

LIMITATIONS OF THE “FOUR-FIFTHS RULE” AND STATISTICAL PARITY TESTS FOR MEASURING FAIRNESS

Manish Raghavan* & Pauline T. Kim**

To ensure the fairness of algorithmic decision systems, such as employment selection tools, computer scientists and practitioners often refer to the so-called “four-fifths rule” as a measure of a tool’s compliance with anti-discrimination law. This reliance is problematic because the “rule” is in fact not a legal rule for establishing discrimination, and it offers a crude test that will often be over- and under-inclusive in identifying practices that warrant further scrutiny. The “four-fifths rule” is one of a broader class of statistical tests, which we call Statistical Parity Tests (SPTs), that compare selection rates across demographic groups. While some SPTs are more statistically robust, all share some critical limitations in identifying disparate impacts retrospectively. When these tests are used prospectively as an optimization objective shaping model development, additional concerns arise about the development process, behavioral incentives, and gameability. In this Article, we discuss the appropriate role for SPTs in algorithmic governance. We suggest a combination of measures that take advantage of the additional information present during prospective optimization, providing greater insight into fairness considerations when building and auditing models.

* Drew Houston Career Development Professor, MIT Sloan and EECS

** Daniel Noyes Kirby Professor of Law, Washington University School of Law

TABLE OF CONTENTS

INTRODUCTION	94
I. LEGAL BACKGROUND	96
A. Origins of the “Four-Fifths Rule”	97
B. Proving Disparate Impact in Practice	98
II. THE “FOUR-FIFTHS RULE” AS A FAIRNESS METRIC.....	101
III. LIMITATIONS OF THE “FOUR-FIFTHS RULE” AND SPTs GENERALLY....	104
A. The Prima Facie Case: A Retrospective View	105
1. <i>Statistical Significance</i>	105
2. <i>Selectivity</i>	106
3. <i>Beyond Binary Outcomes</i>	107
B. SPTs as Fairness Metrics: A Prospective View	109
1. <i>Limited Measures of Performance and Bias</i>	110
2. <i>Data Representativeness</i>	112
3. <i>Determining the Relevant Pool</i>	114
IV. SPTs AND ALGORITHMIC GOVERNANCE	115
A. Litigating Discrimination Retrospectively.....	115
B. Auditing Algorithms Prospectively	117
C. A Concrete Technical Proposal.....	119
CONCLUSION.....	122

INTRODUCTION

Algorithmic tools have become increasingly common in a variety of social domains like consumer finance, housing, employment, and criminal law enforcement. For example, in the employment context, a typical use case involves algorithms that screen job applicants to determine which candidates should be advanced in the hiring process. The applicant provides information, such as a resume, responses to a questionnaire, or even a recorded video, which is then broken down to discrete data points that are analyzed by the algorithm to make a recommendation.¹ These algorithms typically entail models trained on historical data and are often developed by third-party firms specializing in algorithmic assessments.

¹ See MIRANDA BOGEN & AARON RIEKE, *HELP WANTED: AN EXPLORATION OF HIRING ALGORITHMS, EQUITY AND BIAS* (2018); Manish Raghavan, Solon Barocas, Jon Kleinberg & Karen Levy, *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, PROC. 2020 CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 469 (2020).

As awareness has grown that algorithms can discriminate, computer scientists and practitioners have sought to develop methods to ensure that models are fair. In doing so, many such efforts reference the so-called “four-fifths rule” as a measure of a tool’s compliance with anti-discrimination law. The “four-fifths rule” examines the ratio of selection rates across relevant demographic characteristics. For a given selection tool or practice, it asks whether the selection rate of a disadvantaged group is less than four-fifths, or 80%, of the selection rate of an advantaged group. So, for example, if 67% of Black applicants and 90% of white applicants are selected for a benefit, the ratio of selection is $67/90$ or 74%. Because the ratio is less than four-fifths or 80%, the practice would be judged to have a disparate impact on Black applicants.

The focus on the four-fifths ratio has its origins in law—specifically in employment discrimination law—but its use as a metric for measuring fairness in algorithms is problematic for two primary reasons. First, it is not a legal rule and has never been. To the extent that developers have turned to a four-fifths ratio as a way of ensuring compliance with anti-discrimination law, they are mistaken. It does not provide the legal definition of discrimination, and courts have generally rejected it as a determinative test for finding a *prima facie* case of disparate impact. Second, putting aside the inaccurate understanding of the law, the “four-fifths rule” is a poor measure of discrimination because it is a crude statistical measure that will often be over- and under-inclusive in identifying practices that warrant further scrutiny.

The “four-fifths rule” is one of a broader class of statistical tests that we call Statistical Parity Tests, or SPTs. SPTs seek to measure fairness by comparing positive outcomes across groups—for example, the rate at which Black and white candidates are selected for a job. Examples of SPTs include Fisher’s exact test and the chi-squared test.² Some of the deficiencies of the “four-fifths rule” are addressed by SPTs that are more statistically robust; nevertheless, all of these tests share some critical limitations.

In the legal context, courts look to statistical tests to determine whether a *prima facie* case exists that a particular practice has a disparate impact, warranting further scrutiny. In that sense, the inquiry is retrospective—it asks whether a challenged practice or test has caused disproportionate disadvantage for marginalized groups. When developers use an SPT as a fairness metric, however, they rely on it *prospectively* as an optimization objective shaping model

² See Nancy T. Tippins, *Adverse Impact in Employee Selection Procedures from the Perspective of an Organizational Consultant*, in *ADVERSE IMPACT: IMPLICATIONS FOR ORGANIZATIONAL STAFFING AND HIGH STAKES SELECTION* 201, 204–08 (James Outtz ed., 2010).

development. This new context raises additional concerns about the development process, behavioral incentives, and gameability.

Prospective testing that goes beyond SPTs can provide more comprehensive insights into the properties of a model than are available after the fact. This potential for greater insight results from differences in the availability of information. In a retrospective analysis, certain types of information are not available, making it impossible to test for some sophisticated measures of discrimination. In contrast, a firm or auditor conducting testing during the model development and validation phases can take advantage of the information available *ex ante* to test for different measures of discrimination. In what follows, we argue that regulators can and should leverage this additional information to incentivize the development of fair algorithms.

In this Article, we criticize use of the four-fifths rule of thumb for algorithmic decisions, primarily focusing on employment selection tools as an illustrative use case. Some of the crudeness of relying on a four-fifths ratio is alleviated by using more statistically robust SPTs. We argue, however, that even though SPTs, when properly applied, can be helpful in diagnosing discriminatory effects after the fact, they are more problematic when relied on prospectively as the measure of fairness during model development.

We begin in Part I by explaining that the “four-fifths rule” is not a legal test. We trace the origins of the idea and describe how cases of disparate impact discrimination are in fact established in court. Part II provides context for how a four-fifths rule of thumb has become incorporated in computer science and model development. In Part III, we analyze the limitations of the “four-fifths rule” and other SPTs as applied to algorithms. We argue that in isolation, they are too blunt as instruments for detecting whether a practice caused discrimination, and when used prospectively to optimize algorithmic hiring assessments, they can create further problems. In Part IV, we discuss the appropriate role of SPTs in algorithmic governance, noting some positive aspects of SPTs and suggesting ways to combine them with other types of prospective testing to provide comprehensive insights into the fairness properties of a model.

I. LEGAL BACKGROUND

This Part explores the role of SPTs in legal doctrine. Part I.A describes the origins of the four-fifths ratio as a rule of thumb guiding government enforcement efforts and how the so-called rule became a focal point of attention when considering disparate impact discrimination. Part I.B explains how disparate impact cases are actually litigated, emphasizing that courts generally rely on more sophisticated SPTs, not a simple four-fifths ratio.

