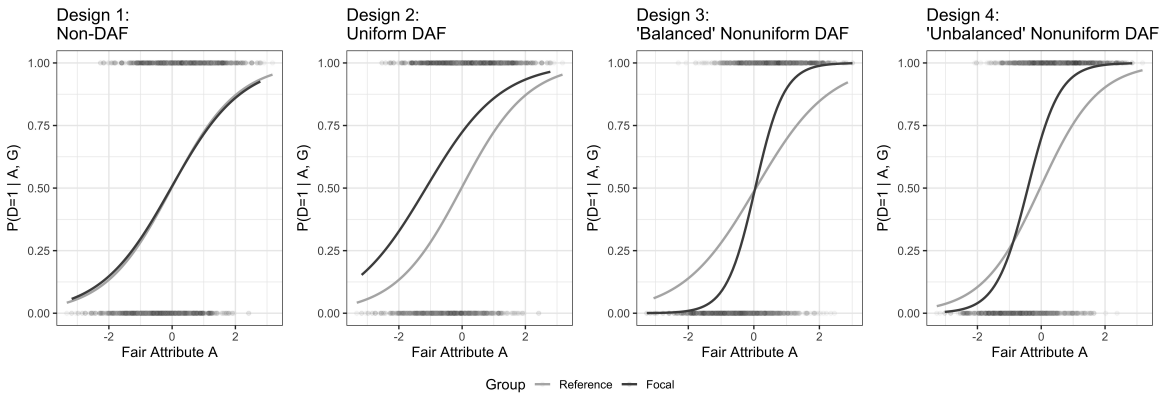


Figure 4: Simulation Designs

For all the designs, we examined the performance of DAF methods by repeating the simulation 1,000 times, and we evaluated the performance of each method by measuring average detection rates ($\sum I(\text{P-value} \leq 0.05)/1,000$). A DAF test statistic that is significant at the 5% alpha level indicates evidence of DAF, and a test statistic for statistical parity is significant at the 5% alpha level indicates evidence of the marginal difference in decisions. Specifically, the average detection rates indicate the power rates if a test statistic is designed for detecting a particular type of DAF (or marginal difference for statistical parity); otherwise, they indicate Type-1 error rates.

## A.2 Results

Table 3 summarizes the average detection rates under Designs 1, 2, 3, and 4. If the average detection rates can be interpreted as Type-1 error, we highlight them in gray in the table. Under Design 1 that assumes non-DAF, the performances of all DAF detection methods—the Mantel-Haenszel test, logistic regression, and residual-based DAF—are similar across different measurement scales of the fair attribute, and they show very low detection rates of DAF, i.e., well controlled Type 1 error rates (.¡07). The statistical parity test (i.e., $Z$ test), which focuses on a marginal difference in decision between groups, also shows very low detection rates under this design.

Design 2 assumes uniform DAF, and in this design, all the DAF methods generally perform as expected. DAF test statistics that specialize in detecting uniform DAF or any type of DAF show high detection rates (i.e., ¿.9 power rates) across measurement scales of the fair attribute. When we focus on logistic regression and residual-based DAF methods, test statistics for detecting uniform DAF (i.e., $LR_{\alpha_2}$ and $RDAF_R$) show slightly higher power rates than test statistics for detecting any type of DAF (i.e., $LR_{\alpha_2,\alpha_3}$ and $RDAF_{RS}$). In contrast, test statistics for detecting nonuniform DAF (i.e., $LR_{\alpha_3}$ and $RDAF_S$) show low detection rates, which are desirable, but $RDAF_S$ shows somewhat over-estimated Type-1 error rates. The statistical parity test shows high detection rates of the marginal difference, but its power rates are smaller than those from any DAF methods.

For Design 3 with balanced unnonuniform DAF, logistic regression and residual-based DAF methods perform well; the test statistics that specialize in detecting nonuniform DAF (i.e., $LR_{\alpha_3}$ and $RDAF_S$) show large power rates (¿.8) regardless of the measurement scales, and have higher power rates than the test statistics for detecting any type of DAF (i.e., $LR_{\alpha_2,\alpha_3}$ and $RDAF_{RS}$). Also, the test statistics for detecting uniform DAF (i.e., $LR_{\alpha_2}$ and $RDAF_R$) show well-controlled Type-1 error rates. However, the Mantel-Haenszel statistics fail to detect nonuniform DAF. This is because the Mantel-Haenszel procedure is not able to detect DAF in particular when the group advantage is cancelled out across the levels of the fair attribute. Also, the statistical parity test shows very low retention rates because no marginal difference is expected under the balanced nonuniform design. Our last design of Design 4 assumes unbalanced nonuniform DAF, and under this design, all the average detection rates are high, ranging from 88.1% to 99.5%. We also find that $RDAF_S$ shows high power rates than $LR_{\alpha_3}$, and DAF methods generally higher power rates compared to the statistical parity test.

