South Carolina Law Review

Volume 71 | Issue 2 Article 9

Winter 2019

Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms

Matthew U. Scherer

Allan G. King

Marko J. Mrkonich

APPLYING OLD RULES TO NEW TOOLS: EMPLOYMENT DISCRIMINATION LAW IN THE AGE OF ALGORITHMS

Matthew U. Scherer,* Allan G. King** & Marko J. Mrkonich***

I. INTRODUCTION

Companies, policymakers, and scholars alike are paying increasing attention to the use of machine learning (ML) in recruitment and hiring, most notably in the form of ML-based employee selection tools that use algorithms in place of traditional employment tests and the judgment of human recruiters.[1] To its advocates, ML-based selection processes can be more effective in choosing the strongest candidates, increasing diversity, and reducing the influence of human prejudices.[2] Many observers, however, express concern about other forms of bias that can infect algorithmic selection procedures, leading to fears regarding the potential for algorithms to create unintended discriminatory effects, reinforce existing patterns of discrimination, or mask more deliberate forms of discrimination.[3]

In the Authors' experiences, most employers very much want to improve diversity and inclusion, from company leadership down to the most junior hourly employees. Companies pursue these objectives not just to avoid legal liability for violating antidiscrimination statutes, but also because they have concluded that a more diverse and inclusive workforce is better from both a

* Analytics Associate, Littler Mendelson, P.C.
** Shareholder, Littler Mendelson, P.C.
*** Shareholder, Littler Mendelson, P.C.

1. *See* Gil Press, *120 AI Predictions for 2019*, FORBES (Dec. 9, 2018, 12:00 PM), https://www.forbes.com/sites/gilpress/2018/12/09/120-ai-predictions-for-2019/#54aa4326688c [https://perma.cc/XTS4-EL9P]; Neelie Verlinden, *Machine Learning in Recruitment & How to Do It Right*, HARVER (Oct. 23, 2018, 3:53 PM), https://harver.com/blog/machine-learning-in-recruitment/ [https://perma.cc/RPR3-T29M] (discussing how machine learning uses algorithms).

2. *See* HAIYAN ZHANG ET AL., THE ROLE OF AI IN MITIGATING BIAS TO ENHANCE DIVERSITY AND INCLUSION 6 (2019).

3. *See id.* at 8.

business perspective[4] and an ethical perspective.[5] Indeed, many (if not most) of the employers who are turning to algorithmic and data-driven selection tools are doing so in part because they want to guard against human biases that can serve as barriers to employment for disadvantaged groups.[6]

Not coincidentally, the eradication of such barriers is, as the Supreme Court long ago recognized, the overarching objective of antidiscrimination laws.[7] Because this is an area where the objectives of the law and of America's businesses are well-aligned,[8] one would think that the law should serve as an inducement rather than a deterrent to companies who wish to deploy algorithmic selection tools that will allow them to improve both the quality and diversity of their employees. Unfortunately, that has not been the case.[9]

The rules governing employment tests and other employee selection procedures were developed in the 1970s and have remained largely unchanged in the decades since.[10] Those rules, written as they were for paper-and-pencil tests and other in-person examinations, are ill-suited for selection procedures that rely on a candidate's historical data rather than real-time observations and firsthand assessments. Complicating matters further, the complexity of the algorithms that underlie ML-based selection tools makes it

---

4. *See, e.g.*, Dennis Nally, *Five Reasons Why Diversity and Inclusion Matter to Every Business—and Every Employee*, PWC: CEO INSIGHTS (June 15, 2015), https://pwc.blogs.com/ceoinsights/2015/06/five-reasons-why-diversity-and-inclusion-matter.html [https://perma.cc/N7W9-EPGJ] (noting that 85% of surveyed CEOs whose companies have diversity and inclusion strategies say that it has "improved their bottom line"); DELOITTE, 2017 BOARD DIVERSITY SURVEY: SEEING IS BELIEVING 8 (2017) (noting that over 90% of respondents believe that greater diversity on a company's board of directors enables an organization to improve its ability to innovate, ability to manage disruptions, and overall business performance). *See generally* MARK KAPLAN & MASON DONOVAN, THE INCLUSION DIVIDEND (2013) (noting that diversity and inclusion should lead to a decrease of expenses or an increase of revenue).

5. *See* Nally, *supra* note 4 ("Diversity and inclusion are quite simply the right thing to do[.] It's about creating equal opportunities for everyone—and we can all see signs of progress. But the statistics make it equally clear that there's still a long way to go.").

6. *See* Nicole Lewis, *Will AI Remove Hiring Bias?*, SHRM (Nov. 12, 2018), https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/will-ai-remove-hiring-bias-hr-technology.aspx [https://perma.cc/FM3Y-4WL8].

7. Griggs v. Duke Power Co., 401 U.S. 424, 429–31 (1971) ("The objective of Congress in the enactment of Title VII is plain from the language of the statute. It was to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees . . . What is required by Congress is the removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to discriminate on the basis of racial or other impermissible classification.").

8. *See* Lewis, *supra* note 6 ("A study from McKinsey and Company found that companies that have a diverse workforce financially outperform companies that don't.").

9. *See id.* ("Discrimination in hiring, however, is proving difficult to reverse.").

10. Michael A. McDaniel et al., *The* Uniform Guidelines *Are a Detriment to the Field of Personnel Selection*, 4 INDUS. & ORGANIZATIONAL PSYCHOL. 494, 507 (2011).

difficult for employers and employees alike to discern how and why an algorithm came up with its scores, rankings, or recommendations.[11] This Article seeks to both highlight the challenges employers, workers, courts, and agencies will face as companies develop and deploy algorithmic selection tools, and propose a framework through which courts and agencies can assess whether such tools comply with antidiscrimination laws.

Part II begins with a brief overview of the technological concepts that underlie algorithmic employee selection procedures. It continues with a discussion of the development of antidiscrimination laws, along with the broader philosophical and legal principles that animate the two major forms of employment discrimination—disparate treatment and disparate impact. Part III details why algorithmic selection procedures fit poorly into the legal framework that has developed around Title VII and similar antidiscrimination laws.

Part IV proposes a uniform analytical framework through which agencies and courts can analyze whether an employer using a particular algorithmic selection tool has engaged in disparate treatment or disparate impact. The proposed framework is built upon three unifying themes:

- In light of the low practical value of statistical significance tests in the age of Big Data, flexible tests of reasonableness should replace inflexible tests of statistical significance when assessing whether a correlation is legally meaningful;

- Courts and agencies should recognize certain forms of ML-based fairness techniques as acceptable ways for employers to mitigate disparate impacts without exposing themselves to disparate treatment liability; and

- Standards for assessing the validity of an algorithmic selection procedure should focus on whether the procedure is based on the essential and important job functions of a particular position, as identified through an adequate job analysis and incorporated into a properly constructed model.

This framework, we posit, would give full effect to the objectives of antidiscrimination laws without discouraging employers from using machine learning and Big Data not only to increase efficiency, but also to improve

11. *See* Gregory Barber, *Shark or Baseball? Inside the 'Black Box' of a Neural Network*, WIRED (Mar. 6, 2019, 12:00 PM), https://www.wired.com/story/inside-black-box-of-neural-network/ [https://perma.cc/43U5-TMVD] (noting that how neural networks work is still a mystery).

diversity and reduce the effects of human biases on the recruitment and hiring process.

II. BACKGROUND

A. Algorithmic Selection Tools

Many of the principles discussed in this Article will be relevant to all forms of data-driven employee selection procedures. But the primary focus will be on algorithmic tools[12] that utilize machine learning with a particular emphasis on those that use deep learning. Tools that rely on these sophisticated algorithmic methods pose challenges that exceed those of earlier generations of data-driven employee selection tools.[13]

1. Machine Learning and Deep Learning

Machine learning is a branch of artificial intelligence consisting of algorithms that learn from data.[14] In this context, "learn" means that the algorithm uses statistical methods and "data-driven insights" to allow an AI system to improve itself without human intervention.[15] A learning algorithm uses training data to build a statistical model that can then be used to make predictions or other decisions about new data.[16] Learning algorithms may entail varying levels of mathematical and computational complexity. The highest profile breakthroughs in artificial intelligence over the past several years have come from a subfield of machine learning known as deep

12. This Article will generally use the term "tool" to refer to such algorithmic programs both as a convenient shorthand and to suggest that for the foreseeable future, employers likely will be using algorithmic selection procedures primarily to supplement or improve their existing human-driven employee selection process. This Article will also use the terms "tool," "test," and "selection procedure" interchangeably to refer to algorithmic systems and programs that are used to make, or to help an employer make, personnel decisions. For the most part, this Article will focus on hiring, but in most cases, the same principles will apply to the use of algorithms to make decisions regarding placement, compensation, promotion, termination, transfer, or other actions that affect aspects of employment.

13. *Cf.* ZHANG ET AL., *supra* note 2 (noting that AI algorithms are reliant on human data and can lead to biases).

14. IAN GOODFELLOW ET AL., DEEP LEARNING 96 (2016); Verlinden, *supra* note 1.

15. Tobias Baer & Vishnu Kamalnath, *Controlling Machine-learning Algorithms and Their Biases*, MCKINSEY & CO. 1 (Nov. 2017), https://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases# [https://perma.cc/745L-C45M].

16. *Learning Algorithm*, TECHOPEDIA, https://www.techopedia.com/definition/33426/learning-algorithm [https://perma.cc/53NE-ACH8/].

learning.[17] Deep learning involves the use of artificial neural networks that are inspired by how neurons in the human brain are thought to interact with each other.[18] A neural network operates by taking certain data as an input and passing that data through one or more layers of artificial "neurons" that analyze and transform the data.[19]

Most machine learning approaches can be classified under one of two broad headings: supervised learning or unsupervised learning.[20] In supervised learning, the training data is labeled by humans.[21] In unsupervised learning, by contrast, the algorithm proceeds by looking for patterns in unlabeled data.[22] In general, supervised learning techniques are better suited for applications where the developers are interested in predicting a specific outcome.[23] For example, to build an algorithm that takes photographic images as inputs and that will output a prediction as to whether the image contains a cat, a sensible approach would be to use supervised learning where the training data consists of images that humans have reviewed and labeled as "cat" or "not a cat." On the other hand, an unsupervised learning algorithm might be an appropriate choice for a more general object-recognition algorithm, where the algorithm would receive unlabeled images as input, examine the content of each image, and identify groups of images that it identifies as having shared characteristics.[24]

In technical parlance, the data sets used to train learning algorithms are said to consist of "instances" (also known as examples, observations, subjects,

17. Tom Simonite, *The WIRED Guide to Artificial Intelligence*, WIRED (Feb. 1, 2018, 9:22 AM), https://www.wired.com/story/guide-artificial-intelligence/ [https://perma.cc/QR4E-CGXB].

18. *See* Chris Nicholson, *A Beginner's Guide to Neural Networks and Deep Learning*, SKYMIND, https://skymind.ai/wiki/neural-network [https://perma.cc/4YFU-XS7N] [hereinafter *A Beginner's Guide*].

19. *Id.*

20. *Id.* A third type of machine learning is *reinforcement learning*. A reinforcement learning algorithm is conceptually similar to a supervised learning algorithm except that, rather than using human-labeled training data for optimization, it optimizes itself by updating its decision-making strategy in response to feedback received on prior decisions, usually via a reward function. *See* Chris Nicholson, *A Beginner's Guide to Deep Reinforcement Learning*, PATHMIND, https://pathmind.com/wiki/deep-reinforcement-learning [https://perma.cc/4N7A-PMEP]. This type of machine learning has become prominent in many spheres, but it seems unlikely to have much application in the context of recruitment and hiring algorithms in the near future and thus is not discussed further in this Article.

21. *Id.*

22. *Id.*

23. *See* Chris Nicholson, *A Beginner's Guide to Supervised Learning*, SKYMIND, https://skymind.ai/wiki/supervised-learning [https://perma.cc/57JY-JAN3].

24. *See A Beginner's Guide*, *supra* note 18.

or units) and "attributes" (also known as features or covariates).[25] Instances generally correspond to the rows on a spreadsheet[26] and, for purposes of the types of tools that are the subject of this Article, most often represent individual persons. Attributes are the measurable properties and characteristics of interest for each instance and are analogous to column headings in spreadsheets, such as "educational attainment" or "years of experience."[27] The number of attributes included in a dataset is referred to as the "dimensionality" of the data set.[28]

While a detailed description of deep learning architectures is beyond the scope of this Article, a few characteristics are notably relevant to the legal challenges that employers using algorithmic selection tools will likely face. Deep learning uses neural networks and various mathematical and statistical techniques to determine a set of parameters that an algorithm can use to make predictions based on a given set of input attributes.[29] To determine that optimal set of parameters, deep learning uses the neural network to combine, abstract (and recombine and re-abstract), and otherwise transform the input attributes as they pass through multiple layers of the neural network.[30] This process is repeated thousands or millions of times, with the algorithm making slight adjustments to the parameters during each iteration.[31] The process continues until the model finds an optimal set of parameters—that is, until the model reaches a point where further slight adjustments to the parameters will no longer improve the model's accuracy on the training data.[32] The resulting parameters are what the algorithm ultimately uses to make predictions.[33]

Importantly for legal purposes, the optimized parameters cannot be expressed easily and reliably in terms of the original attributes that were used as inputs, particularly if the algorithm regularly receives new training data. The complexity of the calculations embedded in the deep learning process

---

25. Jason Brownlee, *Data, Learning and Modeling*, MACHINE LEARNING MASTERY (Jan. 6, 2017), https://machinelearningmastery.com/data-learning-and-modeling/ [https://perma.cc/3VTW-XK3P].

26. *Id.*

27. *See id.*

28. Stephanie Glen, *Dimensionality & High Dimensional Data: Definition, Examples, Curse of*, STATISTICS HOW TO (Oct. 10, 2016), https://www.statisticshowto.datasciencecentral.com/dimensionality/ [https://perma.cc/F363-VY4S].

29. *See A Beginner's Guide*, *supra* note 18.

30. *See id.*

31. *See id.*

32. *See id.* The optima generally referred to here are local optima, rather than global or absolute optima. This feature has important implications for how courts should assess algorithmic selection tools under antidiscrimination laws. *See infra* Section IV.E.

33. *See A Beginner's Guide*, *supra* note 18.

means that the algorithm generates the parameters that will not be readily interpretable, and the exact path through which the algorithm arrived at those parameters might not be practically traceable or capable of reconstruction. Consequently, even if the developer of an algorithm knows and understands all of the input variables (hardly a given in the age of Big Data) and also knows the target variables (or criteria) on which the algorithm optimizes, the algorithmic tool may nevertheless be effectively opaque even to the developer, much less the broader public. That is why deep learning algorithms are often referred to as "black box" algorithms.[34] Once the developer has specified the target (or criterion) by which success is judged, and selected the attributes that are potential predictors, the means by which the algorithm determines the parameters that result in the most accurate predictions is opaque.[35]

2. *Algorithmic Employee Selection*

This Article is focused on algorithmic tools designed to make predictions about job candidates' suitability for particular jobs. Today, building such a prediction system is generally best accomplished through supervised learning. The training data for a particular job will generally consist of historical examples of employees who have held the same job or a similar job, and possibly candidates who have applied for such jobs but who were not ultimately hired. In such a data set, the instances in the training data are individual employees or candidates, while the attributes consist of data on various characteristics of those employees or candidates. The labels for this training data would be information indicating each employee or candidate's actual or projected performance in the job.

As an example, say that an employer wants to predict job candidates' future job performance based on their educational attainment and experience. For training data, the employer has a data set consisting of 100 current employees' educational attainment and years of experience at the time of hire, with each employee labeled with their most recent performance rating. In this example, the 100 employees are the instances for the training data, whereas educational attainment, performance rating, and years of service are attributes. If this data were used to build a standard statistical model (not necessarily one that uses machine learning),[36] performance rating would be termed the target

34. *See* Barber, *supra* note 11.

35. *See generally id.* (discussing the difficulty of determining how neural networks generate complex outcomes).