A. Origins of the “Four-Fifths Rule”

When the Civil Rights Act of 1964 outlawed discrimination in employment, uncertainty arose about the legality of pre-employment tests which were widely used by employers. These tests were not necessarily intentionally discriminatory, which would have clearly violated the law as a form of disparate treatment. Nevertheless, they often had a disparate racial impact on hiring. Different federal agencies, each having some enforcement responsibilities, developed different guidelines regarding pre-employment tests. In an effort to produce a consistent government position, the principal agencies involved in enforcing employment discrimination laws (Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice, Department of Labor) issued the 1978 Uniform Guidelines on Employee Selection Procedures (“Guidelines”).³

The Guidelines confirmed the basic framework to be followed in evaluating employee selection procedures. Where a practice had an adverse impact on protected groups, the agencies would consider it discriminatory unless justified.⁴ The Guidelines then explained in considerable technical detail how a test could be validated under existing professional standards established by industrial psychologists.

Importantly, if a test or procedure had no adverse impact, the agencies would not require validity studies.⁵ Thus, the question of what constituted an adverse impact became salient. The Guidelines explained that if the ratio of selection rates between two groups was less than four-fifths, it would “generally be regarded by the [f]ederal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by [f]ederal enforcement agencies as evidence of adverse impact.”⁶

The four-fifths ratio was never intended to be a rule of law, but rather a “rule of thumb.”⁷ It offered a “practical device” to guide the enforcement priorities of the relevant agencies, focusing their attention on practices that caused “serious discrepancies” in hiring and promotion rates.⁸ The Guidelines specifically disclaimed

³ Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607 (1978).

⁴ *Id.* § 1607.3.

⁵ *Id.* § 1607.1(B).

⁶ *Id.* § 1607.4(D).

⁷ Uniform Guidelines on Employment Selection Procedures, 43 Fed. Reg. 38290, 38291 (Aug. 25, 1978).

⁸ *Id.*

application to the resolution of individual complaints alleging discrimination.⁹

Even for the enforcement agencies, the “rule” was not controlling. The Guidelines recognized that, rigidly applied, the four-fifths ratio was both under- and over-inclusive. Smaller differences in selection rates could suffice to show adverse impact when requirements of statistical and practical significance were met, and larger differences might not constitute adverse impact when the sample size was small.¹⁰ The Guidelines further recognized that the context mattered. Special recruitment efforts might increase the number of applicants from disadvantaged groups; discriminatory action might discourage them.¹¹ In either case, the pool of applicants would change in ways that would affect the selection ratio.

Because a finding of adverse impact triggered the requirement of validation and the risk of government scrutiny, the four-fifths ratio became a focal point of attention. It was immediately the subject of criticism by scholars and advocates on all sides, who argued that it was highly problematic if deployed as a rule for identifying discrimination.¹² Because validating a test under the Guidelines was technically complex and costly, employers had strong incentives to try to avoid triggering scrutiny in the first place. After the 1970s, however, the federal government’s efforts to combat systemic employment discrimination receded, and the role of the four-fifths ratio in guiding agency discretion became less salient.¹³

B. Proving Disparate Impact in Practice

⁹ 29 C.F.R. § 1607.16(I).

¹⁰ *Id.* § 1607.4(D).

¹¹ *See id.*

¹² *See, e.g.,* Anthony E. Boardman & Aidan R. Vining, *The Role of Probative Statistics in Employment Discrimination Cases*, 46 L. & CONTEMP. PROBS. 189 (1983); Elaine W. Shoben, *Differential Pass-Fail Rates in Employment Testing: Statistical Proof Under Title VII*, 91 HARV. L. REV. 793 (1978); Marion G. Sobol & Charles J. Ellard, *Evaluating the Four-Fifths Rule vs. A Statistical Criterion for the Determination of Discrimination in Employment Practices*, 10 LAB. STUD. J. 153 (1985).

¹³ In a recently issued technical assistance document, the EEOC reiterated that the four-fifths rule is merely a rule of thumb that is not always appropriate to rely on and should not substitute for a test of statistical significance. *See* U.S. EQUAL EMPLOYMENT OPPORTUNITY COMM’N, SELECT ISSUES: ASSESSING ADVERSE IMPACT IN SOFTWARE, ALGORITHMS, AND ARTIFICIAL INTELLIGENCE USED IN EMPLOYMENT SELECTION PROCEDURES UNDER TITLE VII OF THE CIVIL RIGHTS ACT OF 1964 (2023), <https://www.eeoc.gov/select-issues-assessing-adverse-impact-software-algorithms-and-artificial-intelligence-used> [https://perma.cc/N9GD-VSSL].

In the courts, most of the law around disparate impact liability evolved through cases brought by civil rights groups or private litigants. Cases alleging disparate impact discrimination entail a three-step analysis.¹⁴ First, plaintiffs must establish a *prima facie* case, usually by producing statistical evidence that shows an employer practice disproportionately screens out a protected group. Second, employers have the opportunity to defend their practice by showing that it is job-related and consistent with business necessity¹⁵ or “validating” it in the terminology of the Guidelines. Even if they succeed in doing so, plaintiffs may nevertheless prevail by pointing to a less discriminatory alternative that would meet the employer’s business needs.¹⁶ Thus, as with agency enforcement decisions, an important first step is deciding whether there is sufficient evidence—a *prima facie* case—to warrant further legal scrutiny.

Although some have suggested that the “four-fifths rule” should apply, courts have generally not adopted it as the test for establishing a *prima facie* case of disparate impact. The Supreme Court specifically noted that “[the rule] has been criticized on technical grounds . . . and it has not provided more than a rule of thumb for the courts.”¹⁷ Federal courts of appeals have similarly refused to treat selection ratios below four-fifths as the legal test of disparate impact.¹⁸ While a selection ratio that falls below that threshold *can* be sufficient to establish a *prima facie* case,¹⁹ it does not always do so, particularly when sample sizes are small.²⁰ On the other hand, selection ratios above that cutoff do not automatically absolve an

¹⁴ See 42 U.S.C. § 2000e-2(k); *Albemarle Paper Co. v. Moody*, 422 U.S. 405 (1975).

¹⁵ Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(k)(1)(A)(i).

¹⁶ *Id.* § 2000e-2(k)(1)(C).

¹⁷ *Watson v. Fort Worth Bank & Tr.*, 487 U.S. 977, 995 n.3 (1988).

¹⁸ See, e.g., *Jones v. City of Boston*, 752 F.3d 38, 51 (1st Cir. 2014) (“Although the four-fifths rule may serve as a helpful benchmark in certain circumstances, both the Supreme Court and the EEOC have emphasized that courts should not treat the rule as generally decisive.”); *Stagi v. Nat’l R.R. Passenger Corp.*, 391 F. App’x 133, 138 (3d Cir. 2010) (noting that the “four-fifths” rule “has come under substantial criticism, and has not been particularly persuasive, at least as a prerequisite for making out a *prima facie* disparate impact case.”); *Clady v. County of Los Angeles*, 770 F.2d 1421, 1428 (9th Cir. 1985) (“The [‘four-fifths rule’ is] not legally binding.”).

¹⁹ See, e.g., *Vulcan Pioneers, Inc. v. N.J. Dep’t of Civ. Serv.*, 625 F. Supp. 527, 544 (D.N.J. 1985); *M.O.C.H.A. Soc’y, Inc. v. City of Buffalo*, 689 F.3d 263, 274 (2d Cir. 2012).

²⁰ See, e.g., *Mems v. City of St. Paul, Dep’t of Fire and Safety Servs.*, 224 F.3d 735, 740 (8th Cir. 2000); *Frazier v. Consol. Rail Corp.*, 851 F.2d 1447, 1451 (D.C. Cir. 1988); *Fudge v. City of Providence Fire Dep’t*, 766 F.2d 650, 658 n.10 (1st Cir. 1985).

employer.²¹ Simply put, a selection ratio below four-fifths is neither necessary nor sufficient for a finding of disparate impact under current law.