Overall, the logistic regression and residual-based DAF methods perform well by detecting both uniform DAF and nonuniform DAF. Using the Mantel-Haenszel procedure may not be desirable when balanced nonuniform DAF exhibits where the group advantage is cancelled out over the levels of the fair attribute. We also find that satisfying statistical parity does not guarantee achieving DAF-free; specifically, the statistical parity is met even when DAF (in particular, balanced uniform DAF) is present.

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

| | Continuous | Discrete: 15 | Discrete: 5 |
|---|---|---|---|
| Design 1: non-DAF ($\alpha_2 = 0, \alpha_3 = 0$) | | | |
| $Z$-test | 0.040 | 0.044 | 0.049 |
| MH | 0.041 | 0.052 | 0.055 |
| $LR_{\alpha_2}$ | 0.044 | 0.050 | 0.050 |
| $LR_{\alpha_3}$ | 0.061 | 0.049 | 0.063 |
| $LR_{\alpha_2,\alpha_3}$ | 0.056 | 0.046 | 0.054 |
| $RDAF_R$ | 0.040 | 0.051 | 0.050 |
| $RDAF_S$ | 0.060 | 0.049 | 0.064 |
| $RDAF_{RS}$ | 0.052 | 0.045 | 0.057 |
| Design 2: uniform DAF ($\alpha_2 \neq 0, \alpha_3 = 0$) | | | |
| $Z$-test | 0.945 | 0.955 | 0.942 |
| MH | 0.972 | 0.975 | 0.984 |
| $LR_{\alpha_2}$ | 0.974 | 0.980 | 0.979 |
| $LR_{\alpha_3}$ | 0.054 | 0.047 | 0.055 |
| $LR_{\alpha_2,\alpha_3}$ | 0.953 | 0.960 | 0.954 |
| $RDAF_R$ | 0.974 | 0.984 | 0.982 |
| $RDAF_S$ | 0.103 | 0.089 | 0.091 |
| $RDAF_{RS}$ | 0.951 | 0.958 | 0.956 |
| Design 3: balanced nonuniform DAF ($\alpha_2 = 0, \alpha_3 \neq 0$) | | | |
| $Z$-test | 0.041 | 0.029 | 0.042 |
| MH | 0.048 | 0.058 | 0.048 |
| $LR_{\alpha_2}$ | 0.033 | 0.037 | 0.045 |
| $LR_{\alpha_3}$ | 0.881 | 0.880 | 0.858 |
| $LR_{\alpha_2,\alpha_3}$ | 0.807 | 0.810 | 0.784 |
| $RDAF_R$ | 0.038 | 0.036 | 0.047 |
| $RDAF_S$ | 0.857 | 0.870 | 0.861 |
| $RDAF_{RS}$ | 0.769 | 0.803 | 0.785 |
| Design 4: unbalanced nonuniform DAF ($\alpha_2 \neq 0, \alpha_3 \neq 0$) | | | |
| $Z$-test | 0.889 | 0.885 | 0.899 |
| MH | 0.938 | 0.957 | 0.946 |
| $LR_{\alpha_2}$ | 0.978 | 0.979 | 0.976 |
| $LR_{\alpha_3}$ | 0.881 | 0.901 | 0.889 |
| $LR_{\alpha_2,\alpha_3}$ | 0.995 | 0.995 | 0.994 |
| $RDAF_R$ | 0.955 | 0.957 | 0.956 |
| $RDAF_S$ | 0.944 | 0.956 | 0.953 |
| $RDAF_{RS}$ | 0.995 | 0.995 | 0.994 |

Note. MH = Mantel-Haenszel, LR = logistic regression, and RDAF = residual-based differential algorithmic functioning (DAF) statistics. The average detection rates are interpreted as power rates when test statistics are designed for detecting a particular type of DAF (or a marginal difference); otherwise, they are interpreted as Type-1 error (highlighted in gray in the table). The true effect size for $\alpha$ on odds ratio scale is 1.49.