36. In fact, it would make little sense to use deep learning, rather than simple regression analysis, to create a predictive model on such a simple data set.

variable while educational attainment and years of service would be termed the predictor variables.[37]

But in deep learning, and as stated above, the algorithm ultimately makes its predictions by using the final set of parameters that the trained algorithm generates rather than directly using the original input attributes.[38] Those original inputs are the raw materials for the resulting model, but the algorithm transforms them into something unrecognizable when it actually constructs the model.[39] Consequently, the final parameters, rather than the original attributes that the employer included in the training data are, in some sense, the true predictors. For that reason, we more accurately refer to the original attributes as input variables rather than predictor variables for the remainder of this Article.

In recruitment and hiring, available attributes most often include job-relevant characteristics such as certifications and prior employers—i.e., information that can be drawn from a candidate's resume or application. If it is being developed by a third party, the training data may include employees from several different companies. In either case, employers may have the ability to access or acquire data from other sources on many more attributes—which may or may not be job related—such as a candidate's social media profiles, criminal history, and web browsing history. Consequently, the data sets on which the models are trained may have a very high dimensionality and include inputs with no obvious connection to job performance. Some may contain thousands of candidates with thousands of attributes (or more). This makes algorithmic selection procedures considerably more complex than aptitude tests and other traditional employee selection tools.

B. Law of Discrimination

The seminal event in the history of employment discrimination law was the passage of the Civil Rights Act of 1964. Title VII[40] of that statute made it unlawful for employers to, among other things, "fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of

37. *See* ANDREW BRUCE & PETER BRUCE, PRACTICAL STATISTICS FOR DATA SCIENTISTS ch. 1 (2017) (ebook) (describing the process of linear regression).

38. *See* John Villasenor, *Artificial Intelligence and Bias: Four Key Challenges*, BROOKINGS: TECHTANK (Jan. 3, 2019), https://www.brookings.edu/blog/techtank/2019/01/03/artificial-intelligence-and-bias-four-key-challenges/ [https://perma.cc/3HMX-ADU3].

39. *See id.*

40. Civil Rights Act of 1964, Pub. L. No. 88-352, tit. VII, 78 Stat. 241, 253-66 (codified as amended at 42 U.S.C. §§ 2000e–2000e-17 (2012)).

employment, because of such individual's race, color, religion, sex, or national origin."[41] Various other federal statutes have been passed over the years creating additional protected categories, including age (under the Age Discrimination in Employment Act, or ADEA) and disability (under the Americans with Disabilities Act, or ADA), and many states have their own antidiscrimination laws covering different or additional protected categories.[42]

But more than half of a century after its enactment, Title VII remains the most prominent antidiscrimination law and has the most fully developed legal framework for assessing employee selection procedures. Title VII is generally described as having two basic prohibitions, termed "disparate treatment" and "disparate impact."[43] The Supreme Court introduced the disparate impact doctrine in 1971, framing it, in essence, as a logical corollary to the general bar on discrimination "because of" a protected characteristic.[44] But in the ensuing decades, courts drew increasingly stark contrasts between the disparate treatment and disparate impact theories of discrimination,[45] culminating in a 2009 Supreme Court decision, *Ricci v. DeStefano*, where the high court described the inclusion of both theories in Title VII as a "statutory conflict."[46] Navigating this intersection will be a key challenge for employers seeking to implement algorithmic selection procedures.

41. 42 U.S.C. § 2000e-2(a)(1) (2012).

42. Age Discrimination in Employment Act of 1967, 29 U.S.C. § 623(a) (2012) ("It shall be unlawful for an employer . . . to fail or refuse to hire or to discharge any individual or otherwise discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's age . . . ."); Americans with Disabilities Act of 1990, 42 U.S.C. § 12112(a) (2012) (stating that it is unlawful for an employer to "discriminate against a qualified individual on the basis of disability in regard to job application procedures, the hiring, advancement, or discharge of employees, employee compensation, job training, and other terms, conditions, and privileges of employment."). *See generally Discrimination – Employment Laws*, NCSL (July 27, 2015), http://www.ncsl.org/research/labor-and-employment/discrimination-employment.aspx [https://perma.cc/VD6E-W79D] (listing employment discrimination laws from the states).

43. JOSEPH A. SEINER, EMPLOYMENT DISCRIMINATION: PROCEDURE, PRINCIPLES, AND PRACTICE 80, 170 (2015) [hereinafter EMPLOYMENT DISCRIMINATION].

44. *See* George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313, 2314 (2006) (citing Griggs v. Duke Power Co., 401 U.S. 424, 429–30 (1971)); 42 U.S.C. § 2000e-2(a)(1).

45. *See* Rutherglen, *supra* note 44 ("A strictly chronological account of these developments would reveal a very checkered history, with decisions to adopt or reject liability for disparate impact soon followed by qualifications and limitations.").

46. Ricci v. DeStefano, 557 U.S. 557, 580 (2009).

1. Disparate Treatment

Title VII's prohibition against disparate treatment derives from the original text of § 703(a), which prohibits employers from taking any adverse action against an employee or applicant "because of" a protected characteristic.[47] Another provision in § 703 reinforces this primary prohibition by stating the following:

> Nothing contained in this subchapter shall be interpreted to require any employer . . . to grant preferential treatment to any individual or to any group . . . on account of an imbalance which may exist with respect to the total number or percentage of persons of any race, color, religion, sex, or national origin.[48]

These two provisions lie at the core of what became known as disparate-treatment discrimination, although that precise terminology did not become common until the Supreme Court recognized the disparate impact theory of discrimination.[49]

The vast majority of disparate treatment case law focuses on intentional acts of discrimination. Courts generally follow the *McDonnell Douglas* burden-shifting framework to demonstrate circumstantial evidence of discriminatory intent—a near necessity in light of the fact that in most discrimination cases, there is no direct evidence of discriminatory intent.[50] Rare is the case where there is a "smoking gun" demonstrating that the employer used race or some other characteristic as the explicit justification for an adverse employment action. The *McDonnell Douglas* framework allows plaintiffs to create an inference of intent without such direct evidence of discriminatory animus.[51]

But on its face, § 703(a) does not actually require an intent to discriminate; it bars all discrimination made because of a protected characteristic.[52] The absence of an explicit intent requirement created—and continues to generate—ambiguity regarding Title VII's scope.[53] The most

47. EMPLOYMENT DISCRIMINATION, *supra* note 43 (quoting 42 U.S.C. § 2000e-2(a)(1)).

48. 42 U.S.C. § 2000e-2(j).

49. *See* Joseph A. Seiner, *Disentangling Disparate Impact and Disparate Treatment: Adapting the Canadian Approach*, 25 YALE L. & POL'Y REV. 95, 104–05 (2006) [hereinafter *Disentangling Disparate Impact*].

50. *See id.* at 81, 84–85 (citing McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973)).

51. *See McDonnell*, 411 U.S. at 802 (describing the prima facie case that a plaintiff is required to establish in order to show an inference of intent where direct evidence of discrimination is lacking).

52. 42 U.S.C. § 2000e-2(a) (2012).

53. *Disentangling Disparate Impact*, *supra* note 49, at 96.

important consequence of the broad language of § 703(a) was the creation of the disparate impact doctrine.

2. Disparate Impact

a. The Development of the Disparate Impact Doctrine

The standards governing disparate-impact discrimination are considerably more complex and ambiguous than those governing disparate treatment. The Supreme Court first established the disparate impact doctrine in *Griggs v. Duke Power Co.*, a class action by a group of black employees challenging their employer's requirement that new employees, in all but the lowest paying departments, have a high school diploma or pass a general intelligence test.[54] Both requirements operated to disproportionately exclude black workers—an outcome that likely was intended, given that many of the new requirements were imposed immediately after the passage of the Civil Rights Act of 1964.[55] The court of appeals concluded that the education and intelligence test requirements did not violate Title VII because they were facially neutral—that is, that they made no express distinction between employees on the basis of race—and because there was "no showing of a racial purpose or invidious intent."[56]

The Supreme Court reversed with an opinion that reshaped the legal landscape for employment discrimination law.[57] The Supreme Court began by rejecting the court of appeals' holding that the absence of intent to discriminate insulates a facially neutral employment condition under Title VII:

> The objective of Congress in the enactment of Title VII is plain from the language of the statute. It was to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees. Under the Act, practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they

54. Griggs v. Duke Power Co., 401 U.S. 424, 424–28 (1971).

55. *Id.* at 426–28.

56. *Id.* at 428–29. Apparently, the court of appeals did not perceive the close proximity between the enactment of Title VII and the imposition of the new requirements to be evidence of racially discriminatory intent.

57. *Id.* at 436.

> operate to "freeze" the status quo of prior discriminatory employment practices.[58]

In ruling that the intelligence test and high school diploma requirements were unlawful, the *Griggs* court emphasized the systemic disadvantages that African Americans face as a result of receiving "inferior education in segregated schools."[59] The Court explained the rationale for its new doctrine by analogy to one of Aesop's fables:

> Congress has now provided that tests or criteria for employment or promotion may not provide equality of opportunity merely in the sense of the fabled offer of milk to the stork and the fox. On the contrary, Congress has now required that the posture and condition of the job-seeker be taken into account. It has—to resort again to the fable—provided that the vessel in which the milk is proffered be one all seekers can use. The Act proscribes not only overt discrimination but also practices that are fair in form, but discriminatory in operation.[60]

The *Griggs* decision also announced what would become known as the business necessity defense: "The touchstone is business necessity. If an employment practice which operates to exclude Negroes cannot be shown to be related to job performance, the practice is prohibited."[61] The Court then concluded that "neither the high school completion requirement nor the general intelligence test is shown to bear a demonstrable relationship to successful performance of the jobs for which it was used."[62]

Four years after *Griggs*, the Supreme Court laid out what remains the basic framework for disparate impact litigation in *Albemarle Paper Co. v. Moody*.[63] In that case, the Court introduced a three-step rubric for disparate-impact cases that roughly corresponds to the *McDonnell Douglas* burden-shifting framework that it had adopted two years earlier for disparate treatment claims.[64] First, the complaining party must "[make] out a prima facie case of discrimination, i.e. . . . show[] that the tests in question select

58. *Id.* at 429–30, 436.

59. *Id.* at 430 (citing Gaston County v. United States, 395 U.S. 285, 287 (1969)).

60. *Id.* at 430–31.

61. *Id.*; *see* Susan S. Grover, *The Business Necessity Defense in Disparate Impact Discrimination Cases*, 30 GA. L. REV. 387, 387 (1996).

62. *Griggs*, 401 U.S. at 430–31.

63. Albemarle Paper Co. v. Moody, 422 U.S. 405, 425–36 (1975).

64. *Id.* at 425 (citing McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801–02, 804–05 (1973)).

applicants for hire promotion in a racial pattern significantly different from that of the pool of applicants."[65] If a prima facie case is established, the employer then can rebut by showing that the tests are "job related."[66] Finally, if the defendant establishes job relatedness, the plaintiff may still prevail by demonstrating "that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer's legitimate interest in 'efficient and trustworthy workmanship.'"[67] These three stages of a disparate impact case are explored further below.

b. Prima Facie Case

Albemarle Paper states that a plaintiff makes out a prima facie case of disparate impact by showing "that the tests in question select applicants for hire or promotion in a racial pattern significantly different from that of the pool of applicants."[68] The Court did not indicate, however, whether "significantly different" was intended to be a reference to significance in a formal statistical sense, or if it instead meant significant in some more colloquial sense.[69] This ambiguity has led to divergent interpretations of the nature and magnitude of the disparity necessary to establish a prima facie case of disparate-impact discrimination.

The Uniform Guidelines on Employee Selection Procedures (Guidelines) adopted the "four-fifths" or "80%" rule, under which:

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.[70]

65. *Albemarle Paper*, 422 U.S. at 425.

66. *Id.*

67. *Id.*; *Albemarle Paper* made no mention of the fact that the *McDonnell* framework was initially adopted in a case involving an allegation of intentional discrimination in what would today be termed a disparate treatment case. This further underscores the Court's initial conception of the disparate impact doctrine as merely a corollary to—and not different-in-kind from—disparate treatment discrimination. *See Griggs*, 401 U.S. at 431 (describing scope of what Congress proscribed in Title VII).

68. *Albemarle Paper*, 422 U.S. at 425.

69. *Id.*

70. 29 C.F.R. § 1607.4(D) (2019).

At first glance, this rule appears to focus exclusively on differences in selection rates and examines only the magnitude of the differences rather than their statistical significance. But the Guidelines hedge this rule to the point of meaninglessness, noting that smaller differences "may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms," and that greater differences may not constitute adverse impact "where the differences are based on small numbers and are not statistically significant . . . ."[71] The Guidelines offer no guidance on how enforcement agencies or the courts should determine whether an adverse impact exists where the four-fifths rule and a statistical significance test point in opposite directions.[72]

Courts have generally shunned the four-fifths rule as a test for prima facie disparate impact, preferring instead to rely on statistical significance tests. In *Hazelwood School District v. United States*, the Supreme Court indicated in a footnote that a difference of "more than two or three standard deviations" between the expected and actual number of protected class employees selected would make "the hypothesis that [employees] were hired without regard to race . . . suspect."[73] In the forty years since *Hazelwood*, courts have more often looked to the social science standard of statistical significance at the 5% level (1.96 standard deviations) than to *Hazelwood*'s less precise "two or three standard deviation" standard.[74] But no particular statistical method or threshold has been established as the *sine qua non* of disparate impact analysis.

Indeed, many courts have been openly hesitant to rely on statistical significance alone when attempting to assess adverse impact. Just as the Guidelines suggest that their four-fifths rule may be disregarded if observed disparities "are significant in both statistical and practical terms,"[75] courts have occasionally sought to inject a requirement of "practical" or "legal" significance—usually, akin to the four-fifths rule, by looking to the raw magnitude of the disparity—in addition to statistical significance.[76] But courts

71. *Id.*

72. *See* § 1607.4.

73. *Id.* at 308 n.14 (quoting Castaneda v. Partida, 430 U.S. 482, 496–97 n.17 (1977)).

74. *See, e.g.*, Smith v. Xerox Corp., 196 F.3d 358, 366 (2d Cir. 1999) ("Courts generally consider [significance at the 5%] level sufficient to warrant an inference of discrimination."); Palmer v. Shultz, 815 F.2d 84, 96 (D.C. Cir. 1987) ("[S]tatistical evidence must meet the 5% level . . . for it alone to establish a prima facie case under Title VII.").

75. 29 C.F.R. § 1607.4(D).

76. *See, e.g.*, Jones v. City of Boston, 752 F.3d 38, 49 (1st Cir. 2014) (characterizing the district court's use of the four-fifths rule as an examination of practical significance); Apsley v. Boeing Co., 691 F.3d 1184, 1199–1201 (10th Cir. 2012). The *Apsley* court affirmed the district court's rejection of plaintiff's prima facie disparate impact claim due to a lack of practical significance. *See Apsley*, 691 F.3d at 1199–1200. The court also noted that: [T]he

have reached no consensus on what practical significance entails or even whether it need be examined at all.

The Supreme Court arguably closed the door on practical significance requirements in *Ricci v. DeStefano*, where the Court stated in passing that a prima facie case of disparate impact requires showing a statistically significant disparity "and nothing more."[77] Although this statement is arguably dicta, it nevertheless suggests that statistical significance tests remains the primary means by which courts determine whether a prima facie case of disparate impact discrimination exists. This bodes ill for employers seeking to leverage Big Data because, as discussed in greater detail below, large data sets can render even the slightest differences in selection rates statistically significant, even if they have minimal real-world importance.[78]

c. Business Necessity Defense

Under the amendments to § 703 enacted in the Civil Rights Act of 1991, employers faced with a prima facie case of disparate impact discrimination must "demonstrate that the challenged practice is job related for the position in question and consistent with business necessity" to escape liability.[79] The concepts of "job relatedness" and "business necessity" first appeared in *Griggs*,[80] but in the five decades since, courts, agencies, and Congress alike have struggled with the meaning, relative importance, and interplay between the two concepts.[81]

Employees' own figures show that the Companies recommended and hired over 99% of the older employees they would have been expected to recommend and hire in the absence of any discrimination. While this disparity might still lead a social scientist to suspect that the divestiture process was not wholly free of age-based discrimination, it would not permit a jury to find that such discrimination was the Companies' standard operating procedure.

Id. at 1200–01 (first citing Apsley v. Boeing Co., 722 F. Supp. 2d 1218, 1239 (D. Kan. 2010); then citing Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977); and then citing International Brotherhood of Teamsters v. United States, 431 U.S. 324, 336 (1977)).