In the litigation context, the role of the *prima facie* case is to determine whether there is sufficient evidence of a disparate impact to warrant requiring employers to defend their employment practice. This inquiry is a retrospective one. It asks whether a particular employment practice, though facially neutral, systematically disadvantaged a marginalized group, and therefore requires justification. Because there is some inevitable randomness in any process, a key question is whether the employer’s practice caused the observed difference in selection rates between groups, or whether those differences could have occurred by chance.²²

One problem with relying on the “four-fifths rule” to identify discrimination is that simply looking at the selection ratio does not consider the statistical significance of the effect it is trying to measure. Put concretely, consider two firms that each have selection rates of 30% and 20% for men and women, respectively. In both cases, the ratio between the selection rates is 0.2/0.3 or 0.67, which is less than 0.8, or four-fifths. In other words, both firms would be considered in violation of the “four-fifths rule.” Suppose, however, that Firm 1’s applicant pool contained 100 men and 100 women (of which it hired 30 men and 20 women), while Firm 2’s applicant pool contained 10 men and 10 women (of which it hired 3 men and 2 women). The disparity in selection rates is clearly more meaningful for Firm 1 than Firm 2, even though they have the same selection ratio. If Firm 2 happened to hire one more woman and one fewer man, it would have an adverse impact against men instead of women according to the “four-fifths rule.” In technical terms, the “four-fifths rule” considers only effect size (how far apart the selection rates are) and not statistical significance (the likelihood of observing the same results by random chance).

Recognizing this limitation, courts have looked to a variety of tests (which we broadly term SPTs) to determine whether an

²¹ See, e.g., *Bew v. City of Chicago*, 252 F.3d 891 (7th Cir. 2001); *Isabel v. City of Memphis*, 404 F.3d 404 (6th Cir. 2004); *City of Boston*, 752 F.3d 38.

²² See, e.g., *Ricci v. DeStefano*, 557 U.S. 557, 587 (2009) (noting that a *prima facie* case of disparate impact is “essentially, a threshold showing of a significant statistical disparity.”); *N.Y.C. Transit Auth. v. Beazer*, 440 U.S. 568, 584 (1979) (explaining that a *prima facie* case is established by statistical evidence showing that the challenged practice has the effect of denying equal access to employment opportunities); *City of Boston*, 752 F.3d 38 (explaining that statistical significance tests whether a correlation could have been observed by chance).

observed difference in selection rates is statistically significant.²³ For example, they have used formal statistical tests such as Fisher's exact test and the chi-squared test to ask whether an observed difference in selection rates is statistically significant using a conventional cutoff like 10%, 5%, or 1%.²⁴ These tests seek to determine whether observed disparities in the selection rates of different groups could have arisen by chance, or were more likely caused by the challenged employment practice.

In addition to statistical significance, some courts also ask whether the magnitude of differences in selection rates is meaningful—i.e., whether it is practically significant. To measure practical significance, courts may refer to a four-fifths ratio, but often use other measures.²⁵

As discussed below, there are other ways to compare the impact of a selection procedure on different groups. Although the law has traditionally focused on SPTs, it does not, as commonly assumed, always use a four-fifths ratio as a cutoff. To be clear, the outcome of a statistical test does not establish whether discrimination occurred. Rather, it is the first step in the legal process of determining whether a given practice is considered discriminatory. Statistical evidence is used to establish a *prima facie* case, which then shifts the burden to the employer to justify the practice. If the employer can demonstrate the validity and necessity of its practice, it generally will avoid liability for discrimination.²⁶ However, facing a *prima facie* case is costly to an employer, who must mount a legal defense and gather evidence supporting its practices. As such, employers face strong incentives to avoid legal jeopardy in the first place by tailoring their practices to avoid a *prima facie* case.²⁷ The legal standard for establishing a *prima facie* case will thus shape how algorithmic hiring tools are developed.

II. THE "FOUR-FIFTHS RULE" AS A FAIRNESS METRIC

²³ See, e.g., *Castaneda v. Partida*, 430 U.S. 482 (1976).

²⁴ See, e.g., *City of Boston*, 527 F.3d at 43.

²⁵ Kevin Tobia, *Disparate Statistics*, 126 YALE L. J. 2382, 2399–2403 (2017).

²⁶ Even if the employer meets its burden of justification, it might nevertheless be found liable if the plaintiff identifies a less discriminatory alternative that the employer refuses to adopt. See 42 U.S.C. § 2000e-2(k)(1)(A)(ii); *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425 (1975). Very few cases, however, succeed by following this route.

²⁷ Raghavan et al., *supra* note 1; Manish Raghavan & Solon Barocas, *Challenges for Mitigating Bias in Algorithmic Hiring*, BROOKINGS (Dec. 6, 2019), <https://www.brookings.edu/research/challenges-for-mitigating-bias-in-algorithmic-hiring/> [<https://perma.cc/7ZP6-Y5JB>].

As machine learning techniques are applied to decision-making in social domains like employment, concerns have grown that predictive algorithms may be discriminatory and unfair. Computer scientists and practitioners have sought methods to ensure that models are fair, and some have looked to the law for guidance.²⁸

As a result, the “four-fifths rule” has sometimes been adopted as a common metric by which the fairness of models is evaluated. Much of the academic research that makes reference to the “four-fifths rule” is not specific to the employment context; instead, compliance with a four-fifths ratio is seen as one possible property for characterizing model fairness across varying applications and contexts.

When developers invoke a statistical test like the “four-fifths rule,” its function is different than in the legal context where the focus is retrospective. Instead of measuring the amount of impact that has occurred after the deployment of a test, it is used as a way of defining fairness or nondiscrimination prospectively when building models. Rather than a way of examining the connection between a practice and an observed disparity, the selection ratio is used as an optimization objective.

A common goal of this research is to develop machine learning algorithms to automatically train models that comply with a fairness metric, such as the “four-fifths rule.” To a first approximation, we can think of traditional machine learning algorithms as following the instruction: “find me the model that makes the most accurate predictions on this data.” Algorithmically enforcing the “four-fifths rule” amounts to modifying that instruction to: “find me the model that makes the most accurate predictions on this data, subject to the constraint that the ratio of selection rates does not fall below four-fifths.”

In the literature, a variety of techniques have been developed to achieve this end. However, many of them directly rely on individual demographic characteristics as an input feature. While there is debate in the legal literature on this point,²⁹ developers are sufficiently

²⁸ See Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger & Suresh Venkatasubramanian, *Certifying and Removing Disparate Impact*, PROC. 21ST ACM SIGKDD INT’L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 259 (2015); Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez & Krishna P. Gummadi, *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment*, PROC. 26TH INT’L CONF. ON WORLD WIDE WEB 1171 (2017).

²⁹ See Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 803–53 (2020); Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action*, 110 CAL. L. REV. 1539 (2022).

concerned that this practice would itself be considered a form of discrimination to limit its practical application in the employment context.

A separate class of techniques, sometimes called “Disparate Learning Processes” or DLPs, also seek to ensure compliance with fairness metrics,³⁰ but do not rely on an applicant’s demographic characteristics when making predictions.³¹ While some scholars have argued that DLPs could bypass legal concerns about relying on protected characteristics,³² these strategies have yet to gain traction in practice.

The “four-fifths rule” is often invoked by practitioners as well. For example, vendors who build predictive algorithms for use in social domains like consumer finance, housing, employment, and criminal law enforcement sometimes refer to the “four-fifths rule” as a measure of legal compliance. Others have promoted “toolkits” that offer generalized approaches to ensure fair algorithms across use cases that refer to a four-fifths ratio.³³

As concerns about algorithmic discrimination have moved into the policy sphere, auditing is emerging as an important governance tool.³⁴ In the absence of well-established auditing standards, once again, some have looked to the “four-fifths rule” as a measure of

³⁰ Faisal Kamiran & Toon Calders, *Classifying without Discriminating*, 2ND INT’L CONF. ON COMPUT., CONTROL AND COMMC’N (2009); Zachary C. Lipton, Alexandra Chouldechova & Julian McAuley, *Does Mitigating ML’s Impact Disparity Require Treatment Disparity?*, 31 ADVANCES NEURAL INFO. PROCESSING SYS. (2018); Zafar et al., *supra* note 28.