# B A list of Covariates

Table 4: A list of Covariates in a Decision-Making Algorithm for Grade Retention

| | Variables | Variable Names in ECLS-K |
|---|---|---|
| 1 | Gender | GENDER |
| 2 | Race | RACE |
| 3 | Hispanic | WKHISP |
| 4 | Poverty | W1POVRTY |
| 5 | C1 reading IRT scale score | C1RSCALE |
| 6 | C1 math IRT scale score | C1MSCALE |
| 7 | C1 general knowledge IRT scale score | C1GSCALE |
| 8 | C2 reading IRT scale score | C2RSCALE |
| 9 | C2 math IRT scale score | C2MSCALE |
| 10 | C2 general knowledge IRT scale score | C2GSCALE |
| 11 | Age at kindergarten entry | P1AGEENT |
| 12 | Spring, K child in in-class ESL program | T2INCESL |
| 13 | SES | WKSESL |
| 14 | Mother's education | WKMOMED |
| 15 | English as home language | WKLANGST |
| 16 | Number of siblings | P1NUMSIB |
| 17 | Fmaily type | P1HFAMIL |
| 18 | How many books child has | P1CHLBOO |
| 19 | Fall, K parent report of child's frequency of reading books outside school | P1CHREAD |
| 20 | Spring, K parent report of child's frequency of reading books outside school | P2CHREAD |
| 21 | Home computer for child use | P2HOMECM |
| 22 | Parent educational expectation | P1EXPECT |
| 23 | Child ever in center-based care | P1CENTER |
| 24 | Child receiving special service/education | P2SPECND |
| 25 | Spring, K child fell behind due to health | T2FLBHND |
| 26 | Child with disability | P1DISABL |
| 27 | Fall, K child literacy ARS score | T1RARSLI |
| 28 | Fall, K child math ARS score | T1RARSMA |
| 29 | Fall, K child general knowledge ARS score | T1RARSGE |
| 30 | Fall, K teacher rating on child approaches to learning | T1LEARN |
| 31 | Fall, K teacher rating on child self control | T1CONTRO |
| 32 | Fall, K teacher rating on child interpersonal skills | T1INTERP |
| 33 | Fall, K teacher rating on child externalizing problem behaviors | T1EXTERN |
| 34 | Fall, K teacher rating on child internalizing problem behaviors | T1INTERN |
| 35 | Spring, K child literacy ARS score | T2RARSLI |
| 36 | Spring, K child math ARS score | T2RARSMA |
| 37 | Spring, K child general knowledge ARS score | T2RARSGE |
| 38 | Spring, K teacher rating on child approaches to learning | T2LEARN |
| 39 | Spring, K teacher rating on child self control | T2CONTRO |
| 40 | Spring, K teacher rating on child interpersonal skills | T2INTERP |
| 41 | Spring, K teacher rating on child externalizing problem behaviors | T2EXTERN |
| 42 | Spring, K teacher rating on child internalizing problem behaviors | T2INTERN |
| 43 | Spring, K teacher rating on child language skills | T2RTLANG |
| 44 | Spring, K teacher rating on child science/social studies skills | T2RTSCI |
| 45 | Spring, K teacher rating on child math skills | T2RTMTH |
| 46 | Spring, K teacher report on child not working at best ability | T2ABIL |
| 47 | Number of class hours per day in Fall, K | A1HRSDA |
| 48 | Days per week in Fall, K | A1DYSWK |
| 49 | Time on teacher-directed whole class activity in Fall, K | B1WHLCLS |
| 50 | Spring, K percentage of minority students enrolled | S2KMINOR |
| 51 | Spring, K school percentage of Hispanic students | S2PCTHSP |
| 52 | Spring, K school is public versus private | S2KPUPRI |
| 53 | Spring, K school enrollment requiring academic records | S2ACADRC |
| 54 | Spring, K School receives Title 1 funding | S2TT1 |
| 55 | Spring, K school number of FTE bilingual-ESL teachers | S2ESLFTE |

| | | |
|---|---|---|
| 56 | Spring, K principal report of school being successful in providing help to low achievers | S2SUCC7 |
| 57 | Spring, K principal report of raising performance level of low-achieving students influencing evaluation of principal performance | S2PRFLVL |
| 58 | Spring, K principal report of teacher and staff support influencing evaluation of principal performance | S2STFSPP |
| 59 | Spring, K school safety rating | K2Q3 |
| 60 | Spring, K school with decorated hallways | K2Q6_A |