77. Ricci v. DeStefano, 557 U.S. 557, 587 (2009) (citing Connecticut v. Teal, 457 U.S. 440, 446 (1982)).

78. *See* discussion *infra* Section III.B.1.

79. 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

80. Griggs v. Duke Power Co., 401 U.S. 424, 424, 431 (1971).

81. The *Principles for the Validation and Use of Personnel Selection Procedures* provides an oblique definition of job relatedness by stating how job relatedness must be shown:

> The job relatedness of a selection procedure has been demonstrated when evidence supports the accuracy of inferences made from scores on, or evaluations derived from, those procedures regarding some important aspect of work behavior (e.g., quality or

In *Albemarle Paper*, the Court described the employer's burden as solely that of demonstrating job relatedness, without reference to business necessity.[82] But the Supreme Court appeared to return to the concept of business necessity two years later in *Dothard v. Rawlinson*, which stated that "a discriminatory employment practice must be shown to be necessary to safe and efficient job performance to survive a Title VII challenge" under a disparate impact theory.[83] But courts in the late 1970s and 1980s generally continued to follow the *Albemarle Paper* approach, seemingly disregarding *Griggs*'s description of business necessity as the touchstone of the analysis and instead focusing on job relatedness.[84] The Supreme Court then attempted to put the final nail in the business necessity coffin in *Wards Cove Packing Co. v. Atonio*.[85] There, the Court held that "the dispositive issue is whether a challenged practice serves, in a significant way, the legitimate employment goals of the employer" and that "there is no requirement that the challenged practice be 'essential' or 'indispensable' to the employer's business for it to pass muster."[86]

Just two years later, however, Congress overrode the Supreme Court's *Wards Cove* decision in the Civil Rights Act of 1991, which enshrined both job related and business necessity in the text of Title VII.[87] The latter term appeared not by itself, but instead as part of the phrase "job related for the position in question and *consistent with* business necessity."[88] The statutory text, like the case law that inspired it, does not clarify how job relatedness differs (if at all) from business necessity, nor does it indicate how demonstrating that something is "consistent with business necessity" differs

---

quantity of job performance; performance in training, advancement, tenure, turnover, or other organizationally pertinent behavior).

SOC'Y FOR INDUS. & ORGANIZATIONAL PSYCHOLOGY, PRINCIPLES FOR THE VALIDATION AND USE OF PERSONNEL SELECTION PROCEDURES 4 (5th ed. 2018) [hereinafter PRINCIPLES]. This definition may be adequate for social scientific purposes, but, as described further in the discussion of validation below, a test must measure a representative set of job behaviors and outcomes to satisfy the business necessity defense; it does not suffice that a test measures merely *some* important aspect of work behavior.

82. Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975).

83. Dothard v. Rawlinson, 433 U.S. 321, 331 n.14 (1977).

84. *See* LEX K. LARSON, 2 LARSON ON EMPLOYMENT DISCRIMINATION § 23.04[1] (2017), LEXIS.

85. Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989), *superseded by statute*, Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1074, *as recognized in* Raytheon Co. v. Hernandez, 540 U.S. 44 (2003).

86. *Wards Cove Packing Co.*, 490 U.S. at 659.

87. Civil Rights Act of 1991, Pub. L. No. 102-166, § 105, 105 Stat. 1074 (codified as amended at 42 U.S.C. §§ 2000e–2000e-17 (2012)).

88. *Id.* (emphasis added).

(if at all) from demonstrating that it is an actual business necessity.[89] Further, Congress expressly limited the legislative history that may be used to elucidate these distinctions.[90]

The original source for the pairing of *job related* with the phrase *consistent with business necessity* appears to be the Department of Labor regulations for federal contractors under the Rehabilitation Act, a predecessor to the ADA that applied to federal employees and contractors.[91] The relevant Rehabilitation Act regulations, which predated the Civil Rights Act of 1991 by more than a decade, stated that "to the extent qualification requirements tend to screen out qualified handicapped individuals," the requirement must be "job-related . . . and . . . consistent with business necessity and the safe performance of the job."[92]

The same wording that now appears in Title VII also appears, almost verbatim, in the Americans with Disabilities Act, which Congress enacted a year before the 1991 amendments to Title VII.[93] Specifically, the ADA prohibits employers from:

> [U]sing qualification standards, employment tests or other selection criteria that screen out or tend to screen out an individual with a disability or a class of individuals with disabilities unless the standard, test or other selection criteria, as used by the covered entity, is shown to be *job-related for the position in question and is consistent with business necessity*.[94]

89. *See id.* (emphasis added).

90. *Id.* § 105(b) ("No statements other than the interpretive memorandum appearing at Vol. 137 Congressional Record S 15276 (daily ed. Oct. 25, 1991) shall be considered legislative history of, or relied upon in any way as legislative history in construing or applying, any provision of this Act that relates to *Wards Cove*—Business necessity/cumulation/alternative business practice.").

91. 41 C.F.R. § 60-741.44(c)(1) (2019); Exec. Order No. 11,758, 39 Fed. Reg. 2,075 (Jan. 15, 1974); Americans with Disabilities Act of 1990, Pub. L. No. 101-336, 104 Stat. 327; Rehabilitation Act of 1973, Pub. L. No. 93-112, 87 Stat. 355.

92. § 60-741.44(c)(1); *see also* 29 C.F.R. § 32.14(a) (2019) (stating that qualifications for jobs with programs receiving federal financial assistance must be "related to the performance of the job and . . . consistent with business necessity and safe performance" if they tend to exclude individuals with disabilities).

93. *See* 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012); Americans with Disabilities Act of 1990 § 12101, 42 U.S.C. § 12112(b)(6) (2012).

94. § 12112(b)(6) (emphasis added).

According to case law,[95] ADA regulations,[96] and Equal Employment Opportunity Commission (EEOC) guidance,[97] this provision is closely linked to the ADA's central inquiry into whether an individual can perform the "essential functions" of a position.

Title VII makes no explicit reference to the essential functions of a job, and the ADA's linking of essential functions to the business necessity defense remains mostly foreign to Title VII jurisprudence. Nevertheless, the general concept—that job relatedness and business necessity require linking the selection criteria to specific, articulable, and important job functions—is one of the few common themes pervading the scattershot judicial and administrative interpretations of Title VII's business necessity defense.[98] The Guidelines emphasize careful job analysis, with a particular focus on identifying the "critical or important job duties, work behaviors or work outcomes."[99] And courts have generally refused to countenance challenged selection procedures where the employer fails to demonstrate a connection

95. *See, e.g.*, Bates v. United Parcel Serv., Inc., 511 F.3d 974, 996 (9th Cir. 2007) ("To show 'job-relatedness,' an employer must demonstrate that the qualification standard fairly and accurately measures the individual's actual ability to perform the essential functions of the job."); EEOC. v. Exxon Corp., 203 F.3d 871, 875 (5th Cir. 2000) ("[T]he business necessity defense . . . involves whether the individual can perform the 'essential functions' of the job . . . .").

96. *See, e.g.*, 29 C.F.R. § 1630.14(b)(3) (2019) ("[I]f certain criteria are used to screen out an employee or employees with disabilities as a result of such an examination or inquiry, the exclusionary criteria must be job-related and consistent with business necessity, and performance of the essential job functions cannot be accomplished with reasonable accommodation as required in this part."); 29 C.F.R. § Pt. 1630, app. (2019) ("As part of the showing that an exclusionary criteria is job-related and consistent with business necessity, the employer must also demonstrate that there is no reasonable accommodation that will enable the individual with a disability to perform the essential functions of the job.").

97. *See, e.g.*, EEOC, *The Americans With Disabilities Act: Applying Performance And Conduct Standards To Employees With Disabilities*, https://www.eeoc.gov/facts/performance-conduct.html [https://perma.cc/NBS4-YECZ] ("If an applicant or employee cannot meet a specific qualification standard because of a disability, the ADA requires that the employer demonstrate the importance of the standard by showing that it is 'job-related and consistent with business necessity.' This requirement ensures that the qualification standard is a legitimate measure of an individual's ability to perform an essential function of the specific position the individual holds or desires."); EEOC, A TECHNICAL ASSISTANCE MANUAL ON THE EMPLOYMENT PROVISIONS OF THE ADA § 9.4 (1992) ("If a worker has an on-the-job injury which appears to affect his/her ability to do essential job functions, a medical examination or inquiry is job-related and consistent with business necessity.").

98. *See Bates*, 511 F.3d at 996 (describing the relationship between essential functions of the job and job relatedness); *Exxon Corp.*, 203 F.3d at 875 (stating that essential job functions are part of the criteria for the "business necessity" defense); 29 C.F.R. § 1630.14(b)(3).

99. § 1607.14(B)(2).

between the selection procedure and specific, key aspects of job performance,[100] a process known as validation.

Establishing the *validity* of a selection procedure thus is the central task of employers faced with a prima facie case of disparate impact discrimination. The Supreme Court stated in *Griggs* that any test or screening mechanism for job applicants "must measure the person for the job and not the person in the abstract" to survive a Title VII challenge.[101] *Albemarle Paper* refined this rule by casting doubt on the usefulness of generic or subjective measures of performance to validate selection criteria.[102] The Court refused to accept an employer's attempt to validate its test by showing that the results correlated with supervisorial ratings, holding that those ratings were "extremely vague and fatally open to divergent interpretations":

> There is no way of knowing *precisely what criteria of job performance* the supervisors were considering, whether each of the supervisors was considering the same criteria or whether, indeed, any of the supervisors actually applied a focused and stable body of criteria of any kind. There is, in short, simply no way to determine whether the criteria actually considered were sufficiently related to the Company's legitimate interest in *job-specific ability* to justify a testing system with a racially discriminatory impact.[103]

Albemarle Paper narrowed the permissible focus of employment tests in other ways as well, effectively requiring employers to use tests that measure essential aspects of job performance.[104] The Court held that employers cannot use selection procedures that hold applicants to a higher standard than

---

100. *See, e.g.*, Ernst v. City of Chicago, 837 F.3d 788, 805 (7th Cir. 2016) ("Chicago failed to establish that its physical-skills entrance test reflects 'important elements of job performance.' And this lack of connection between real job skills and tested job skills is, in the end, fatal to Chicago's case.") (quoting EEOC v. Dial Corp., 469 F.3d 735, 743 (8th Cir. 2006)). Federal courts have at times endorsed, sometimes explicitly, an "essential job functions" interpretation of Title VII's business necessity defense. Most notably, in *Dothard v. Rawlinson*, 433 U.S. 321, 331–2 (1977), the Supreme Court rejected a prison system's argument that its minimum height and weight requirements for corrections counselors were job related. The Court acknowledged that height and weight requirements may "have a relationship to strength, a sufficient but unspecified amount of which is essential to effective job performance as a correctional counselor." *Id.* at 331. But the Supreme Court found this insufficient, noting that the employer "produced no evidence correlating the height and weight requirements with the requisite amount of strength thought *essential to good job performance*." *Id.* (emphasis added).

101. Griggs v. Duke Power Co., 401 U.S. 424, 436 (1971).

102. Albemarle Paper Co. v. Moody, 422 U.S. 405, 432–33 (1975).

103. *Id.* (emphasis added).

104. *Id.* at 433–34.

successful people currently in the position by imposing requirements those current workers could not satisfy.[105] It also endorsed the contemporary EEOC guidelines' rule that employers could not test for criteria relevant only to higher level jobs unless the "job progression structures and seniority provisions are so established that new employees will probably, within a reasonable period of time and in a great majority of cases, progress to a higher level."[106] The Court reasoned that the flaw in such an approach is that:

> The fact that the best of those employees working near the top of a line of progression score well on a test does not necessarily mean that that test, or some particular cutoff score on the test, is a permissible measure of the *minimal qualifications of new workers* entering lower level jobs.[107]

While never using the term essential functions, the Court implied that the criteria used to assess job candidates must be based on key aspects of job performance.[108]

Determining what those key aspects of job performance are—and demonstrating that the selection process effectively measures them—is the crux of test validation. The current version of the *Standards for Educational and Psychological Testing* (*Standards*) defines validity as "the degree to which *evidence* and *theory* support the interpretations of test scores for proposed uses of tests."[109] In the context of employee selection procedures,

105. *Id.* at 429, 435–36 ("The record shows that a number of white incumbents in high-ranking job groups could not pass the tests.").

106. *Id.* at 434 (quoting 29 C.F.R. § 1607.4 (c)(1) (2019)). This rule was later incorporated almost verbatim into the Uniform Guidelines. 29 C.F.R. § 1607.5(I) (2019).

107. *Albemarle Paper*, 422 U.S. at 434 (emphasis added). The Third Circuit has made a "minimal qualifications" requirement the crux of its construction of the business necessity defense, holding that "employers may not use criteria which have a discriminatory effect unless those criteria define the minimum qualifications necessary to perform the job." NAACP v. N. Hudson Reg'l Fire & Rescue, 665 F.3d 464, 477 (3d Cir. 2011). The Uniform Guidelines did not go quite so far, allowing employers to rely upon tests that measure important components of job performance. 29 C.F.R. § 1607.5(B) (2019) ("Evidence of the validity of a test or other selection procedure by a criterion-related validity study should consist of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance . . . . Evidence of the validity of a test or other selection procedure by a content validity study should consist of data showing that the content of the selection procedure is representative of important aspects of performance on the job . . . .").

108. *Albemarle Paper*, 422 U.S. at 433.

109. AM. EDUC. RESEARCH ASS'N ET AL., STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING 11 (4th ed. 2014) (emphasis added). "Theory" is a key element of this formulation because data-driven algorithms are largely, if not entirely, empirical. *Id.* As a

the *evidence* is the information indicating how well the selection procedure actually measures the fitness of candidates for that particular job. The *theory* is the chain of logic that links the selection procedure to the job requirements. For example, there is a logical relationship between the requirement that a programmer be conversant with a particular computer language and the ability of that candidate to efficiently write code in that language. But no logic or theory suggests that the car one drives ought to predict a candidate's ability to succeed as a coder. Consequently, identifying the critical and important aspects of job performance—as well as metrics that have an evidentiary and theoretical connection to those aspects of job performance—lies at the heart of the validation process under the Guidelines.[110]

The Guidelines discuss three different types of validity, presenting each as an independent path through which an employer can establish the job relatedness of a selection procedure: criterion-related validity, which is based on correlations between performance on the test and performance on the job and is by far the validation method most frequently used for employee selection procedures; content validity, which requires designing a test that adequately simulates job performance; and construct validity, which is based on measuring more abstract characteristics that are important for successful job performance.[111] Of these, only criterion-related validation represents a plausible path to establishing the validity of an algorithmic selection procedure. Content validity is a poor match for most algorithmic selection tools, which do not attempt to directly test an applicant's job-related knowledge or ability to perform specific tasks central to the job. The Guidelines assume that evidence for construct validity will come from criterion studies;[112] because the Guidelines also recognize criterion-related studies alone as a basis for establishing the validity of a test, it rarely is efficient or even useful for an employer to pursue construct validation (at least as presented in the Guidelines)[113] rather than criterion validation.

But even criterion-related validation is an arduous process under the Guidelines.[114] Moreover, the Guidelines were promulgated in the 1970s[115] and reflect half-century-old conceptions both of the nature and format of

result, this raises the question of whether the APA would endorse predictive methods that lack a theoretical foundation.

110. *See* 29 C.F.R. §§ 1607.14 B(2), C(4) (2019).

111. § 1607.5(A)–(B).

112. § 1607.14(B)(2)–(3).

113. *See* discussion *infra* Section III.A.2. As discussed in greater detail in Section III.A.2, scientific concept of construct validity has evolved considerably in the decades since the Guidelines were issued.

114. *See* discussion *infra* Section III.A.1.