³¹ Traditionally, a developer selects a model by choosing one that achieves maximal accuracy over a particular dataset. In contrast, DLPs select a model with both high accuracy and low demographic disparity. The resultant model itself does not take demographic characteristics as inputs, but the development process relies on demographic characteristics when selecting among possible models. These processes could reduce accuracy, though recent evidence suggests that the trade-offs involved may be mild. For a more detailed discussion, see Emily Black, Manish Raghavan & Solon Barocas, *Model Multiplicity: Opportunities, Concerns, and Solutions*, ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 850 (2022).

³² See Zach Harned & Hanna Wallach, *Stretching Human Laws to Apply to Machines: The Dangers of a “Colorblind” Computer*, 47 FLA. ST. U. L. REV. 617 (2022).

³³ See examples cited in Elizabeth Anne Watkins, Michael McKenna & Jiahao Chen, *The Four-Fifths Rule is Not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness*, in PARITY TECHS, INC., TECH. REP. (2022).

³⁴ See, e.g., NEW YORK CITY, N.Y., N.Y.C. ADMIN. CODE 5 §§ 20-870 to -874 (2023).

compliance.³⁵ Other proposals have suggested the use of pre-certification requirements or licensing standards, but implementation of those regimes will also require reference to a substantive measure of nondiscrimination.

The role of the “four-fifths” rule and other SPTs differs somewhat when employment selection tools are developed in practice compared with the automated techniques found in the computer science literature. After developing a machine learning model, developers often run a suite of SPTs that measure differences in selection rates across demographic groups. If significant differences are found, the firm will remove data attributes that contribute to these differences, re-build the model, and repeat until the model passes the tests.³⁶

While firms claim that this procedure is a good-faith attempt to ensure non-discrimination, it may also be motivated by litigation avoidance: if a firm produces a model that passes the statistical test it believes courts will use, then it may be difficult or impossible for a plaintiff to demonstrate a *prima facie* case, and the firm can avoid scrutiny. As discussed above, however, a procedure that satisfies the four-fifths rule of thumb may still warrant further legal scrutiny. The prospective use of the “four-fifths rule” can thus become a strategy for model developers to minimize legal risk without addressing the substantive harms of potential discrimination. In what follows, we demonstrate the limitations of over-reliance on the “four-fifths rule” and other SPTs as tests for discrimination in algorithms.

III. LIMITATIONS OF THE “FOUR-FIFTHS RULE” AND SPTS GENERALLY

As explained above, the four-fifths ratio has become a common metric for evaluating model fairness, even though it does not actually reflect the legal test for disparate impact discrimination. But does that discrepancy matter? One might argue that regardless of its legal status, the four-fifths ratio offers a useful metric for measuring discriminatory effects that should guide model development. This Part explores some of the problems with relying on the “four-fifths rule” or SPTs more generally.

Part III.A explains the limitations of the “four-fifths rule” as a retrospective test of discrimination in the litigation context. Although more statistically robust SPTs alleviate some of these problems,

³⁵ See Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel & Frida Polli, *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, PROC. 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 666 (2021).

³⁶ See BOGEN & RIEKE, *supra* note 1; Raghavan et al., *supra* note 1.

SPTs as a class share certain limitations when applied to algorithms. Part III.B next considers how the use of SPTs as fairness metrics to guide model development *prospectively* raises additional concerns, particularly because they may distort incentives, leading developers to prioritize formal compliance without actually addressing model unfairness.

A. The Prima Facie Case: A Retrospective View

When statistical parity tests are used to determine whether a prima facie case of discrimination exists, they are retrospective in orientation. They take data about actual applicants, examine the outputs of the model, and determine whether it has disproportionately screened out disadvantaged or marginalized groups. When used for this type of retrospective examination, SPTs may be a poor tool for deciding which cases warrant further legal scrutiny, as they can be both over- and under-inclusive. For example, in the case of the “four-fifths rule,” some practices that produce ratios less than four-fifths may not constitute discrimination; other practices with ratios above four-fifths may still warrant close scrutiny. Even when statistical significance is taken into account, SPTs have limitations when used to establish a prima facie case due to several factors: measuring statistical significance is limited by the size of the dataset; the conclusions of SPTs are heavily dependent upon the selectivity of the underlying practice; and SPTs are not easily applied to practices that do not produce binary outcomes.

1. Statistical Significance

On its own, the four-fifths ratio only measures effect size, not statistical significance. For small sample sizes, the four-fifths ratio is thus quite unreliable; the exclusion or inclusion of a single data point can easily alter the conclusions of the test. Recognizing this, courts have relied on other SPTs that provide information on both the magnitude of an effect (how different the selection rates are) and its statistical significance (the likelihood of such an effect occurring due to chance).

While more robust SPTs overcome some of the limitations of relying on the four-fifths ratio, reporting statistical significance is not a panacea. SPTs test whether there is sufficient evidence to reject a “null hypothesis” that a practice is not discriminatory. Even if a practice is discriminatory (meaning the “null hypothesis” is false), a dataset may simply be too small for an SPT to draw a statistically significant conclusion. And, on the other hand, given enough data, an SPT is likely to find statistically significant effects simply because with large amounts of data from the real world, a test can detect even

small, idiosyncratic biases that may not warrant legal action.³⁷ In effect, a statistically significant result from an SPT can be as indicative of sufficiently large samples as it is of discriminatory behavior. In practice, courts do not use a fixed rule when establishing a *prima facie* case; instead, they take both effect size and statistical significance into account when making judgements about whether practices warrant deeper scrutiny, and such an approach would also be appropriate if the challenged practice was an algorithmic system.

2. Selectivity

The effect of applying the “four-fifths rule” also depends upon the selectivity of an employer’s process.³⁸ Examining the *ratio* of selection rates means that more selective processes will more likely result in a finding of disparate impact. Compare, for example, a highly selective process in which 2% of white applicants and 1.5% of Black applicants are hired with a less selective process in which 90% of white and 73% of Black applicants are hired. The first scenario, where the difference in selection rates is 0.5%, falls below the four-fifths ratio because the ratio of selection rates is 0.75. In the second scenario, despite the larger absolute difference in selection rates of 17%, the practice does not violate the “four-fifths rule.” Without knowing more about the specific selection procedure and the type of job at issue, it is difficult to assess whether these judgments are correct. However, it is not at all obvious that the more selective procedure, which affects far fewer Black applicants, poses the more serious threat to equal opportunity.

Some SPTs—e.g., the z-test—are designed to detect differences in selection rates instead of analyzing the *ratio* of selection rates. Again, the selectivity of the procedure influences whether a test will produce statistically significant results. In particular, differences in selection rates are easier—i.e., require less data—to detect when selection rates are extremely low or high—e.g., close to 0 or 1.³⁹ Thus, the sensitivity of an SPT to differences in selection rates depends on the selectivity of the practice, regardless of whether the

³⁷ Rick Jacobs, Kevin Murphy & Jay Silva, *Unintended Consequences of EEO Enforcement Policies: Being Big is Worse than Being Bad*, 28 J. BUS. PSYCH. 467 (2013).

³⁸ Jerard F. Kehoe, *Cut Scores and Adverse Impact*, in ADVERSE IMPACT, *supra* note 2, at 289.

³⁹ This is because distinguishing between distributions is harder when they have high variance, and variance is maximized when selection rates are close to 0.5. For example, detecting differences between selection rates of 0.001 and 0.051 requires less data than distinguishing between 0.5 and 0.55.

test considers the ratio of or difference between the selection rates of different groups.

3. *Beyond Binary Outcomes*

SPTs are easiest to apply when an algorithm sorts candidates into or out of a pool—a binary decision. Algorithmic prediction tools, however, generally produce continuous scores. For example, they may predict the likelihood of particular outcomes—e.g., the probability that this individual will be a successful employee—rather than categorical judgments—e.g., hire or don’t hire. It is up to the humans who design or implement these tools to decide what cut-off score to use to make the selection decision. If the distribution of scores looks different for different groups, then the relative selection rates of different groups will vary depending upon which cut-off score is chosen. Under a legal regime narrowly focused on SPTs, the cutoff might be selected with an eye to equalizing selection rates, even though the underlying rankings significantly favor one group over another. Very often, algorithms are used as screening tools, with subsequent decisions further narrowing the pipeline of candidates. If rankings influence these later decisions, the fact that an early screening tool passes an SPT may obscure from view the effect that unequal predictions may have down the road.