115. McDaniel et al., *supra* note 10, at 507.

employment tests and of what makes a test valid. As discussed further below,[116] this makes it difficult to predict how courts and agencies will assess the validity of algorithmic selection procedures.

d. Least Discriminatory Alternative

If an employer meets its burden in establishing the job relatedness of the selection procedure, the final stage of disparate impact analysis requires a plaintiff to demonstrate that a less discriminatory alternative was available that would meet the employer's business needs.[117] This test traces its roots to *Albemarle Paper*, which stated that a plaintiff could prevail on a disparate impact claim by demonstrating "that other tests or selection devices, without a similar undesirable racial effect, would also serve the employer's legitimate interest in 'efficient and trustworthy workmanship.'"[118]

A key question that remains largely unresolved is how effective the plaintiff's proposed alternative must be to defeat an employer's showing of business necessity. *Albemarle Paper*'s standard—that the procedure need only "serve the employer's legitimate interest in 'efficient and trustworthy workmanship'"—appeared to set the bar rather low, suggesting that the proposed alternative need not be *exactly* as effective as the challenged procedure, so long as it is adequate to meet the employer's needs.[119] In *Wards Cove*, the Supreme Court attempted to reject this low bar, holding that an alternative practice "must be equally effective as [the employer's] chosen hiring procedures in achieving [the employer's] legitimate employment goals."[120] But as with *Wards Cove*'s alteration of the business necessity defense, Congress overrode the Court through the 1991 amendments to Title VII.[121] In the Civil Rights Act of 1991, Congress explicitly restored the law governing alternative employment practices to "the law as it existed on June 4, 1984," the day before *Wards Cove* was decided.[122]

Unfortunately, the exact nature of the "less discriminatory alternative" standard was far less than clear even before *Wards Cove*.[123] The only type of modification to a selection procedure that seems to have gained wide recognition as an adequate alternative is the practice of "banding" test scores,

116. *See* discussion *infra* Section III.A.1.

117. Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975) (quoting McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801 (1973)).

118. *See id.* (quoting *McDonnell*, 411 U.S. 792, 801 (1973)).

119. *Id.*

120. Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 661 (1989).

121. *See id.*; Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1074.

122. *Id.*; 42 U.S.C. § 2000e-2(k)(1)(C) (2012).

123. *See Disentangling Disparate Impact*, *supra* note 49, at 103–04.

where candidates are grouped together in bands based on differences between scores that are considered insignificant.[124] Because of the paucity of cases clarifying the standards by which alternative selection procedures should be judged, courts have generally been reluctant to decide cases on the basis of a plaintiff's showing of a less discriminatory alternative.[125]

3. Ricci v. DeStefano *and the Interplay Between Disparate Treatment and Disparate Impact*

While *Griggs* cast the disparate impact theory as simply a logical corollary of the disparate treatment that Title VII clearly prohibited, these legal theories in fact spring from quite separate views on the thrust and purpose of antidiscrimination laws. Disparate treatment, as presently interpreted, reflects an anticlassification view of discrimination, which holds that the purpose of antidiscrimination laws is to prohibit classifying or differentiating between individuals on the basis of a protected characteristic.[126] Disparate impact, by contrast, reflects an antisubordination perspective on discrimination, under which the purpose of such laws is to "prohibit practices that enforce the social status of oppressed groups and allow practices that challenge oppression."[127] The antisubordination roots of disparate impact theory can be seen in *Griggs*, where the Supreme Court emphasized the long-running and systemic disadvantages that blacks had endured, and rejected the notion that an employment test complies with Title VII so long as it is "fair in form."[128]

The conceptual tension between disparate treatment and disparate impact causes practical problems for employers who observe that their policies are having disparate impacts (or anticipate that they will have a disparate impact in the future) and perceive that the most logical way to stop such adverse impacts from arising is to take direct steps to correct for the disparate impact. But the very act of correcting disparate impacts may itself be a form of

124. *See, e.g.*, Officers for Justice v. Civil Serv. Comm'n, 979 F.2d 721, 723–24, 728 (9th Cir. 1992) ("Today we hold that the banding process is valid as a matter of constitutional and federal law."); Chi. Firefighters Local 2 v. City of Chicago, 249 F.3d 649, 656 (7th Cir. 2001) ("[Banding is] a universal and normally an unquestioned method of simplifying scoring by eliminating meaningless gradations.").

125. LARSON, *supra* note 84, at § 24.02.

126. *See* Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 10 (2003).

127. Bradley A. Areheart, *The Anticlassification Turn in Employment Discrimination Law*, 63 ALA. L. REV. 955, 961 (2012).

128. Griggs v. Duke Power Co., 401 U.S. 424, 430–31 (1971).

disparate treatment.[129] That dilemma made its way to the Supreme Court in the 2009 case *Ricci v. DeStefano*.[130]

The plaintiffs in *Ricci* were white and Hispanic firefighters who had taken and passed an examination administered by the City of New Haven that determined the firefighters' eligibility for promotion to lieutenant or captain.[131] The City worked with an outside consulting firm to develop the test over a period of several years.[132] But the City's first real-life administration of the test showed that using the results of the exam would have an adverse impact on black and Hispanic firefighters; thirty-four of the seventy-seven firefighters who took the examination were black or Hispanic, but all ten of the candidates who scored high enough to be considered for promotion were white.[133] Based on these disproportionate outcomes, the City believed that using the results of the test would have an unlawful disparate impact and subject them to liability under Title VII.[134] Consequently, the City chose not to certify the examination results.[135]

The firefighters who passed the test challenged the City's decision as expressly race based, and sought review by the Supreme Court.[136] The Court ruled for the firefighters and held that the City's decision, because it was driven by concern over the adverse impact on minority firefighters, was a decision made because of race in violation of Title VII's disparate treatment prohibition:

> All the evidence demonstrates that the City chose not to certify the examination results because of the statistical disparity based on race—*i.e.*, how minority candidates had performed when compared to white candidates. As the District Court put it, the City rejected the test results because "too many whites and not enough minorities would be promoted were the lists to be certified." Without some other justification, this express, race-based decision-making violates Title VII's command that employers cannot take adverse employment actions because of an individual's race.[137]

129. *See* Ricci v. DeStefano, 557 U.S. 557, 593 (2009).
130. *Id.*
131. *Id.* at 562–63.
132. *Id.* at 564.
133. *Id.* at 566.
134. *Id.* at 563, 566.
135. *Id.* at 574.
136. *Id.* at 563, 574–75.
137. *Id.* at 579 (citation omitted).

Notably, the Supreme Court's reasoning did not focus on racial animus or an intent to discriminate in the usual sense.[138] On the contrary, the Court acknowledged that the employer's objective had been avoiding disparate-impact liability—in other words, to avoid committing unlawful discrimination.[139] But that objective did not insulate the employer from liability because it ignored "the City's conduct in the name of reaching that objective."[140] The Court reasoned:

> Whatever the City's ultimate aim—however well-intentioned or benevolent it might have seemed—the City made its employment decision because of race. The City rejected the test results solely because the higher scoring candidates were white. The question is not whether that conduct was discriminatory but whether the City had a lawful justification for its race-based action.[141]

The Court rejected the City's argument that its violation of the disparate treatment prohibition should be excused because the City only did so to avoid the prospect of disparate impact liability.[142] But in doing so, the Court explicitly left open the possibility that an employer, although not the employer in *Ricci* itself,[143] would be able to use the prospect of disparate impact liability as a defense to a disparate treatment claim.[144] The Court rejected the plaintiff firefighters' blanket argument that "avoiding unintentional discrimination

138. *Id.* at 592.

139. *Id.* at 579.

140. *Id.*

141. *Id.* at 579–80.

142. *Id.* at 563.

143. As the dissent noted, despite its adoption of the "strong basis in evidence" standard discussed below, the *Ricci* majority did not remand the matter for further proceedings so that evidence could be presented on the strength of a potential business necessity defense and the availability (or not) of less discriminatory alternative selection procedures. *Id.* at 563, 631 (Ginsburg, J., dissenting). Instead, it ruled as a matter of law that plaintiffs were entitled to summary judgment. *Id.* at 592. The majority's reasoning implies that it concluded either that the test was actually valid (which appeared to be a disputed factual question) or presumptively valid (which runs contrary to the Guidelines and the statute, which places the burden on the employer to establish validity). Given the absence in the record of criterion-related validity evidence, the basis of the Court's validity finding is unclear. The Court seemed to imply that the rigorous job analysis that the City had performed, coupled with its efforts to craft a test based on that evidence, established the test's validity. That line of reasoning most closely tracks a content validity argument, but a paper-and-pencil multiple-choice test would not be a direct test of job performance for a firefighter, as the dissenters in *Ricci* pointed out. *See id.* at 634 (Ginsburg, J., dissenting).

144. *See id.* at 593. The narrowness of the holding was to the apparent chagrin of Justice Scalia, who wrote a brief concurrence chiding the Court for declining to confront the constitutionality of the disparate impact doctrine directly. *Id.* at 594 (Scalia, J., concurring).

cannot justify intentional discrimination."[145] Going a step further, it also declined to adopt a standard under which "an employer in fact must be in violation of the disparate-impact provision before it can use compliance as a defense in a disparate-treatment suit."[146]

Instead, borrowing from the Court's constitutional Equal Protection Clause jurisprudence, the Court held that an employer could escape disparate treatment liability if it "can demonstrate a *strong basis in evidence* that, had it not taken the action, it would have been liable under the disparate impact statute."[147] The City failed in this regard because it did not adequately consider evidence of the validity and job relatedness of the test—and job relatedness is a complete defense to a disparate impact claim.[148] After concluding—dubiously, given the case's posture as an appeal from a summary judgment motion—that the City had failed to make such a "strong basis in evidence" showing, it ruled that the plaintiffs were entitled to summary judgment and, in effect, ordered that the City certify the examination results.[149]

At first blush, *Ricci* seems a very ominous portent for employers considering whether and how to implement novel selection procedures—and it certainly is for employers who discover an adverse impact only after a selection procedure has been designed and administered. Such employers face a catch-22, where attempting to mitigate the disparate impact could subject them to disparate treatment liability, while inaction would leave them vulnerable to a disparate impact claim.[150] But the Court appeared to leave open an avenue through which employers could mitigate anticipated disparate impacts without necessarily violating Title VII.[151]

Specifically, the Court held that "Title VII does *not* prohibit an employer from considering, *before* administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race."[152] Explaining the dissonance between that principle and the Court's disposition of the firefighters' examination results, the Court stated:

145. *Id.* at 580.

146. *Id.* at 580–81.

147. *Id.* at 563 (emphasis added).

148. *See id.* at 578.

149. *See id.* at 593 (holding that petitioners are entitled to summary judgment and remanding for further proceedings consistent with the opinion).

150. *See id.* at 629 (Alito, J., concurring).

151. *See id.* at 585.

152. *Id.* at 585 (emphasis added).

> [W]e [do not] question an employer's affirmative efforts to ensure that all groups have a fair opportunity to apply for promotions and to participate in the process by which promotions will be made. But once that process has been established and employers have made clear their selection criteria, they may not then invalidate the test results, thus upsetting an employee's legitimate expectation not to be judged on the basis of race.[153]

The City's actions, according to the Court, fell into the latter category.[154] The Court emphasized the "high, and justified, expectations of the candidates who had participated in the testing process on the terms the City had established for the promotional process," many of whom "had studied for months, at considerable personal and financial expense."[155] The unfairness of the City's decision lay not in its desire to avoid using a test that would have a disparate impact, but in the fact that the City only decided to discard the results after the examination design process was complete and the promotion candidates developed something akin to a reliance interest in having the examination used as a basis for promotion decisions.[156]

The Court's reasoning seems consistent with the text of the most on-point provision in Title VII, § 703(*l*).[157] That provision makes it unlawful for employers to "adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment related tests" on the basis of protected class status.[158] Technically, designing a selection procedure to avoid disparate impacts would not be adjusting test scores or using different cutoffs because the scoring rubric for a selection procedure is not yet finalized during the test design stage.[159]

That said, there is no case law squarely addressing the issue of how much license employers have to protect against disparate impacts by designing a selection procedure in a manner that explicitly takes protected class status into account. Is it permissible for employers to choose a suboptimal selection device, as measured by its accuracy, because it results in a more diverse workforce? It is safe to assume that there are limits—not least from § 703(a)'s general prohibition against making employment decisions because of protected class status—on the degree to which employers can be race or

153. *Id.* at 585.
154. *Id.* at 593.
155. *Id.*
156. *See id.*
157. *See* 42 U.S.C. § 2000e-2(*l*) (2012).
158. *Id.*
159. *See id.*

gender conscious when designing a selection procedure.[160] Using quotas or granting bonus points on the basis of protected class status, for instance, surely would not survive a disparate treatment challenge, even if an employer adds those features as part of initial test design.[161] But it is not clear where courts will draw lines between permissible and impermissible methods of designing around disparate impacts.

This ambiguity is a source of concern for employees considering algorithmic selection procedures.[162] Algorithms offer the potential for employers to design a selection procedure that reduce or eliminate disparate impacts using methods that are far more sophisticated and subtle than the blunt instruments available for traditional tests.[163] The degree to which those methods are deemed consistent with Title VII will likely determine how quickly employers adopt algorithmic selection procedures in the coming years.

III. THE UNIQUE CHALLENGES FOR ALGORITHMIC SELECTION TOOLS

Designing algorithmic selection tools that leverage the ability to generate unique data-driven insights while maintaining legal compliance will prove challenging under current law. A comprehensive treatment of all the practical and legal ambiguities surrounding algorithmic selection tools would be prohibitively lengthy, but sections A through D of Part III identify four overarching categories that encompass the most vexing legal compliance issues for algorithmic tools: challenges relating to the validation process; those stemming from algorithmic tools' reliance on correlation and use of Big Data; those relating to the opacity of models generated by deep neural networks; and those stemming from the bare fact that the deployment of algorithmic tools will provide plaintiffs' lawyers with a clear target for bringing discrimination claims.

A final issue that starkly illustrates the "square peg in a round hole" dynamic of algorithmic selection tools and current employment discrimination law is whether Title VII's disparate treatment doctrine can even be applied to machines that do not possess conscious intentions—or, indeed, consciousness at all. As explained in the final section of this part, despite the intent-focused tilt of case law on disparate treatment, the broad language of the statutory text and the equally broad early Supreme Court

160. *See generally* 42 U.S.C. § 2000e-2(a) (2012) (discussing unlawful employment practices for employers based on an "individual's race, color, religion, sex, or national origin").

161. *See* NAT'L ARCHIVES, *EEO Terminology*, https://www.archives.gov/eeo/terminology.html [https://perma.cc/G4DQ-84JA].

162. *See* ZHANG ET AL., *supra* note 2, at 8.

163. *See id.* at 10.

decisions interpreting it mean that employers are unlikely to escape disparate treatment liability if they deploy algorithms that make facially discriminatory classifications.

A. Validation

1. The Difficult Path to Criterion-Related Validity Under the Guidelines

The Guidelines establish rigorous standards for criterion-related validation studies.[164] These standards correctly ensure that a selection procedure has a demonstrable relationship to each job for which it is used, but the expense and complexity of establishing and maintaining the criterion-related validity of an algorithmic tool will blunt the efficiency gains that algorithmic tools promise.[165]

a. Job Analysis

A criterion-related validation study begins with a careful job analysis conducted by industrial psychologists or other trained professionals to identify the critical and important elements of job performance.[166] Courts emphasize the thoroughness and attention to detail that a job analysis entails and often reject validation studies that are not supported by adequate job analyses.[167] One court described a job analysis: "A thorough survey of the relative importance of the various skills involved in the job in question and the degree of competency required in regard to each skill. It is conducted by interviewing workers, supervisors and administrators; consulting training manuals; and closely observing the actual performance of the job."[168]

---

164. *See generally* 29 C.F.R. § 1607.14 (2019) ("The following minimum standards, as applicable, should be met in conducting a validity study.").

165. *See* 29 C.F.R. § 1607.14(B)(3) (2019).

166. *See* 29 C.F.R. § 1607.5(B)(2) (2019); § 1607.14(B)(2).

167. *See, e.g.*, Albemarle Paper Co. v. Moody, 422 U.S. 405, 432 (1975) ("The study in this case involved no analysis of the attributes of, or the particular skills needed in, the studied job groups. There is accordingly no basis for concluding that 'no significant differences' exist among the lines of progression, or among distinct job groupings within the studied lines of progression."); Rogers v. Int'l Paper Co., 510 F.2d 1340, 1351 (1975) ("The . . . absence of proper and careful job analyses . . . is fatal to the validation study.").