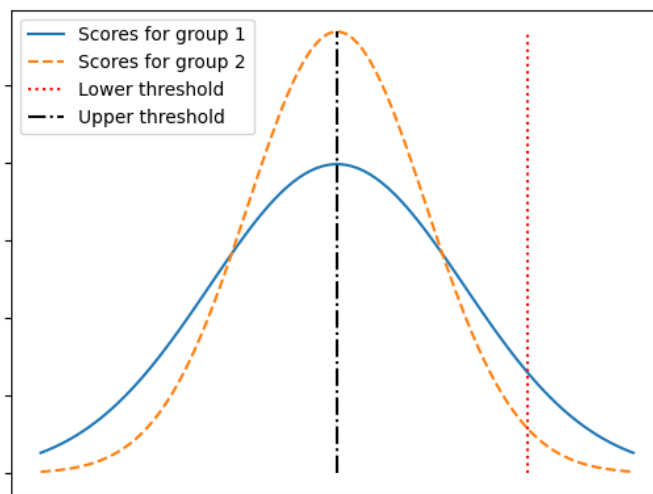


Figure 1: Relative passing rates for different groups depend on the threshold applied.

In some cases, choosing a cutoff score is a reasonable approximation to how employers use models in practice. Some employers simply set thresholds and interview all applicants who score above the threshold. But in other cases, model predictions are used in far more complex ways. An employer might simply rank applicants and interview them sequentially until they make an offer. Or a human evaluator may take the scores into account as one of many factors in their decision-making process. In such cases, running an SPT on the selection rates resulting from a threshold score does not reflect the way in which the algorithm is deployed, and as a result, the test may not accurately capture when disadvantaged groups are adversely impacted.

Finally, a new and growing class of AI techniques does not directly produce numerical estimates of candidate quality, but instead, seeks to infer relevant information for use in making judgments downstream. For example, commercially available resume parsers extract candidates’ skills from their resumes, and newer technologies can generate free-form text about candidates.⁴⁰ These applications raise new concerns, because there is growing evidence that AI-generated text can and often does reflect societal biases.⁴¹ Discussion of bias in these types of systems are beyond the scope of this Article. The important point here is that SPTs are ill-suited for detecting discrimination in these types of AI systems. SPTs do not naturally generalize beyond binary selection decisions, and identifying discrimination in these applications will require a more nuanced approach.

* * * *

When considering retrospective liability, the “four-fifths rule” is a poor test for determining whether a practice caused a discriminatory effect because it is simply too crude a measure. Even more robust SPTs that take into account statistical significance have significant limitations. Thus, although SPTs can be helpful tools, when they are applied rigidly according to hard-and-fast rules, they

⁴⁰ See, e.g., AFFINDA, <https://www.affinda.com/resume-parser> [<https://perma.cc/5D7J-QYVK>]; PARSIO, <https://parsio.io/> [<https://perma.cc/2FRU-HQJA>].

⁴¹ Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani & Adina Williams, “I’m Sorry to Hear that”: Finding New Biases in Language Models with a Holistic Descriptor Dataset, PROC. 2022 CONF. ON EMPIRICAL METHODS NAT. LANGUAGE PROCESSING 9180 (2022); Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency & Ruslan Salakhutdinov, *Towards Understanding and Mitigating Social Biases in Language Models*, PROC. 38TH INT’L CONF. ON MACH. LEARNING 6565 (2021).

lose sight of important context and nuance that are relevant to determining whether a practice warrants further scrutiny.

B. SPTs as Fairness Metrics: A Prospective View

When computer scientists or practitioners invoke the “four-fifths rule” or other SPTs as fairness metrics, they are relying on those statistical measures to guide model development prospectively. Used in this way, additional concerns come into play. The crucial inquiry is no longer “are these tools appropriate for diagnosing discriminatory systems?” but “do they create the right incentives for developing fair models?”

The concern is that developers will focus too narrowly on SPTs, making choices keyed to these metrics, rather than try to understand why disparities are arising and where substantive unfairness may be affecting the selection process. In other words, they may build models to pass statistical tests rather than look for models that will actually reduce inequities when implemented in the real world.

Of course, industrial-organizational psychologists have long considered the four-fifths ratio and other statistical tests prospectively, using them to evaluate selection instruments.⁴² In these traditional practices, however, the test designer evaluates an instrument with a suite of tests including SPTs to see if it is suitable for deployment. The designer typically seeks a qualitative understanding of the performance of the instrument, and makes judgments whether to adjust it, or to adopt some alternative, by weighing validity, adverse impact, and other job-related considerations.

Novel algorithmic techniques to automatically enforce SPTs like the “four-fifths rule” short-circuit this process by substituting qualitative judgments with a mechanically enforced rule. Instead of a person with substantive expertise making a reasoned decision about the trade-offs from one instrument to another, the developer pre-specifies tradeoffs to optimize for compliance with a four-fifths ratio. This automated optimization process reduces search costs but comes at the cost of qualitative understanding. The developer may have little intuition as to what alternative models the algorithm failed to produce.⁴³ If the developer can precisely specify their objective—e.g., the potential trade-off they are willing to make between predictive accuracy and differences in selection rates, then this lack

⁴² See SOC’Y FOR INDUS. AND ORGANIZATIONAL PSYCH., PRINCIPLES FOR THE VALIDATION AND USE OF PERSONNEL SELECTION PROCEDURES (2018); Tippins, *supra* note 2.

⁴³ See Black et al., *supra* note 31, at 850.

of intuition may have little practical impact. But to the extent that a developer is unable to completely specify their preferences—e.g., that the resultant model refrain from heavy reliance on a candidate’s place of education—the developer has little control over the resultant model.

Below, we highlight several substantive limitations of relying on SPTs prospectively. SPTs are insensitive to accuracy: they make no attempt to determine whether an assessment is “correct.” Using SPTs in the model development process thus can create a trade-off between accuracy and fairness, a potentially misleading binary that more nuanced methods might avoid. Because developers have information about outcomes in the training dataset, they can measure properties that are impossible to observe in retrospect. Thus, these more nuanced methods could be incorporated into prospective audits, which we discuss in further detail in Part IV. In addition, because SPTs depend heavily on the representativeness of the test data and how the relevant pool is defined, they can create incentives for data curation and gaming, which undermines their utility in evaluating model fairness prospectively.

1. *Limited Measures of Performance and Bias*

Whether or not a model passes an SPT has little bearing on whether it accurately predicts outcomes. A model that outputs purely random predictions for two demographic groups has no predictive validity, yet it passes any SPT. AI models are typically built to predict “labels” (often denoted by Y), which are simply values for the target outcome of interest that the model is trying to predict. Examples of labels used in employment models include employee retention, job performance measures like sales numbers, and psychometric traits.⁴⁴ A developer seeks to predict the correct label for each individual—e.g., whether the person will still be employed after two years—using available data about that person—e.g., their features.

In order to build a model, a developer uses a dataset—the training data—that contains information about numerous individuals. For each candidate, the data contains information about features (X , composed of x_1 , x_2 , x_3 , etc.) as well as information about the class labels that capture the outcome of interest (Y , employed after two years / no longer employed after two years) for that individual. In employment settings, this dataset often includes each candidate’s demographic information (A), like race and sex, in addition to the features (X) and labels (Y). In this notation, the

⁴⁴ Raghavan et al., *supra* note 1.

developer's goal is to build a model that, given a new candidate's features X , generates a prediction ($\hat{Y}$) of the label for that candidate.

Recall that the four-fifths ratio, and SPTs more generally, tests for disparities in selection rates. Selection rates depend only on predictions ($\hat{Y}$) and demographic characteristics (A), since they only measure the rates at which members from different demographic groups receive positive predictions. Crucially, they do not depend on labels (Y)—i.e., the actual value of the target of interest. This limitation is inherent to ex post evaluation. To see why, consider the concrete case where a model developer seeks to predict Y defined as employee retention time. They build a model based on employee retention data from the past several years. After deploying the model, they form a prediction of retention ($\hat{Y}$) for each applicant, but they do not observe true employee retention (Y) for all applicants. They simply cannot observe retention for candidates who were not hired, meaning it is impossible to evaluate whether their predictions were “correct” for applicants who were screened out. Thus, while labels (Y) are key to model *development*, they cannot be observed across all candidates after model deployment.