168. Guardians Ass'n of N.Y.C. Police Dep't v. Civil Serv. Comm'n of N.Y., 633 F.2d 232, 242 (2d Cir. 1980), *aff'd sub nom*, Guardians Ass'n of N.Y.C. Police Dept, v. Civil Serv. Comm'n of New York, 463 U.S. 582 (1983) (quoting Vulcan Soc'y of N.Y.C. Fire Dep't v. Civil Serv. Comm'n of New York, Inc., 360 F. Supp. 1265, 1274 (S.D.N.Y. 1972)).

From these observations and information, the experts conducting the study then "break[] down an observed task into a set of component skills, abilities and knowledge," and indicate what level of competence or proficiency is required for each component.[169] According to the American Psychological Association's *Principles for the Validation and Use of Personnel Selection Procedures*, which apply the more generally applicable *Standards* to the employee selection setting, a proper job analysis "may include different dimensions or characteristics of work, including work complexity, environment, context, tasks, behaviors and activities performed, and worker requirements (e.g., KSAOs [Knowledge, Skills, Abilities, and Other Characteristics] or competencies)."[170]

b. Selecting Criteria

From the critical and important job duties, work behaviors, and work outcomes identified during the job analysis, an employer must then select or develop measurable criteria that serve as metrics of how well an individual can perform the key functions of a job.[171] Employees' real-world performance with respect to those job related criteria then serve as the benchmarks for validation[172]—and, in the case of algorithmic tools, as target variables for building a model.

Needless to say, criterion selection is crucial to a proper criterion-related validity study. "Criteria should be chosen on the basis of work relevance, freedom from contamination, and reliability rather than availability or convenience,"[173] and "should represent important or critical work behavior(s) or work outcomes," as identified in the job analysis.[174] Where courts have refused to recognize proffered criterion validity studies, it has not usually been because the employer failed to show the proper correlation between the selection procedure and the criteria, but because the employer failed to select proper criteria in the first place.[175]

169. *See* Jones v. N.Y.C. Human Res. Admin. 391 F. Supp 1064, 1080 (S.D.N.Y. 1975).

170. PRINCIPLES, *supra* note 81, at 1, 7.

171. *See id.* at 7.

172. *See id.* at 10.

173. *Id.* at 11.

174. 29 C.F.R. § 1607.14(B)(3) (2019).

175. Albemarle Paper Co. v. Moody, 422 U.S. 405, 432–34 (1975) (stating that Albemarle Paper failed to show that its testing was job related when the criteria were subjective); *see also, e.g.*, Green v. U.S. Steel Corp., 570 F. Supp. 254, 273–77 (E.D. Pa. 1983) (stating that the hiring standard U.S. Steel adopted was a series of subjective criteria and found that because the employer failed to rebut plaintiffs' prima facie case of disparate impact, the plaintiffs were entitled to judgment in their favor).

As a threshold matter, the criteria must be direct measures of job performance, and not separate on-the-job tests or assessments that have not themselves been validated.[176] Courts generally expect criteria to be specific and reasonably objective markers of job performance and frown on criteria that are vague, generic, or subjective. In *Albemarle Paper*, the Supreme Court found a purported criterion validity study inadequate in large part because the criteria consisted of subjective supervisory employee rankings that were made according to "a 'standard' that was extremely vague and fatally open to divergent interpretations."[177] Consequently, supervisory ratings and assessments—which are often the only available measures of an employee's on-the-job performance—may not be adequate to support criterion-related validity.[178]

But it is difficult, and often impossible, to capture all essential and important job behaviors and job outcomes using readily available data. More general signals of employee performance such as statistics on hiring, retention, and tenure are generally available to employers. Formal performance reviews in some form may also be available, but if these include narrative sections or are not subject to a uniform rubric that ensures the reviews have consistent meaning, the reliability of the reviews (and the ability of an algorithm to make sense of them) as target variables might be limited.

Some jobs may have reasonably reliable performance metrics that seem to capture the essence of the job. But a closer examination often reveals that available metrics do not adequately measure job performance.[179] For example, a district attorney's office may track the number of cases that its prosecutors try and the percentage of cases they win. These statistics, which can be tracked reliably at little or no cost, may make attractive target variables. But a prosecutor's win-loss record may be a poor indicator of the quality of their lawyering. The best prosecutors might be the ones who take on the most difficult and time-intensive cases, and thus try fewer cases and have a lower

176. Ernst v. City of Chicago, 837 F.3d 788, 802 (7th Cir. 2016) ("Chicago created a skills test and a work-sample test, found a strong correlation between the skills test and the work-sample test, and thus concluded that the skills test is a good measure of job-related skills. As the plaintiffs argue, this is a statistical form of self-affirmation. There is no evidence that the work-sample test, which Chicago used to validate the skills test, is a proper validation of job skills.").

177. *Albemarle Paper*, 422 U.S. at 432–33; *see also, e.g.*, *Green*, 570 F. Supp. at 275–76 (discussing subjective hiring criteria). In *Green*, the court struck down a company's "'best qualified' hiring standard," which "consist[ed] of about twenty subjective criteria applied as an 'amalgam,'" and where evidence showed that "each individual decision-maker essentially simply consulted his or her 'gut level' reaction to an individual applicant." *Green*, 570 F. Supp. at 275–76.

178. *See Albemarle Paper*, 422 U.S. at 432–33.

179. Dick Grote, *The Myth of Performance Metrics*, HARV. BUS. REV. (Sept. 12, 2011), https://hbr.org/2011/09/the-myth-of-performance-metric [https://perma.cc/RZ7R-HW7W].

rate of positive outcomes than less skilled lawyers who shy away from such cases. But if the district attorney lacks the resources to closely observe the work of most prosecutors, the flawed trial statistics may be the only metrics available.[180]

Similarly, employers are often tempted to search for readily observable characteristics that can serve as proxies for attributes essential to the job in question. But this too carries risk. In *Dothard v. Rawlinson*, an employer attempted to justify its minimum requirements for height and weight—metrics that were readily available—on the claimed basis that those requirements were meant to ensure that corrections counselors had the requisite physical strength, which was the job-relevant attribute of interest.[181] The Court rejected this argument, reasoning that "[i]f the job-related quality that the appellants identify is bona fide, their purpose could be achieved by adopting and validating a test for applicants that measures strength directly."[182]

Having a representative set of participants is another key requirement for criterion-related validation.[183] The subjects must broadly reflect of the characteristics of the pool of actual applicants.[184] Thus, the sample must consist of entry-level employees if it is for an entry-level job; using employees from higher in the line of progression is not sufficient.[185] Representativeness across protected classifications is also required; the Guidelines state that the sample "should insofar as feasible include the races, sexes, and ethnic groups normally available in the relevant job market."[186] Ultimately, an employer establishes criterion validity under the Guidelines by demonstrating that performance on the selection procedure correlates with a representative set of performance measures tied to the job criteria identified during the job analysis.[187]

As the above discussion suggests, a proper criterion-related validity study is a major undertaking even for large and sophisticated employers. This may explain, in part, why most employers have shied away from using

180. The tendency to turn to easily available metrics as a substitute for deeper analysis is hardly unique to the hiring process. Before the recent explosion in sports analytics, Michael Lewis observed that "[f]or most of its history basketball has measured not so much what is important as what is easy to measure—points, rebounds, assists, steals, blocked shots—and these measurements have warped perceptions of the game." Michael Lewis, *The No-Stats All-Star*, N.Y. TIMES MAG., Feb. 13, 2009, https://www.nytimes.com/2009/02/15/magazine/15Battier-t.html [https://perma.cc/9LTB-ZRS7].

181. Dothard v. Rawlinson, 433 U.S. 321, 331 (1977).

182. *Id.* at 332.

183. 29 C.F.R. § 1607.14(B)(1) (2019).

184. *Id.*

185. Albemarle Paper Co. v. Moody, 422 U.S. 405, 434 (1975).

186. § 1607.14(B)(4).

187. § 1607.14(B)(2)–(3).

employment tests altogether;[188] relying on human judgment, however flawed, generates neither the cost nor the discoverable paper trail that validation entails.

Because many employers wishing to deploy an algorithmic selection procedure will not have ready access to a properly developed set of criteria that can serve as the basis for a criterion-related validity study, the process of developing and validating an algorithmic tool may take several years. That timetable that may prove problematic given the vintage of the Guidelines and the likelihood that courts and agencies will introduce new standards for validation in the coming years.

2. *The Guidelines: Behind the Times*

The long and difficult process of criterion-related validation under the Guidelines will be challenging enough for employers testing new hiring tools. But the Guidelines' forty-year-old standards are overdue for revamping or replacement to bring them in line with the modern social science of test validity, which has evolved considerably in the decades since the Guidelines first appeared. This adds an additional layer of legal uncertainty.

The EEOC and four other federal agencies and departments[189] jointly adopted the Guidelines in 1978. Recognizing that theories of test validity were still evolving, the Guidelines state that "[n]ew strategies for showing the validity of selection procedures will be evaluated as they become accepted by the psychological profession."[190] But the Guidelines' validation standards have, in fact, remained unchanged in the four decades since their promulgation. In the interim, the American Psychological Association (APA) has issued revised versions of the *Standards* three times (in 1985, 1999, and 2014). Starting with the 1985 edition, the *Standards* moved away from the Guidelines' trichotomous separation of test validity into content, criterion, and construct validity.[191] Consequently, even before the advent of Big Data and the prospect of completely new types of employee selection procedures, many of the Guidelines' provisions and much of their terminology seemed dated.

Comparing the descriptions of construct validity in the Guidelines with those in modern scientific literature provides a stark example of how much the social science of test validation has evolved since the Guidelines were

188. *See infra* note 238 and accompanying text.

189. These federal agencies and departments include the Office of Personnel Management, Department of Justice, Treasury Department, and the Department of Labor's Office of Federal Contract Compliance Programs. *See* 29 C.F.R. § 1607.2(A) (2019).

190. § 1607.5(A).

191. *See* Samuel Messick, *Validity*, *in* EDUCATIONAL MEASUREMENT 13, 18–20 (Robert L. Linn ed., 3d ed. 1989).

issued. The Guidelines refer to construct validity as "a relatively new and developing procedure in the employment field," for which there was, as of 1978, "a lack of substantial literature extending the concept to employment practices."[192] But the literature surrounding construct validity developed rapidly in the 1980s and 90s; today, far from an undeveloped and novel theory, construct validity is generally recognized as *the* overarching validity concept.[193] Where the Guidelines present construct and content validity as separate types of validity, modern social science treats test content and criterion relatedness as categories of evidence for demonstrating the broader concept of test validity.[194]

Modern test literature treats test bias and fairness as potential threats to validity, and the vocabulary surrounding what constitutes test bias relies—sometimes explicitly—on the concept of the constructs that represent whatever the selection procedure is ultimately attempting to measure.[195] One specific threat to validity extensively studied by modern social scientists—construct-irrelevant variance—will take on particular importance in the age of Big Data and with the rise of algorithmic selection procedures, as discussed further below.[196] But the Guidelines and the existing case law on validation are bereft of meaningful discussion of these threats to validity, leaving employers to guess if, when, and how courts and agencies will take them in into account.

Many courts continue to cite the Guidelines when discussing proper validation methods, and the EEOC's Fact Sheet on Employment Tests and Selection Procedures still references the Guidelines as the primary source of regulatory guidance on validation of selection procedures.[197] Employers seeking to implement algorithmic selection procedures thus have little choice but to pursue validation that complies with the Guidelines. But the stringent requirements for criterion validation under the Guidelines can take many years to complete. The law may well change in the interim, which makes reliance on the Guidelines' validity standards an inherently unstable proposition as long as they lag decades behind the prevailing social science.

---

192. § 1607.14(D)(1).

193. *See, e.g.*, WILLIAM M. TROCHIM ET AL., RESEARCH METHODS 128–30 (2d ed. 2015); AM. EDUC. RESEARCH ASS'N ET AL., *supra* note 109, at 11 ("The term *construct* is used in the *Standards* to refer to the concept or characteristic that a test is designed to measure."). *But see* Jerry A. Colliver et al., *From Test Validity to Construct Validity... and Back?*, 46 MED. EDUC. 366, 367–70 (2012) (criticizing the increasingly broad use of construct validity despite its rising popularity in social science).

194. AM. EDUC. RESEARCH ASS'N ET AL., *supra* note 109, at 14–19.

195. *Id.* at 5 ("Fairness and accessibility, the unobstructed opportunity for all examinees to demonstrate their standing on the construct(s) being measured, are relevant for valid score interpretations for all individuals and subgroups in the intended population of test takers.").

196. *See* discussion *infra* Section III.B.2.

197. § 1607.14(B)(3).

B. The Pitfalls of Correlations and Big Data

The sheer size of data sets in the era of Big Data deepens the challenges that employers, agencies, and courts will face when attempting to analyze whether a particular algorithmic selection tool is legally compliant. Some of these challenges relate to algorithmic and data-driven selection tools' reliance on correlation rather than causation. In some ways, using correlative techniques across a huge number of attributes allows for a richer and more holistic analysis of candidates. But correlative techniques fit awkwardly (if at all) with existing legal frameworks, many of which—including antidiscrimination laws—rest on cause-and-effect relationships.[198] Reliance on correlation alone is also discouraged in modern test validity theory. This could complicate efforts to validate selection procedures that have an adverse impact.[199]

A related challenge is that even fairly small gaps in selection rates will be statistically significant given a sufficiently large number of observations. The large number of attributes stored regarding candidates introduces additional dangers, most notably that the risks of construct-irrelevant variance and redundant encoding of protected class status, explained below, increase with the dimensionality of a data set.

1. The Ubiquity and Meaninglessness of Statistical Significance in Large Data Sets

If an employer uses an algorithmic tool to assess hundreds or thousands of candidates, rejected candidates who sue may find that the bar for making out a prima facie case of disparate impact discrimination under current law is remarkably low. Recall that the primary inquiry for prima facie disparate impact focuses on the differences in the rates at which members of protected class groups are selected, and that courts have most often focused on whether those differences are statistically significant.[200] For selection procedures that are used on a few dozen candidates, the magnitude of the difference required for statistical significance is fairly large.

But, all else being equal, the magnitude of the difference necessary for statistical significance diminishes as the number of observations in a data set increases. If a data set has thousands of observations, even very small differences—say a 0.5% difference in selection rates between men and

198. *See* Allan G. King & Marko Mrkonich, *"Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555, 563 (2016).

199. *Id.*

200. *See* discussion *supra* Section II.B.2.b.

women—may nevertheless be statistically significant. Under some interpretations of current law, such a statistically significant difference may, by itself, establish a prima facie case of disparate impact.[201]

Consider the First Circuit's 2014 decision in *Jones v. City of Boston*.[202] In that case, the First Circuit reversed a district court decision that had relied on the four-fifths rule in granting summary judgment to an employer, with the circuit court holding that the four-fifths rule cannot be used to "trump a showing of statistical significance," particularly in cases with a large sample size.[203] Indeed, the court ultimately rejected the notion of an additional "practical significance" requirement for prima facie disparate impact altogether, finding that "any theoretical benefits of inquiring as to practical significance outweighed by the difficulty of doing so in practice in any principled and predictable manner."[204]

Employers seeking to leverage the power of Big Data at scale must either hope for a change in the prevailing winds of case law, or else find ways of eliminating statistically significant disparities between protected groups. But it may be devilishly difficult to reduce differences in selection rates to statistically insignificant levels without using techniques that make direct adjustments on the basis of protected characteristics—a technique that could constitute disparate treatment discrimination.[205] Also, even if a selection procedure were designed and confirmed to have no disparate impacts during testing, disparate impacts may arise over time if the characteristics of the applicant pool diverge from the characteristics of the candidates in the training data. Current case law provides no clear guidance on whether making additional adjustments to the algorithm to reduce such later arising disparate impacts would constitute disparate treatment.

2. *Construct-Irrelevant Variance and Construct Underrepresentation*

The large number of attributes that are available in the age of Big Data will also present novel challenges as courts, agencies, and employers attempt to assess what a business necessity defense might look like in the context of algorithmic tools. A high-dimensionality data set presents an increased risk of construct-irrelevant variance, that is, nonrandom differences in test results that

201. Ricci v. DeStefano, 557 U.S. 557, 587 (2009) (citing Connecticut v. Teal, 447 U.S. 440, 446 (1982)) ("[A] prima facie case of disparate-impact liability [is] essentially, a threshold showing of a statistical disparity . . . and nothing more.").