However, many important measures of model performance and bias towards or against particular demographic groups, depend not just on $\hat{Y}$ (the model's predictions) and A (demographic information), but also on Y , the *true* value of the outcome of interest. In other words, assessing the performance and bias of a model that predicts retention depends not only on its prediction of whether a given individual will still be employed after two years, but whether that prediction would be correct—i.e., whether that individual would be retained if actually hired.

Measures of predictive validity, or the accuracy of a model, typically involve comparisons between Y and $\hat{Y}$. The closer $\hat{Y}$ is to Y , the greater the predictive validity. Additionally, many widely used notions of test bias from the psychology literature depend on Y , $\hat{Y}$, and a demographic attribute A .⁴⁵ We focus on two in particular: subgroup calibration, which measures whether a given prediction corresponds to similar outcomes of interest for members of different demographic groups, and differential validity, which measures discrepancies in predictive accuracy across groups. These are both important notions that describe whether an assessment unfairly favors one group over another and are typically measured using these

⁴⁵ See James L. Outtz & Daniel A. Newman, *A Theory of Adverse Impact*, in *ADVERSE IMPACT*, *supra* note 2, at 53; Herman Aguinis & Marlene A. Smith, *Balancing Adverse Impact, Selection Errors, and Employee Performance in the Presence of Test Bias*, in *ADVERSE IMPACT*, *supra* note 2, at 403.

three attributes: Y , $\hat{Y}$, and A .⁴⁶ These measures provide a more nuanced understanding of how a model performs for different demographic groups, and have been commonly used throughout both the psychology and computer science literatures.⁴⁷ SPTs cannot capture these important concepts simply because they lack information about Y . We build on this observation in Part IV.

2. Data Representativeness

Algorithm developers sometimes use prospective testing to claim that their models pass SPTs, often the “four-fifths rule.”⁴⁸ Without further context, however, this claim is ill-defined: whether or not a model passes an SPT depends crucially on the data on which it is evaluated. A model may pass an SPT on one source of data and fail it on another. Thus, we cannot conclude that a model in isolation either passes or fails an SPT; instead, we can only evaluate whether a model passes *with respect to a particular dataset*. And, as a result, evaluating models requires examining not only the results of SPTs, but also the dataset to which they were applied.

In the litigation context, in order to determine whether a *prima facie* case exists, courts scrutinize hiring decisions in retrospect. The relevant dataset consists of actual applicants to the position and the decisions made about them. We can determine whether hiring practices satisfied an SPT by examining the outcomes they produced for real people. Algorithm developers, however, are often interested in prospective, as opposed to retrospective, analyses. They want to determine whether a model *will* pass an SPT when it is deployed, not whether it has already done so. In order to make this assessment, an algorithm developer effectively needs to guess what the distribution of future candidates will be, evaluate the model based on this guess, and hope that the guess wasn’t too far off when deploying the model. In practice, a developer might use a dataset comprised of past applicants or collect data from a population that they believe to be

⁴⁶ Christopher M. Berry, *Differential Validity and Differential Prediction of Cognitive Ability Tests: Understanding Test Bias in the Employment Context*, 2 ANN. REVS. ORG. PSYCH. & ORGANIZATIONAL BEHAV. 435 (2015); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, PROC. 1ST CONF. ON FAIRNESS, ACCOUNTABILITY AND TRANSPARENCY 77 (2018); Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 67 LEIBNIZ INT’L PROC. INFORMATICS (2017).

⁴⁷ See, e.g., Berry, *supra* note 46; Kleinberg et al., *supra* note 46; ADVERSE IMPACT, *supra* note 2.

⁴⁸ See, e.g., BOGEN & RIEKE, *supra* note 1; Raghavan et al., *supra* note 1.

representative—i.e., their best guess of what the actual applicant pool will look like.

Importantly, the dataset must be representative of the true population in all respects. Finding a dataset that is demographically representative does not guarantee that the data are representative for other attributes—e.g., education level or work history—that are relevant to the model’s predictions. If well-qualified members from some demographic groups are overrepresented in the dataset, a model may pass an SPT on that dataset but fail to achieve it in practice when the prevalence of qualified applicants drops. Similar challenges exist for techniques like propensity score reweighting designed to make a dataset representative:⁴⁹ while they can re-weight or modify a dataset to be representative along a few known axes, they cannot in general enforce representativeness on all attributes.

Moreover, even if a model passes an SPT for a dataset representative of one context, the same model may fail the SPT for a dataset representative of another. A model that yields statistical parity in New York City may not in Atlanta or Phoenix. Changing conditions over time may mean that this same model would fail to achieve statistical parity in New York City a few years later. As a result, a firm cannot certify that a model passes an SPT in general; it must re-evaluate the model in each context in which it will be deployed. There may be no non-trivial model that simultaneously passes an SPT in two different contexts. This dramatically complicates model evaluation for firms seeking to create off-the-shelf models.

Because firms have considerable discretion in selecting the dataset on which to evaluate a model, it is difficult to know if they have done so in good faith. Regulations that rely on prospective auditing can create incentives to curate datasets that make it “easier” for a model to pass an SPT. If a firm is worried that its model under-selects applicants from a particular demographic group, it may simply add more qualified applicants from that demographic group to its dataset, thereby increasing the group’s measured selection rate on that dataset. For SPTs that measure statistical significance in addition to effect size, firms may rely on smaller datasets since they are less likely to lead to statistically significant results. Of course, simply changing the dataset on which a model is evaluated will not affect its potential for adverse impact in practice; it simply makes the model appear (for the sake of prospective analysis) less discriminatory.

One tempting response to the problems introduced by data collection is to attempt to centralize collection. If a third party—e.g.,

⁴⁹ See Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, 66 J. EDUC. PSYCH. 688 (1974).

a regulator—collects and maintains data, firms will lose their ability to manipulate datasets used for SPTs. This approach faces a major hurdle: datasets used to evaluate a predictive model must contain exactly the information required as input to that model. A model that makes predictions based on recorded video interviews requires a dataset containing such interviews. A model that makes predictions based on questionnaires requires a dataset of responses to questionnaires. Thus, the dataset used for a firm’s model must be specific to the firm in question; a regulator cannot simply collect a common dataset to be used by all firms. Centralized data collection would require the regulator to collect a new dataset for each firm or model to be evaluated, which may be prohibitively expensive or simply infeasible.

3. *Determining the Relevant Pool*

The problem of data representativeness is compounded in situations in which the applicant pool is affected by employer behavior or is otherwise difficult to define. For example, the Guidelines recognize that if an employer has discouraged minority or female candidates from applying, differences in passage rates on an SPT screening test may not accurately measure the overall effect of the employer’s selection practices. And conversely, employers who engage in special recruiting efforts to increase the number of minority or female applicants should not necessarily have their practices judged solely by disparities in selection rates. Similarly, when assessing some algorithmic hiring tools, it may be difficult to determine who should be considered part of the candidate pool.

Consider an algorithm designed to search a platform’s inventory of candidate profiles and recommend to a recruiter the ten best matches for their position. How should we think about the relevant candidate pool in this case? The pool cannot be defined by who submitted an application, because no one did in this type of situation. So, should the relevant pool be the set of all candidates on the platform? Just those who work in the same industry? Or only consider those with appropriate qualifications? Whether or not a model passes an SPT will depend heavily on how we construct this baseline. When used for prospective analysis, an SPT gives a fair amount of flexibility, offering firms an opportunity to choose a baseline that increases their likelihood of passing the test.

* * * *

The “four-fifths rule,” or SPTs more generally, has significant limitations when used prospectively as an optimization metric. Relying solely on such measures ignores other relevant metrics that

may be important for fairness, such as differential validity. At the same time, it risks creating incentives for gaming, encouraging developers to make choices designed to satisfy the tests rather than seeking substantive understanding of the sources of unfairness and addressing them directly.