202. Jones v. City of Boston, 752 F.3d 38 (1st Cir. 2014).

203. *Id.* at 46, 49, 52–53.

204. *Id.* at 53.

205. *See* discussion *infra* Sections III.E, IV.C.

are the result of factors unrelated to the intended construct.[206] This can happen for a variety of reasons, including when a criterion or predictor measures something more or different than the target construct (e.g., if the scores on a mathematical aptitude test are affected by a test-taker's proficiency in written English); or when scores reflect cultural differences rather than (or in addition to) differences in job related competencies. The inverse of construct-irrelevant variance is construct underrepresentation or construct deficiency.[207] This occurs when criterion measures or predictors fail to reflect construct-relevant sources of variance because the criteria or predictors are unrepresentative or otherwise do not capture important aspects of the target construct.[208] Both construct-irrelevant variance and construct deficiency can generate adverse impacts if members of certain subgroups perform differently on the improperly included or excluded aspects of job performance.[209]

The manner in which predictors and the test sample are selected in an algorithmic selection tool creates a risk of construct deficiency and introduces a potential source of construct-irrelevant variance in addition to those that affect traditional employment tests. According to modern test validation literature, the proper method for selecting predictors involves not just searching for statistical relationships between predictors and criteria, but also examining whether there are theoretical and logical reasons to suppose that the predictors are related to the criterion—in other words, that they are related in more than a mere correlational sense.[210]

This was not a major issue for the sorts of employee selection procedures that existed at the time the Guidelines were promulgated because having a conceptual basis for predictor selection is a practical necessity for paper-and-

206. *See* AM. EDUC. RESEARCH ASS'N ET AL., *supra* note 109, at 12–13; Messick, *supra* note 191, at 34.

207. *See* AM. EDUC. RESEARCH ASS'N ET AL., *supra* note 109, at 12–13; Messick, *supra* note 191, at 34.

208. *See* AM. EDUC. RESEARCH ASS'N ET AL., *supra* note 109, at 12–13; Messick, *supra* note 191, at 34; PRINCIPLES, *supra* note 81, at 11–12.

209. *See* PRINCIPLES, *supra* note 81, at 11–12.

210. *See, e.g.*, Messick, *supra* note 191, at 17 ("[E]mpirical relationships between the predictor scores and criterion measures should make theoretical sense in terms of what the predictor test is interpreted to measure and what the criterion is presumed to embody . . . [E]ven for purposes of applied decision making, reliance on criterion validity or content coverage is not enough. The meaning of the measure, and hence its construct validity, must always be pursued . . . ."); PRINCIPLES, *supra* note 81, at 12 ("The rationale for a choice of predictor(s) should be specified. A predictor is more likely to provide evidence of validity if there is good reason or theory to suppose that a relationship exists between it and the behavior it is designed to predict. A clear understanding of the work (e.g., via results of a work analysis), the research literature, or the logic of predictor development provides this rationale. This principle is not intended to rule out the application of serendipitous findings, but such findings, especially if based on small research samples, should be verified through replication with an independent sample.").

pencil employment tests; it would be inefficient, to say the least, for the developers of such a test to provide a sample of hundreds or thousands of random questions to current employees and blindly search the results to see which questions correlate with performance on the criterion measures of interest. Instead, the designers of traditional employment tests select or develop questions because they have a prior reason to believe that there is a relationship between the proposed test questions and the criterion of interest. Choosing predictors based on their theoretical relationship with the target construct thereby allows test designers to be alert to potential sources of construct-irrelevant variance and to ensure that the test is measuring a sufficiently representative set of job related criteria. The hypothesized relationship between predictors and criteria is then tested by analyzing the results of the validation study to see if the test responses correlate with the criterion measures.[211]

But the algorithms that drive ML-based selection procedures do not consider theoretical or logical relationships between variables, or whether the training data includes attributes that constitute a representative set of predictors. The training algorithm instead examines numerous individual attributes and combinations of the attributes available in the training data and then develops a model based on correlations with the criterion measures—without regard to whether there was a prior reason to suppose that the attributes would have predictive value with respect to the criterion. This is both a blessing and a curse. It is a blessing because it has the potential to unearth job related predictors that would not have been obvious to humans. But it also creates a heightened risk that an algorithm will discover and capitalize on chance correlations.[212] That risk that is heightened further when data sets contain a large number of observations (because small differences can constitute statistically significant correlations given a large enough sample size) or attributes (because more attributes also means more opportunities for chance correlations).

In the science world, the tendency of algorithmic tools—particularly those that utilize deep learning—to "discover" chance correlations is already

211. *Cf.* AM. EDUC. RESEARCH ASS'N ET AL., *supra* note 109, at 17 ("[T]he test is not a measure of a criterion, but rather is a measure hypothesized as a potential predictor of that targeted criterion. Whether a test predicts a given criterion in a given context is a testable hypothesis.").

212. PRINCIPLES, *supra* note 81, at 13 ("In cases where scores from . . . . algorithms are used as part of the selection process, the conceptual and methodological basis for that use should be sufficiently documented to establish a clear rationale for linking the resulting scores to the criterion constructs of interest. In addition, when some form of empirical keying is used, clear evidence of cross-validity should be provided prior to operational use to guard against empirically driven algorithms' propensity to capitalize on chance.").

causing a "reproducibility crisis," in the words of Rice University statistician Dr. Genevera Allen.[213] Allen discovered in her research several instances where scientists using deep learning algorithms claimed to have identified previously unknown associations between variables, only to find that other researchers were unable to reproduce the results when applying the same techniques to different data sets.[214] They discovered associations between variables that existed only in the particular samples available to the researchers, but those associations had no generality because these correlations were absent from different sets of similar data.[215]

Similar phenomena pose a substantial threat to validity for users of algorithmic employee selection tools. First, as with the genomic and health research that was the focus of Allen's study,[216] there is a risk that algorithmic selection tools will discover correlations between variables in the training data that do not actually exist in the broader real-world applicant pool. While machine learning offers a number of well-accepted techniques for cross-validation, those methods may not be adequate to weed out all of the construct-irrelevant associations between variables in large data sets, particularly if a data set contains information on thousands (or tens or hundreds of thousands) of attributes.

There is another type of correlation that can also afflict employee selection procedures—associations between attributes that do hold in the population at large but that are nevertheless construct irrelevant. The number of such correlations may increase if the training examples tend to come from individuals from the same demographic group or groups, and who therefore share non-job-related attributes in the data. For example, if musical tastes differ by race, and the best incumbent job performers for a particular position are predominantly from a given race, then a high correlation between musical taste and job performance may exist—but only due to demographics, and not because musical taste is an accurate and generalizable predictor of job performance. The less representative the training data are of the population at large, the higher the risk that a deep learning model will identify and create a model that relies upon such demographics-dependent correlations.

213. Pallab Ghosh, *AAAS: Machine Learning 'Causing Science Crisis'*, BBC (Feb. 16, 2019), https://www.bbc.com/news/science-environment-47267081 [https://perma.cc/VT3H-KQHW].

214. *Id.*

215. *Id.* The *Principles* allude to this potential problem when discussing validation in the context of algorithmic selection procedures. *See* PRINCIPLES, *supra* note 81, at 13 ("[W]hen some form of empirical keying is used, clear evidence of cross-validity should be provided prior to operational use to guard against empirically driven algorithms' propensity to capitalize on chance.").

216. *See* Ghosh, *supra* note 213.

An example of this phenomenon can be seen in the results of the MIT Media Lab Gender Shades study.[217] That study examined the accuracy of gender classification systems—that is, machine learning software that takes a photograph of a person as its input and outputs a predicted classification of that person's gender as male or female.[218] The MIT study used the gender classification systems on photographs of Northern European and African politicians.[219] The study showed that each of the three facial recognition platforms was more accurate in classifying the European legislators than their African counterparts.[220] Not only that; the study also indicated that the accuracy of the tool was generally better for people with skin types typically associated with moderately dark skin than those with very dark skin.[221]

The authors hypothesized that this may be because darker skinned individuals may have been "less represented in the training data."[222] If so, the tool's accuracy might have been diminished either because of the dissimilarity of darker subjects' skin from those that dominated the training data set or because darker skin may be highly correlated with other gender-distinctive attributes that were also underrepresented in the training data.[223] The tool may thus have learned attributes useful for distinguishing white males and white females, while devaluing gender-distinctive attributes present in individuals with darker skin, and underweighting those attributes that actually are useful predictors across the population as a whole.

This is an illustration of a broader challenge with correlation-based selection: the more dissimilar an individual is from the population that served as training examples, the less reliable the tool's output will be for that individual. That could lead to undesirable—and perhaps unlawful—outcomes with algorithmic employee selection tools.[224] In the employment setting, if the positive examples used in the training data are predominantly individuals with a certain set of protected class characteristics, the data may tell the tool that

---

217. JOY BUOLAMWINI & TIMNIT GEBRU, GENDER SHADES: INTERSECTIONAL ACCURACY DISPARITIES IN COMMERCIAL GENDER CLASSIFICATION (Sorelle A. Friedler & Christo Wilson eds., 2018).

218. *Id.* at 1.

219. *Id.* at 5.

220. *Id.*at 8, 10.

221. *Id.* at 7 (noting the difference in the distributions of lighter and darker skinned subjects, labeled according to the Fitzpatrick classification system).

222. *Id.* at 10.

223. *See id.*

224. Among psychologists, the term "subgroup validity" refers to the different validity coefficients that can arise between tested subgroups, and the differences in those coefficients are termed "differential validity." *See* Richard J. Klimoski & Lori B. Zukin, *Psychological Assessment in Industrial/Organizational Settings*, *in* 10 HANDBOOK OF PSYCHOLOGY: ASSESSMENT PSYCHOLOGY 317, 324 (Irving B. Weiner ed. 2003). The issue of differential validity is discussed further in Section IV.C.1.

those individuals' attributes—whether construct-relevant or not—are associated with success for the position in question. The more highly qualified candidates' attributes differ from the training benchmarks, the more the algorithm's ability to identify those candidates would diminish.

As an example, say that a company was training an algorithmic tool to recognize good software engineers using training data that reflects the demographics of their best current network engineers, who are predominantly white males. If these employees share, as is likely, construct-irrelevant characteristics that are reflected in the training data, the tool will learn to associate those characteristics with good job performance. This could have two related adverse impacts on qualified candidates who are not white males. First, if the ablest female and nonwhite candidates have attributes (whether construct-relevant or not) that differ from those of the white males who dominate the current sample, the tool's accuracy will be lower when scoring those candidates, just as the gender classification programs in the MIT study were less accurate when attempting to classify individuals with darker skin. Second, the individuals that the tool identifies as the best candidates from the underrepresented groups may have scored highly not because of characteristics that affect their actual competence, but because of the construct-irrelevant characteristics they share with the current software engineers.

Both of those factors may drive down the number of qualified female and minority candidates that the tool selects. In addition, the candidates who the tool does recommend from the disadvantaged group are less likely to be the most competent candidates from that group, which may reduce the likelihood that they are ultimately hired and retained. Through these mechanisms, an employer's adoption of an algorithmic tool could inadvertently reinforce existing demographics.

If courts and agencies reassess the legal standards of employee selection procedures to bring them in line with modern scientific standards, the resulting new standards will likely include a requirement that an employer demonstrate some level of construct relevance—as opposed to relevance in the correlational sense—for algorithmic selection procedures. In either case, employers may find conducting a legally compliant validation cumbersome at best and infeasible at worst, given the sheer number of attributes that would need to be reviewed. The task would be doubly challenging in the context of a deep learning tool, which may transform the input variables into representations that are not human interpretable.[225] Because courts have never ruled on the requirements of validity studies in the context of algorithmic selection procedures that utilize thousands of features, it simply is not clear

225. *See* discussion *infra* Section III.C.

how courts will treat such tools if they produce a disparate impact and the employer is unable to explain how and why the variables considered and constructed by the tool were relevant to the job in question.

3. Redundant Encodings

On the surface, it may seem easy for the developer of an algorithmic selection tool to design around disparate treatment—simply ensure that gender, race, and other protected class status information is not made available to the selection tool during training. But in the age of Big Data, it may not be that simple. First, it may be difficult to reliably excise protected class status information if the training data pools information on candidates from a variety of sources, each of which may encode the sensitive characteristic differently. Even if employers overcome that hurdle, however, a tool trained on data sets of high dimensionality could effectively reconstruct a protected characteristic from other attributes with which it is correlated, a problem called redundant encoding.[226] When that occurs, the redundant encoding effectively creates a reliable proxy for the protected characteristic, even if it does not use the characteristic itself.[227]

If the tool is able to reconstruct the protected characteristic, has the tool engaged in disparate treatment? Or does the fact that it did not explicitly consider the candidate's gender mean that the redundant encoding is facially neutral, such that disparate impact provides the proper analytical rubric? Unsurprisingly, this issue is not addressed in antidiscrimination case law, meaning that courts and agencies will have to decide which rubric to use when faced with redundant encodings.

Say that redundant encoding allows the algorithm to reconstruct a person's sex with 99.9% accuracy—say, by using the candidate's height, weight, college attended, and recent clothing purchases—and uses the resulting proxy for sex as part of the model. If the model then systematically disfavors women, women may plausibly argue that they were rejected because of their sex. Such a ruling would be consistent with the prevailing trend in case law, under which courts have increasingly held that, because Title VII prohibits discrimination because of sex, the prohibition against disparate treatment covers "not just discrimination based on sex itself, but also discrimination based on traits that are a function of sex."[228] Thus, courts have

226. *See* CYNTHIA DWORK ET AL., FAIRNESS THROUGH AWARENESS 22 (2011).

227. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 695 (2016).

228. Zarda v. Altitude Express, Inc., 883 F.3d 100, 111–12 (2d Cir. 2018); *see also* Hively v. Ivy Tech Cmty. Coll., 853 F.3d 339, 339 (7th Cir. 2017). *See generally* Mary Stuart King,

held that using attributes related to sex, such as life expectancy,[229] conformance to gender norms,[230] and sexual preference[231] constitutes disparate treatment.

But it is not clear how far disparate treatment liability may extend when the discrimination is based on proxy characteristics. One court attempted to draw a distinction between characteristics that are a "proxy" for a protected characteristic and those that merely "correlate" with it.[232] But it is unclear where the line between proxy and correlate lies. It is difficult to imagine a court countenancing a model that uses a predictor variable that perfectly correlates with a protected characteristic. But what about a predictor variable with an R-squared value of 0.99 with respect to the protected characteristic? Or 0.8? Or 0.5? Until these questions are resolved, employers cannot afford to assume that they can insulate themselves from disparate treatment liability risk simply by removing demographic data and related information from the training data.[233]

C. The Black Box Problem

Perhaps the issue that legal commentators raise most frequently when discussing algorithmic selection tools is the black box problem—that is, that it may be difficult or impossible for a human to reconstruct or interpret the logical steps that the tool took when assessing the fitness of a candidate for a particular job. In this way, ML-powered selection tools share much in common with human decision makers, whose reasoning behind a particular selection decision may not be apparent to outside observers. But human

---

Note, *To Protect or Not to Protect: An Empirical Approach to Predicting Where the Fourth Circuit Would Stand on Coverage for Sexual Orientation Discrimination Under Title VII*, 705 S.C. L. REV. 1075, 1076–80 (discussing the expansion and development of Title VII in regards to sex as a protected class).

229. *See* City of L.A. Dep't of Water & Power v. Manhart, 435 U.S. 702, 711 (1978) (finding Title VII violation where employer "require[d] 2,000 individuals to contribute more money into a fund than 10,000 other employees simply because each of them is a woman, rather than a man," even though contributions were based on observed actuarial differences between the sexes in longevity).

230. *Zarda*, 883 F.3d at 112 (citing Price Waterhouse v. Hopkins, 490 U.S. 228, 250–51 (1989)).

231. *Id.* at 113–15.

232. Bowers v. Nat'l Collegiate Athletic Ass'n, 563 F. Supp. 2d 508, 517–18 (D.N.J. 2008).

233. Even if the effect of redundant encodings was subject only to disparate impact analysis, the presence of redundant encodings would still pose liability risks to employers. The presence of such encodings could lead to gaps in selection rates for protected groups. And if the attributes that generated the redundant encodings are construct-irrelevant, an employer would likely be unable to establish a business necessity defense.

decision makers can be put on the witness stand and forced to explain their reasoning. Their underlying motivations for a particular decision may also be illuminated by other evidence, such as emails, text messages, conversations with friends, and social media activity. Machines are, for now at least, not able to testify regarding their decisions, and because ML algorithms are effectively built on the closed universe of their training data, little other evidence will likely be available that could shed light on how an algorithmic selection tool arrived at a particular score or recommendation for a particular candidate.

With the rise of deep learning, this inscrutability is not simply a problem for plaintiffs and courts. One result of the complexity of deep neural networks is that the precise inner workings of an algorithm may be indecipherable even to the algorithm's designers.[234] While Title VII does not prohibit opaque selection procedures per se, the potential opacity of algorithmic tools will present considerable practical challenges for both plaintiffs and employers in discrimination suits based on the use of such tools once adverse impact is established.

For example, consider what would happen if redundant encodings of protected characteristics allowed algorithmic tools to essentially reconstruct the protected characteristics themselves, with discriminatory effects on certain protected groups. Regardless of whether courts characterize any resulting discrimination as disparate treatment or disparate impact, the employer may have difficulty deciphering whether—much less how—redundant encoding arose. This would complicate both efforts to rectify the discrimination and preparation of an adequate legal defense.

Of course, plaintiffs would have difficulty determining how the discriminatory output had been generated as well. In a disparate impact case, plaintiffs are responsible for identifying the subset of attributes responsible for the redundant coding, unless they can prove the attributes are "not capable of separation for analysis."[235] That may seem to suggest that employers who use such systems may escape liability for discrimination. But if the tool is as opaque to the employer as it is to the employee, it is difficult to predict whether employers or employees will suffer the greater disadvantage from the tool's opacity.

If courts view the discrimination through a disparate impact lens, the employer would seem to be at a greater strategic disadvantage than the plaintiff. Because the final output of the tool is not a black box, a plaintiff would have little difficulty determining whether the ultimate effect of the tool

234. *See generally* Will Knight, *The Dark Secret at the Heart of AI*, MIT TECHNOLOGY REVIEW (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [https://perma.cc/F249-2N9D] (discussing the difficulty in assessing results from AI technology).

235. 42 U.S.C. § 2000e-2(k)(1)(B)(i) (2012).

was to disproportionately disfavor a protected class, as necessary to establish a prima facie case. On the other hand, the employer's efforts to establish the validity of the procedure would be complicated by the impracticability of tracing the neural network's transformation of the original input attributes into the final parameters used by the model. This is particularly true if courts require validation of the individual components of an algorithmic selection procedure. In such a situation, the employer might find itself hamstrung by its inability to identify and validate the components of the model that are having an adverse impact. This problem becomes even more serious—and perhaps intractable—if an algorithmic tool is updated frequently or continuously as new data is received. In such situations, the employer may not have a practical way of reconstructing the algorithm's parameters at the relevant time(s). If, as the Supreme Court has held, selection procedures are inadequate when their validation studies rely on ratings that are "vague and fatally open to divergent interpretations,"[236] it is unlikely that a court will be satisfied by a selection procedure whose standards are completely opaque and not open to any human-decipherable interpretation.

If courts hold that the use of a reconstructed protected characteristic constitutes disparate treatment rather than disparate impact, it is not clear that employers would fare much better. While a plaintiff might find it impossible to explain how an algorithmic tool discovered redundant encodings of a protected characteristic, it is not difficult to imagine courts taking a res ipsa loquitur attitude if it appears obvious that a tool is employing an effective proxy for a protected characteristic.[237] If so, current law does not appear to provide employers with an easily identifiable defense; the *McDonnell Douglas* framework is inapplicable if courts determine that the use of a redundant encoding is tantamount to use of the protected characteristic itself, and therefore direct evidence of discrimination.

D. A Clear Target

One of the most striking consequences of the *Griggs* decision and the subsequent development of disparate impact litigation has been the deformalization of employee selection procedures. As Lex Larson has observed, starting with *Griggs*, fear that testing would generate liability for disparate impact has driven many employers toward increased reliance on subjective decision-making:

236. *See* Albemarle Paper Co. v. Moody, 422 U.S. 405, 433 (1975).

237. *See* RESTATEMENT (SECOND) OF TORTS § 328D (AM. LAW INST. 1965) (describing tort doctrine of *res ipsa loquitur* as a basis for negligence liability).

> This dramatic reversal in business' attitude toward testing was tinged with irony; for the most part, businesses had moved toward the use of tests as a way to lend objectivity to the selection process to select the best-qualified personnel. Starting with *Griggs*, the courts began telling employers that these devices, too, could result in discrimination. As a result, many employers went back to using subjective judgment in making employment decisions.[238]

Of course, reliance on human judgment can lead to adverse impacts as well—which is precisely why algorithmic tools represent an appealing alternative. But subjective human judgments leave a lesser paper trail than more formal hiring practices. It is also harder to cast such subjective decision-making by numerous different decision makers as a unified employment practice that could serve as the basis for a class action disparate impact suit.[239] These characteristics make relying on the humans in human resources more appealing, particularly in comparison to the lengthy, costly, and uncertain process of designing and validating a formal selection procedure.

These drawbacks are equally, if not more, apparent in the specific context of algorithmic selection procedures. The very essence of an algorithmic selection procedure is to take the observable characteristics of a candidate and reduce them to rows of data. The output of the selection procedure is essentially a function of complex mathematical formulae. The process simply does not work unless both the candidates and the selection procedure that assesses them are formalized and ultimately reduced to computer code, and the procedure loses its value if it is not used consistently for all candidates under consideration for a given position. When disparate impacts arise, algorithmic selection procedures give potential plaintiffs an obvious target.

The inherent explicitness will also muddy the waters in disparate treatment cases. It is trivial for an employer to ensure that an algorithm does not use a protected characteristic as an input when assessing a candidate. But if the algorithm reconstitutes the protected characteristic through redundant encodings, and if courts hold that using such redundant encodings constitutes disparate treatment, it will be equally trivial for a plaintiff to demonstrate that the algorithmic selection procedure is the source of the disparate treatment. The inner workings of the algorithm may be opaque, which will hinder plaintiffs' ability to demonstrate precisely how a protected characteristic was

238. LARSON, *supra* note 84, at § 25.02.

239. *See* Wal-Mart Stores, Inc., v. Dukes, 564 U.S. 338, 356–57 (2011) (rejecting class certification because "[r]espondents have not identified a common mode of exercising discretion that pervades the entire company" and different managers would likely say they were using different criteria when making decisions).

reconstituted. But as discussed elsewhere in this Article,[240] that may not provide employers with an escape route.

The explicitness of algorithmic selection procedures will also complicate employers' efforts to navigate the intersection of disparate impact with disparate treatment, as considered in the *Ricci* case. In the algorithmic age, it will be easier than ever for employers to eliminate disparate impacts in their selection procedures—but race norming, boosting algorithmic scores of candidates from disadvantaged groups, and other preferential practices risk disparate treatment liability. Less direct methods of eliminating disparate impacts remain untested in court, leaving employers with no clear options regarding how to cure disparate impacts when they arise.

And it is almost inevitable that at least some disparate impacts will arise. Even if an employer succeeds in designing an algorithmic selection procedure that has no disparate impacts during initial training, adverse impacts may creep in as the characteristics of candidates and successful employees in a given position change. Making changes after a tool has already been deployed is problematic under *Ricci*, which held that such modifications may be made only prospectively.[241] Employers will then be forced to make conscious decisions about how to manage those adverse impacts, and any adjustments made to the model in response will themselves have to be reduced to computer code and explained during the course of litigation. Faced with this morass of legal uncertainty, many employers may prefer to continue to rely on subjective human judgment—and with it, the potential effects of human prejudice—rather than risk getting bogged down in the marsh of an unsettled area of law.

E. Disparate Treatment: A Brave New World

Given the manner in which disparate-treatment case law has developed, concerns have been raised regarding whether companies might be effectively immune from disparate treatment liability if they use algorithmic selection devices that learn, without any express human programming, to classify workers in a discriminatory manner on the basis of protected characteristics.[242] The premise is the belief that because machines cannot have "intent" in the human sense, there can be no liability for their actions under Title VII unless the machine was intentionally programmed to discriminate.[243] This concern seems misplaced.

240. *See* discussion *supra* Section III.C; *infra* Section III.E.
241. *See* Ricci v. DeStefano, 557 U.S. 557, 585 (2009).
242. *See* Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395, 404–10 (2018).
243. *See id.*

First, fixating on intent means ignoring the clear anticlassification rule set forth in the statutory text. Under the plain text of § 703(a), a Title VII violation occurs whenever an adverse employment or hiring action is because of a protected characteristic.[244] That is language of causation, not intent. The disparate impact theory of discrimination itself first arose out of this language, with the Supreme Court explicitly holding that employers cannot escape liability under § 703(a) for practices with discriminatory effects simply by pleading lack of intent:

> [G]ood intent or absence of fdiscriminatory intent does not redeem employment procedures or testing mechanisms that operate as 'built-in headwinds' for minority groups and are unrelated to measuring job capability.
>
> The Company's lack of discriminatory intent is suggested by special efforts to help the undereducated employees through Company financing of two-thirds the cost of tuition for high school training. But Congress directed the thrust of the Act to the consequences of employment practices, not simply the motivation.[245]

The frequent connection of disparate treatment with intent in the case law has never been cast as mandated by the statutory text. More likely, it is a consequence of the fact that, up to now, hiring practices have been driven by human decision makers.

To that point—who is to say that courts would necessarily conclude that machines cannot possess intent? True, many definitions of intent reference a "state of mind" or a "conscious" desire to bring about a particular result, terms that seem to refer to distinctly human traits.[246] But other definitions are far broader, focusing only on the party's "objective" or "purpose."[247] Under the criminal laws of many states, an entire category of "general intent" exists where the defendant's state of mind is not relevant so long as the defendant

244. 42 U.S.C. § 2000e-2(b) (2012).

245. *See* Griggs v. Duke Power Co., 401 U.S. 424, 432 (1971).

246. *See, e.g.*, *Intent*, BLACK'S LAW DICTIONARY (11th ed. 2019) (defining intent, in part, as "[t]he state of mind accompanying an act . . . .").

247. *See, e.g.*, WASH. REV. CODE ANN. § 9A.08.010 (West, Westlaw through 2019 Reg. Sess.) ("A person acts with intent or intentionally when he or she acts with the objective or purpose to accomplish a result which constitutes a crime."); State v. Salinas, 423 P.3d 463, 465 (Idaho Sup. Ct. 2018) ("Intent 'is the purpose to use a particular means to effect a certain result.'") (citing State v. Stevens, 454 P.2d 945, 950 (Idaho Sup. Ct. 1969)).

acted volitionally as opposed to accidentally.[248] Tort law treats an act as "intentional" when the actor believes that the consequences of his act are "substantially certain" to result from it.[249] Such definitions of intent could easily apply to decisions made by machines. It will not do to simply assume that because machines generally are not considered to have consciousness in the metaphysical sense, they necessarily cannot possess intent in the legal sense or that intent cannot be imputed to those who deploy them.

Moreover, intent has proven to be quite a malleable concept in the context of Title VII, as in other areas of law. The *Ricci* majority stated that disparate treatment requires intent but, at the same time, acknowledged that the employer's objective in that case was avoiding legal liability;[250] to the extent that race factored into the decision, the employer's intent was not to treat workers differently on the basis of race but rather to avoid discrimination on the basis of race. Nevertheless, citing Title VII's use of the broad term *because of*, the Court treated that motivation as itself a form of disparate treatment.[251]

Employers can also be held liable for sexual harassment even if the harassment was committed by nonemployees and even if the employer had no actual knowledge of the harassment.[252] A number of courts have also upheld the "cat's paw" theory of discrimination,[253] under which "an employer who acts without discriminatory intent can be liable for a subordinate's discriminatory animus if the employer uncritically relies on the biased subordinate's reports and recommendations in deciding to take adverse employment action."[254] If a court is willing to find intent based on a decision maker's uncritical reliance on another person's biased recommendation, it seems highly unlikely it would excuse an employer for uncritically relying on the recommendation of a machine it chose to use, regardless of the metaphysics of whether an algorithm can have intent.

Lastly, even if some element of human intent were an absolute requirement for disparate treatment liability, algorithmic selection tools will

248. *See, e.g.*, United States v. Lamott, 831 F.3d 1153, 1156 (9th Cir. 2016) ("In a crime requiring 'specific intent,' the government must prove that the defendant subjectively intended or desired the proscribed act or result. By contrast, a general intent crime requires only that the act was volitional (as opposed to accidental), and the defendant's state of mind is not otherwise relevant.").

249. RESTATEMENT (SECOND) OF TORTS § 8A (AM. LAW INST. 1965).

250. *See* Ricci v. DeStefano, 557 U.S. 557, 579–80 (2009).

251. *Id.*

252. 29 C.F.R. § 1604.11(e) (2019).

253. Thomas v. Berry Plastics Corp., 803 F.3d 510, 514 (10th Cir. 2015); *see also* Vasquez v. Empress Ambulance Serv., Inc., 835 F.3d 267, 271–73 (2d Cir. 2016); Lust v. Sealy, Inc., 383 F.3d 580, 584 (7th Cir. 2004); *cf.* Staub v. Proctor Hosp., 562 U.S. 411, 421–22 (2011) (upholding a court's use of the cat's paw theory of discrimination, albeit in a case brought under the Uniformed Services Employment and Reemployment Rights Act, not under Title VII).

254. *Thomas*, 803 F.3d at 514.

very much be the product of human motivations and intentions. Because of the need to validate selection procedures that may have a disparate impact, a topic discussed further below, algorithmic selection tools will rely on data that is labeled by humans—ideally, managers or HR employees for the company seeking to use the tool—tasked with assessing the fitness of candidates in the training data for a particular job. Those labelers' motivations and intentions are incorporated, however indirectly, into the final selection procedure. Similarly, the training data itself will ideally include employee performance data, such as supervisor ratings. Because the input of human decision makers will be baked into the algorithm, it is difficult to imagine courts and enforcement agencies shrugging their collective shoulders and holding that employers who rely on the recommendations of algorithmic selection tools are immune from disparate treatment liability.

For these reasons, the disparate treatment doctrine will not fade into legal obscurity in the age of algorithms. But it is true that courts developed the prevailing judicial interpretations of the disparate treatment doctrine with human decision-making in mind and that the contours of disparate treatment liability in the context of algorithmic tools have yet to be established. This means that courts and agencies will have to consider the meaning of the statutory text afresh if or when they are faced with algorithmic selection tools that classify candidates on the basis of a protected characteristic—regardless of whether such classification was intended by the algorithmic tool's designers or users.

IV. NEW RULES FOR THE NEW TOOLS: A PROPOSED LEGAL FRAMEWORK FOR ALGORITHMIC SELECTION TOOLS

A. Overview

Title VII requires (1) that employers avoid making employment decisions because of protected characteristics and (2) that employers establish the job relatedness of any selection tool that has an adverse impact on one or more protected groups. Although the legal regime governing employee selection tools was developed without algorithmic selection tools in mind, the broad principles set forth in the statutory text certainly can be applied to algorithmic selection procedures. What is required is not so much a new legal framework as a new conceptual approach to assessing employee selection procedures in the age of algorithms.

In particular, algorithmic selection procedures require taking the fundamental principles of Title VII, and the landmark Supreme Court cases interpreting them, and developing a set of standards that address the unique challenges posed by AI and Big Data discussed in Part III. The ultimate goal

should be to allow employers to find innovative ways of uncovering talent and building a diverse workforce—objectives fully consistent with Title VII—while remaining true to the purpose of Title VII itself. It should not be difficult to reconcile these objectives because selecting the highest quality candidate for a job while ensuring broad participation by disadvantaged groups is, as the Supreme Court held in *Griggs*, the very essence of Title VII.[255] Algorithmic selection tools create an unprecedented opportunity to advance these goals by excising human prejudice and bias from personnel decisions.[256]

Our proposal weaves the disparate impact and disparate treatment inquiries into a single analytical framework:

Step 1: Determine whether the algorithmic procedure had an unlawful disparate impact.

a. Prima facie disparate impact: Determine whether the gap between protected groups is large enough to give a reasonable employer concern that the algorithmically generated model is unreasonably disadvantaging members of a protected group.
   i. *If no unreasonable gap exists, skip to Step 2.*

b. Employer's defense: If a disparate impact exists, determine whether the tool has been properly validated.
   i. *If the tool has not been validated, the employer is liable for disparate impact, and the court or agency should skip to Step 2 to determine whether the employer also engaged in disparate treatment.*

c. Less discriminatory alternative: Determine whether the employer considered and rejected an alternative modeling method that would have affected a reasonable reduction in adverse impact but would have continued to meet the employer's legitimate objectives.