IV. SPTs AND ALGORITHMIC GOVERNANCE

Given these limitations of the “four-fifths rule” and SPTs more generally, what role should they play in legal and policy efforts to prevent algorithmic discrimination? We argue in Part IV.A that SPTs remain useful in the litigation context when examining practices retrospectively, so long as they are not applied rigidly or mechanically. However, prospective testing of algorithms as part of an auditing requirement presents additional opportunities and limitations. In Part IV.B we assess the advantages and disadvantages of prospective as compared with retrospective testing of algorithms. Drawing on these insights, in Part IV.C, we propose additional concrete steps that should be part of any pre-deployment auditing process to more effectively incentivize fair models.

A. Litigating Discrimination Retrospectively

As explained above, the “four-fifths rule” emerged as a rough indication of when a plaintiff has established a *prima facie* case of disparate impact, warranting further legal scrutiny of an employer’s practice. The “four-fifths rule” has significant limitations, and courts for the most part have recognized that, if mechanically applied, it is far too crude a measure of possible discrimination. More robust SPTs that test for statistical significance also have limitations, but in the litigation context, they provide a reasonable place to start the analysis. Because the inquiry is inherently retrospective and outcomes cannot be observed for the entire population, examining differences in selection rates offers a first cut at the problem. So long as they are interpreted with nuance and attention to context, SPTs can usefully draw attention to situations that warrant greater legal scrutiny.⁵⁰

Despite their limitations, SPTs have some desirable properties. For one, SPTs do not require knowledge of actual outcomes across the population—information that is simply unavailable in some circumstances. Unlike measures such as differential item functioning

⁵⁰ See U.S. EQUAL EMPLOYMENT OPPORTUNITY COMM’N, *supra* note 13 (explaining the role of the “four-fifths rule” and statistical tests in determining whether algorithmic decision-making tools have an adverse impact).

or error rate disparities, SPTs do not measure validity. While that is a limitation in some respects, it has the advantage that SPTs will not be distorted by inaccurate or biased labels. For example, if a firm that discriminated in the past seeks to predict hiring decisions using data about prior decisions, its earlier discrimination will be reflected in how outcomes are labeled. In other words, the labeled data will reflect the results of a biased process, not an objective measure of who would have been the best hires. And to the extent that these labels are systematically biased against one demographic group, measures that take labels into account will fail to detect discrimination. In contrast, SPTs will be unaffected by biased labels because they make no attempt to take labels into account. As a result, SPTs can serve as a check against poor or biased measures of outcomes.

In this sense, SPTs can be viewed as aspirational in nature. Instead of assessing the world as it is by accepting background inequities that may contribute to disparate outcomes, SPTs steer attention towards a world where there are no significant differences between different demographic subgroups. When such differences appear, it triggers questions about why this is occurring—specifically, by requiring employers to show that the differing outcomes are justified because they accurately reflect relevant differences between candidates. Such legal scrutiny creates incentives to examine practices that contribute to inequity and to push for expanded opportunities for those who have historically been underrepresented.

Finally, SPTs can create some benefits by pressuring firms to search for equally accurate models with minimal adverse impact. While the computer science literature has frequently explored trade-offs between reducing adverse impact and validity, recent research indicates that in a given setting, there may exist multiple distinct models that have similar overall performance but varying degrees of outcome disparities across demographic groups.⁵¹ Because models with very similar accuracy can vary dramatically in disparity rates, a litigation framework that looks to SPTs can encourage firms to seek alternative models of comparable performance that minimize adverse impacts.⁵²

⁵¹ See Black et al., *supra* note 31; Charles Marx, Flavio Calmon & Berk Ustunal, *Predictive Multiplicity in Classification*, PROC. 37TH INT’L CONF. ON MACH. LEARNING 6765 (2020).

⁵² Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas & Mingwei Hsu, *Less Discriminatory Algorithms*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4590481 [<https://perma.cc/3QVD-W4C9>].

B. Auditing Algorithms Prospectively

In recent years, it has become increasingly common for researchers and policymakers to call for audits and impact assessments as ways to address algorithmic discrimination.⁵³ An ordinance requiring a pre-deployment audit of employment selection algorithms was passed in New York City,⁵⁴ and numerous proposed laws have included similar provisions.⁵⁵ Auditing requirements, however, are typically vague about what the auditing should entail. In other words, they typically lack detail about what types of analyses should be included in a required audit. In the absence of specific direction, some have turned to the “four-fifths rule” as a pre-deployment test of whether an algorithm discriminates. Given its limitations, we suggest here how pre-deployment audits might go beyond SPTs, maintaining the protections they provide while addressing some of their deficiencies. The goal of auditing should be to provide more meaningful information and incentivize the creation of fair algorithms.

In Part III, we explained how some measures of model performance such as subgroup calibration and differential validity cannot be performed on data generated by a model deployed in the real world. These measures require information about labels (Y), which are unavailable when a model is actually deployed to make decisions because some candidates will not be selected and their outcomes cannot be observed.

But this limitation does not apply to prospective testing during model development. A model developer generally has a dataset that *does* contain information about actual outcomes (Y) for the training data in addition to predicted outcomes ($\hat{Y}$) and demographic

⁵³ See, e.g., Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth & Hannah Wallack, *Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs*, PROC. 2021 AAAI/ACM CONF. ON AI, ETHICS, AND SOC’Y 368 (2021); Sasha Costanza-Chock, Inioluwa Deborah Raji & Joy Buolamwini, *Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem*, ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1571 (2022); Briana Vecchione, Karen Levy & Solon Barocas, *Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies*, CONF. ON EQUITY AND ACCESS IN ALGORITHMS, MECHANISMS, AND OPTIMIZATION 1 (2021).

⁵⁴ NEW YORK CITY, N.Y., N.Y.C. ADMIN. CODE 5 §§ 20-870 to -874 (2023).

⁵⁵ See, e.g., Algorithmic Accountability Act, H.R. 6580, 117th Cong. (2022); American Data Privacy and Protection Act, H.R. 8152, 117th Cong. (2022).

information (A). For example, a developer building a model to predict retention necessarily has historical retention data (Y), which is the objective that the model is designed to predict. Label information (Y) is thus available to a model developer before deployment but is generally unobservable for the actual applicant population because not all applicants will be hired. In other words, a model developer can **prospectively** measure properties that are impossible for a regulator to measure after the fact.

There are two key challenges in using prospective evaluation to assess the potential for a model to discriminate. First, labels are rarely objective. For example, suppose a model developer seeks to predict performance reviews. To the extent that performance reviews in the training set systematically and unfairly undervalue members of a demographic group, a model can reproduce these patterns. This problem, often referred to as “label bias,” is in general difficult to identify in datasets. Below, we discuss heuristics developers can use to look for it, with the caveat that technical methods alone cannot fully capture label bias.

Second, prospective evaluation suffers from a data-dependence problem: a firm cannot be sure that the datasets on which they build and evaluate models will be representative of the true data distribution. If the candidates who apply to a position differ dramatically from those on whom the model was developed and evaluated, a firm cannot guarantee that its conclusions drawn from prospective testing will hold in practice.

This analysis highlights the trade-offs between ex ante and ex post evaluation. Ex ante, we can test for a broader range of relevant properties, particularly those that require information about labels. But the conclusions we draw are only valid insofar as the labels are unbiased and the true candidate distribution resembles the dataset used for ex ante evaluation. In contrast, we perform ex post evaluation on the actual set of applicants, meaning we do not need to guess what the candidate distribution will be. However, the lack of labels makes it impossible to evaluate important measures of test bias, such as differential validity. The following table summarizes the trade-offs between prospective and retrospective analyses:

	Ex ante, prospective evaluation of proposed model	Ex post, retrospective evaluation of deployed model
Labeled data	Available	Not available
Tests of validity, differential validity, and subgroup calibration	Possible	Not possible
Risk from label bias	Yes	No
Data representative of actual population	Not necessarily	Yes

Table 1: Properties of ex ante and ex post model evaluation

C. A Concrete Technical Proposal

Given these trade-offs, SPTs likely should remain a part of the auditing process; however, we propose three additional concrete measures that should be part of pre-deployment audits in order to take advantage of the information available during model development and to fill the gaps if only SPTs are considered.

1. **Predictive validity.** For models with binary outputs, this would include measures of error rates like precision and recall.⁵⁶ For models with continuous outputs, firms should report global performance measures such as ROC-AUC (which aggregates predictive quality across all possible thresholds) in addition to performance measures like precision and recall that are specific to a threshold. Crucially, the thresholds used to report model performance should reflect the intended thresholds for use in practice. If a model is to be used for ranking instead of classification, firms can use performance measures from the information retrieval literature designed to measure ranking systems.⁵⁷

⁵⁶ Precision is defined as (number of correct positive predictions) / (number of positive predictions). Recall (also known as true positive rate) is defined as (number of correct positive predictions) / (number of positive instances).

⁵⁷ See generally CHRISTOPHER D. MANNING, PRABHAKAR RAGHAVAN & HINRICH SCHÜTZ, INTRODUCTION TO INFORMATION RETRIEVAL (Cambridge Univ. Press 2008).

2. **Differential validity.** Firms should report validity disaggregated by demographic groups. Note that this is substantially different from what SPTs measure: a model may select members of different demographic groups at the same rate but have far worse predictive validity on one group than another, leading to negative downstream consequences. While some psychologists have noted that testing for differential validity has historically been uncommon in employment settings,⁵⁸ recent work in machine learning has demonstrated that when models are trained on datasets with disparate amounts of data from different demographic groups, they can exhibit large performance disparities in practice.⁵⁹
3. **Subgroup calibration.** For models with continuous outputs, firms should be required to report whether, conditioned on receiving the same predictions, members of different demographic groups exhibit differences in their labels. In the industrial-organizational psychology literature, this is often known as “test bias,” and researchers have developed a variety of measures to quantify it.⁶⁰ While many of these methods are based on regression, as opposed to the more sophisticated machine learning methods deployed today, the computer science and statistics literatures propose alternative frameworks to assess miscalibration, and other firms have developed tools to report it.⁶¹

⁵⁸ See Michael A. Mcdaniel, Sven Kepes & George C. Banks, *The Uniform Guidelines Are a Detriment to the Field of Personnel Selection*, 4 INDUS. ORGANIZATIONAL & PSYCH. 494 (2011).

⁵⁹ See Buolamwini & Gebru, *supra* note 46; Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky & Sharad Goel, *Racial Disparities in Automated Speech Recognition*, 117 PROC. NAT’L ACAD. SCI. 7684 (2020).

⁶⁰ Berry, *supra* note 46; Ben Hutchinson & Margaret Mitchell, *50 Years of Test (Un)Fairness: Lessons for Machine Learning*, ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 49 (2018); ADVERSE IMPACT, *supra* note 2.

⁶¹ For example, Meta’s Fairness Flow considers subgroup miscalibration. See Isabel Kloumann & Jonathan Tannen, *How We’re Using Fairness Flow to Help Build AI That Works Better for Everyone*, FACEBOOK <https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/> [https://perma.cc/DMF7-BDQG].

Beyond these quantitative measures, firms can and should take additional steps to guard against label bias. Firms should document their label definitions, data sources, and model development process. The computer science literature contains a variety of documentation tools that have been deployed across a wide range of contexts.⁶² Moreover, firms can attempt to identify label bias by measuring whether a dataset admits differential prediction, which occurs when the most predictive model for one demographic group differs dramatically from the most predictive model for another.⁶³ Differential prediction provides evidence that similar individuals (in terms of their features X) receive different labels Y, which can indicate label bias. However, differential prediction is typically measured in the context of linear models; adapting these measures to more general classes of machine learning models remains an area of active research.⁶⁴

While these additional measures go a long way towards addressing some of the limitations of SPTs, they do not solve all of the challenges inherent in relying on pre-deployment audits to identify discriminatory models. For this reason, “passing” an audit should not shield a firm from later inquiries about whether its algorithm discriminates.⁶⁵ If an audit conferred legal immunity, firms would be incentivized to manipulate the audit process, undermining its utility and potentially allowing discriminatory algorithms to escape scrutiny.

One significant limitation is inherent to prospective evaluation: in order to evaluate a model before it is deployed a firm must effectively guess the applicant distribution in order to collect a representative dataset. To the extent that they guess wrong, conclusions derived from an audit may be invalid. Even with the best intentions, a firm may simply not know what the applicant pool will

⁶² See, e.g., Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III & Kate Crawford, *Datasheets for Datasets*, 64 COMM’N ACM 86 (2021); Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elenna Spitzer, Inioluwa Deborah Raji & Timnit Gebru, *Model Cards for Model Reporting*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 220 (2019).

⁶³ Mcdaniel et al., *supra* note 58.

⁶⁴ See *id.*; Black et al., *supra* note 31; Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page & Jude Shavlik, *Support Vector Machines for Differential Prediction*, 8725 MACH. LEARNING KNOWLEDGE DISCOVERY DATABASES 50 (2014).

⁶⁵ The only currently enacted legislation in the US requiring audits for hiring algorithms does not foreclose a later legal challenge if the algorithm turns out to be discriminatory in operation. NEW YORK CITY, N.Y., N.Y.C. ADMIN. CODE 5 §§ 20-870 to -874 (2023).

look like. And if audit requirements guaranteed legal immunity, firms would have an incentive to curate a dataset that yields the desired results rather than to select the most likely representative dataset to accurately diagnose the risks of bias. Data-dependence is thus a key limitation on how useful audits can be. Designing statistical techniques to determine whether the dataset used for an audit is sufficiently representative of the actual applicant pool in hindsight is an important direction for future work.

In order for audits to be effective in preventing biased algorithms, they must be conducted with due diligence and in good faith. Deploying firms and the contexts in which algorithms operate are heterogeneous enough that audits cannot be standardized.⁶⁶ Even where precise technical specifications are possible, firms retain a great deal of latitude to make choices, including the relevant candidate pool, which outcomes to report, thresholds to set, and the exact metrics they choose to report. This discretion creates challenges for designing a meaningful audit process. Audits could be conducted by the firm itself, a third-party, or a government regulator, and each of these approaches has advantages and drawbacks. However these issues are resolved as a matter of regulatory design, at a minimum, firms should be required to document and disclose the choices made in conducting the audit in order to enhance its reliability and trustworthiness.

For all these reasons, the requirements for pre-deployment audits should not be seen as metrics that guarantee that models will not discriminate. Instead, they should be crafted with an eye to creating incentives for firms to understand the risks of bias and to make choices that minimize those risks, while increasing the transparency of the model building process.

CONCLUSION

Firms have relied on compliance with the “four-fifths rule” to develop and optimize models for algorithmic hiring in the hopes of avoiding legal liability. The four-fifths ratio was already a poor test for determining whether practices warrant close legal scrutiny. Its use prospectively as a measure of fairness to shape how algorithms are built raises additional concerns. And yet, as firms and vendors increasingly reference the so-called rule, it risks becoming the de facto standard for legal compliance, creating incentives to comply with what appears to be the letter of the law without addressing substantive questions of discrimination. Even though more robust statistical tests can reduce some of the problems with a crude four-fifths ratio, SPTs generally have limitations when used as a

⁶⁶ Barocas et al., *supra* note 53.

retrospective test for the existence of discrimination. And when it comes to prospective evaluation of algorithms, testing requirements can and should go beyond SPTs. Firms can leverage the additional information available during model development to provide a more nuanced picture, which we have detailed above, of how an assessment performs for members of different demographic groups. Similarly, to the extent that regulators impose auditing requirements on firms, these audits should include a more comprehensive set of tests. While discrimination cannot be reduced to a suite of statistical tests, developers and regulators have at their disposal multiple tools to assess the performance and fairness of algorithmic assessments beyond the simple “four-fifths rule.”