Step 2: Determine whether the ML tool used any methods that make prohibited classifications or that otherwise constitute disparate treatment.

255. Griggs v. Duke Power Co., 401 U.S. 424, 436 (1971) ("Far from disparaging job qualifications as such, Congress has made such qualifications the controlling factor, so that race, religion, nationality, and sex become irrelevant.").

256. *See* Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 163–64 (2018).

In broad strokes, with details to follow below, the first step is to determine whether use of the tool has a prima facie disparate impact on one or more protected groups. Because the currently favored statistical significance approaches to prima facie disparate impact would sweep too broadly in the age of Big Data, however, a modified approach to disparate impact analysis is required. Instead of focusing on the presence or absence of statistical significance, the inquiry should be one of reasonableness—a plaintiff can establish a prima facie case of disparate impact demonstrating by producing evidence demonstrating that the gap between protected groups is large enough to give a reasonable employer concern that the algorithmically generated model may be disproportionately disadvantaging members of a protected group.

The next step in the analytical process depends on whether a prima facie disparate impact exists. If it does not, the disparate impact inquiry ceases, and the only remaining issue is whether the algorithmic tool used techniques that constitute disparate treatment.

If, on the other hand, the plaintiff does present prima facie proof of disparate impact, the inquiry would instead progress to whether the algorithmic assessment is job related and consistent with business necessity. The key inquiry here would be whether the criteria that serve as target variables for the training algorithm represent "essential" and "important" job functions, as identified through standard job analysis.[257] Essential functions can be used as screening criteria for an algorithmic selection tool; that is, employers can use algorithmic selection tools to screen out candidates the tool identifies as lacking the ability to perform essential job functions. Important job functions can also be used as target criteria to be optimized, but they cannot be used as hard screening devices. As in ADA cases, the employer's designation of essential and important job functions would be entitled to some deference. An algorithmic tool would be considered job related if the criteria meet these requirements, if the outputs of the algorithmic assessment are significantly correlated with adequate measures of those criteria, and if the employer demonstrates that it took reasonable steps to guard against construct-irrelevant variance in the results. The employee can rebut this showing with proof that the employer used criteria that were not job related or failed to model these dimensions correctly.

The third step of the current disparate impact analysis—the plaintiff's burden of demonstrating the existence of a less discriminatory alternative—would, in the case of algorithmic tools, require a plaintiff to show that the employer considered and rejected an alternative modeling method that would

257. *See* 29 C.F.R. § 1607.14(B)(2) (2019); § 1630.14(b)(3).

have effected a reasonable reduction in adverse impact but would have continued to meet the employer's legitimate objectives. Once again, we eschew the requirement of a statistically significant reduction because of the likelihood that any reduction in adverse impact, in a Big Data world, would meet that criterion. In the same vein, any reduction in the accuracy with which this alternative modeling method selected the best employees also would be deemed statistically significant in a world of Big Data.

After the disparate impact analysis concludes, attention should turn to whether the algorithm used any methods that constitute unlawful disparate treatment. The framework identifies two techniques through which employers, during the development and training process, should be permitted to take measures to prevent disparate impacts without exposing themselves to disparate treatment liability.

B. Disparate Impact

1. Prima Facie Disparate Impact

The standard approach to determining whether prima facie evidence of a disparate impact exists relies on formal statistical tests, with most courts relying on a bright-line rule that statistically significant differences in selection rates between favored and disfavored groups suffice to prove the first element of a disparate impact claim.[258] In the era of Big Data, this criterion is no longer appropriate because, all else equal, the larger the sample, the smaller the differences that will be deemed statistically significant. At a certain point—which we are fast approaching for practical purposes—all differences, no matter how small, will be statistically significant. That means that a statistical significance requirement will be meaningless. How then should courts assess disparities when statistical significance no longer is a useful criterion for distinguishing discriminatory from nondiscriminatory assessment methods?

One possible policy response to the diminishing meaningfulness of statistical significance would be to abolish the disparate impact doctrine altogether. The doctrine has been criticized by some legal commentators and jurists on constitutional grounds, including Justice Scalia in his *Ricci* concurrence.[259] And in practical terms, one could argue that in the age of Big Data, which allows for a richer analysis of candidates while reducing the practical significance of statistical tests, the doctrine has simply outlived its usefulness. But it usually is not possible to design an employment test, whether algorithmic or not, that is so comprehensive that it captures all

258. *See* discussion *supra* Section III.B.1.

259. *See* Ricci v. DeStefano, 557 U.S. 557, 594–96 (Scalia, J., concurring).

characteristics predictive of good job performance. Moreover, in the context of algorithmic selection tools, the effects of past discrimination may be baked into training data, meaning that unchecked reliance on existing data sets could repeat and reinforce existing patterns of discrimination. Similarly, the amount of statistical noise inherent in large data sets create too many opportunities for an algorithm to settle on parameters that relate more to demographic characteristics than to ability to perform the job. As a result, the concept of disparate impact discrimination still has a place in the age of algorithms.

But rather than rigidly relying upon statistical significance—which is not, in any event, mandated by any statute—courts and agencies should substitute a less formal reasonableness criterion when assessing whether a prima facie disparate impact.[260] In other words, policies and practices would be deemed to have a disparate impact only when selection rates between groups differ unreasonably. Although this dispenses with the certainty that a purely statistical rule provides, the loss of that certainty is more than offset by the benefits of adopting a more flexible standard that can be adapted to the changing nature of algorithmic tools and the data sets that they use.

Applying a more flexible test should not be especially difficult; courts have, after all, hardly adhered to a uniform, bright-line rule with respect to statistical tests in the context of disparate impact suits. The Supreme Court's "two or three standard deviations" formulation is not a bright-line rule, and the Court's endorsement of this standard was arguably in dictum and is weaker than generally supposed.[261] And while courts have generally preferred to use tests of statistical significance, a substantial number of courts have looked to the Guideline's four-fifths rule or otherwise examined the magnitude of the

260. *See generally* Allan G. King, *"Two or Three Standard Deviations" from What?: How* Gross v. FBL Financial Services *Changes the Statistical Benchmark in ADEA Collective Actions*, 37 EMP. REL. L. J. 17 (2011) (describing a system for finding a reason alone, as long as that reason taints the employers decision-making).

261. The Court observed in *Castaneda v. Partida* that the statistical disparity at issue, in excess of 12 standard deviations, was probative because scientists routinely consider differences of just two or three standard deviations sufficient to reject a null hypothesis. *See* Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977) ("Thus, in this case the standard deviation is approximately 12. As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist."). Strictly speaking, this is not the case's holding because its affirmance would be equally consistent with a rule that 10 standard deviations was required to prove discrimination. Indeed, Justice O'Connor subsequently noted in *Watson*: "Our formulations, which have never been framed in terms of any rigid mathematical formula, have consistently stressed that statistical disparities must be sufficiently substantial that they raise such an inference of causation." Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 994–95 (1988).

disparity rather than applying a rigid statistical significance rule.[262] The theoretical certainty that mathematical tests provide thus has not been consistently attained in practice.

In any event, reasonableness tests are eminently workable, as their continuing popularity and ubiquity in law indicate. Criminal law, tort law, contract law, and, indeed, employment law are all replete with reasonableness tests that courts interpret and apply on a regular basis. In employment law, courts routinely assess whether a proposed accommodation for an employee with a disability is reasonable,[263] whether an employment decision in an age discrimination case was motivated by reasonable factors other than age,[264] and what amount of attorney fees are reasonable for a prevailing plaintiff,[265] among many other examples. Reasonableness standards give courts the ability to avoid the unjust results that can accompany hard-and-fast rules.

In the context of assessing whether a prima facie disparate impact exists, the inquiry into whether a gap in selection rates is unreasonably large should not focus on whether the gap is sufficiently justified or explained by the criteria that underlie the selection procedure; that falls more properly within the realm of the business necessity defense. Rather, the test should be whether, in light of the magnitude of the difference in selection rates and the size of the affected candidate pool, is the gap large enough to permit a reasonable fact finder to conclude that the test systematically disadvantages members of a protected group. If the gap raises such a concern, then the employer would be required to demonstrate that the selection procedure is job related and consistent with business necessity. The prima facie case would therefore serve a gatekeeping function, protecting employers from having to validate gaps that, while significant in the statistical sense, are meaningless in practical economic and legal terms.

In assessing whether a gap is reasonable, the statistical significance of a gap would be one factor, but it would be assessed alongside indicators of the magnitude of the gap, such as an odds ratio or other measures of effect size. Courts and agencies could substitute other rules of thumb to serve as benchmarks for magnitude, as the Guidelines did with the four-fifths rule. This would allow courts and agencies to recoup some of the lost efficiencies that come with a bright-line rule.

---

262. *See, e.g.*, M.O.C.H.A. Soc'y, Inc., v. City of Buffalo, 689 F.3d 263, 274 (2d Cir. 2012) ("Consistent with our precedent, the district court properly deferred to [the four-fifths rule] in finding M.O.C.H.A. to have carried its *prima facie* burden"); Allen v. City of Chicago, 351 F.3d 306, 310–12, n.5 (7th Cir. 2003) (holding that "promotions made on the basis of the assessment exercise did have a disparate impact on African-American and Hispanic officers" after applying the four-fifths rule).

263. *See* 42 U.S.C. § 12112(b)(5) (2012).

264. 29 U.S.C. § 623(f)(1) (2012).

265. 42 U.S.C. § 2000e-5(k).

2. Business Necessity Defense

If a plaintiff does establish a prima facie case of disparate impact, the burden shifts to the employer to show that it has validated the selection procedure and demonstrated its job relatedness. Because content-related evidence will not be sufficient to validate a selection procedure based on passive data,[266] the most plausible route to validation for algorithmic tools will rely on criterion-related evidence of validity.

Under the Guidelines, the criterion validation process must begin with a careful job analysis to "determine measures of work behavior(s) or performance that are relevant to the job."[267] These measures can then be used as criteria in the validation study if they "represent important or critical work behavior(s) or work outcomes."[268] There is no reason to depart from these basic principles when validating an algorithmic selection procedure. But a slight change in wording would help ensure consistency across discrimination laws and obviate the need to select different criterion measures for different protected classifications. Specifically, and borrowing from the statutory language of the ADA, the criteria should reflect essential or important job functions.

The EEOC's interpretive "Questions and Answers" on the Guidelines support this substitution. That guidance document states that if a particular work behavior is essential to the performance of a job, that behavior is "critical" within the meaning of the Guidelines, even if a worker does not spend much work time engaged in that behavior.[269] The Q&As use the example of a machine operator for whom the ability to read is "essential" because the worker must be able to read simple instructions, even though the reading of those instructions "is not a major part of the job."[270] The essential nature of being able to read instructions is thus a critical task for purposes of the Guidelines.

The concept of essential job functions is central in the ADA, where it is closely identified with that statute's version of the job relatedness test.[271]

266. Content -related evidence would be a more persuasive indicator of validity could be a viable path for interactive algorithmic tools that incorporate chatbots or other means of directly eliciting information from candidates rather than using static datatesting aspects of job performance, but such interactive assessment tools are beyond the scope of this Article.

267. 29 C.F.R. § 1607.14(B)(2) (2019).

268. § 1607.14(B)(3); *see also* § 1607.14(B)(2) ("[M]easures or criteria [of work behavior or performance] are relevant to the extent that they represent critical or important job duties, work behaviors or work outcomes as developed from the review of job information.").

269. *See* Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11,996, 12,005 (Mar. 2, 1979).

270. *Id.*

271. 42 U.S.C. § 12112(b)(6) (2012).

Given that algorithmic employee selection procedures will have to comply with the ADA no less than Title VII, it would be logical to ensure that criterion standards for purposes of validation have a consistent meaning in both ADA and Title VII cases. Thus, a criterion should be acceptable for purposes of an algorithmic selection procedure if that criterion represents an essential job function, as that term is defined in the ADA.

The Guidelines also provide that noncritical but nevertheless important job duties can also serve as criteria for purposes of establishing the job relatedness of a selection procedure. The question is how an important job function differs from a critical or essential one. Here too, the ADA provides a useful framework. The ADA's job relatedness requirements apply only to criteria that "screen out or tend to screen out an individual with a disability or a class of individuals with disabilities."[272] The ADA does not prohibit an employer from taking important but nonessential job functions into account when designing a selection procedure, but it may not use the ability to perform such functions as a screening device or otherwise apply them in a manner that would effectively bar individuals with disabilities from the position in question.[273] In other words, the ability to perform important, but nonessential, job functions can be a factor—just not an inherently decisive one.

Consistent with this principle, the rule should be that employers can use both essential and important job duties as part of an algorithmic employee selection procedure, but only attributes that strongly correlate with essential job functions can be used in algorithms that act, in form or effect, as screening devices. That is, if a validation study shows that the presence or absence of certain attributes is strongly predictive of a candidate's ability to perform one or more essential functions of a job, then the algorithmic tool can use those attributes to remove candidates from consideration for a position. Important job functions can be used as criteria and serve as target variables and used to score or rank candidates so long as they are not used to screen out candidates altogether. In addition, if the algorithm's target variable represents a composite of multiple criteria, its validity does not rest solely on the proper selection of criteria. Rather, the employer must also assign reasonable weights to the criteria in accordance with their relative importance to the performance of the job in question, as revealed by job analysis.[274]

As in ADA cases, an employer's assessment of which job criteria are essential and important should generally be entitled to deference. The same

272. *Id.*

273. *See* Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. at 12,005.

274. *See* PRINCIPLES, *supra* note 81, at 14 ("If the testing professional combines scores from several criteria into a composite, there should be a rationale to support the rules of combination, and the rules of combination should be described.").

rule should apply to the employer's identification of and assignment of weights to the various job functions that serve as the basis for criteria; a plaintiff should not be able to defeat a finding of job relatedness simply by quibbling about the precise weights the employer chose. As long as the employer demonstrates that the selected criteria and weights are reasonable in light of an adequate job analysis, the employer will have satisfied its burden on criterion selection, and the only remaining question would be whether the test results correlate with those criteria.

In accordance with the evolution of the social science of test validity, the rules governing validation of algorithmic selection procedures should also reflect the need to avoid contamination and reduce construct-irrelevant variance. Even if the chosen job criteria are limited to essential and important job functions, there still may be attributes that correlate with the performance of those functions in the training data simply because those attributes are more prevalent among the demographic groups that predominate in the training data. Here, eliminating differential validity, requiring statistical independence, or both could ensure that predictor-criterion relationships do not tend to unfairly exclude members of protected groups for construct-irrelevant reasons.[275]

Differential validity occurs when a test has substantially greater validity for some tested subgroups than for others.[276] For example, a test that accurately predicts job performance for men but not for women has differential validity. The Gender Shades study showed differential validity for the gender classification systems—the tool predicted gender almost perfectly for light-skinned individuals but was noticeably less accurate for darker

275. Of course, it would be impractical to eliminate all sources of construct-irrelevant variance in the uncontrolled setting of recruitment and hiring. But eliminating differential validation and requiring statistical independence across protected groups would at least help ensure that such variance does not stem from protected group membership itself.

276. *See generally* JOHN W. YOUNG, DIFFERENTIAL VALIDITY, DIFFERENTIAL PREDICTION, AND COLLEGE ADMISSION TESTING: A COMPREHENSIVE REVIEW AND ANALYSIS (2001).

skinned individuals.[277] Differential validity and its cousin, differential prediction,[278] are well-recognized threats to validity in test design.[279]

Two variables are said to be *statistically independent* if knowing the value of one of the variables does not provide any information about the value of the other variable.[280] In the context of employee selection procedures, race would be statistically independent of the outcome of the selection procedure if knowing an individual's race would not help someone ascertain that individual's performance on the selection procedure.

Viewed through a job-relatedness lens, the concepts of differential validation and statistical independence are intertwined; if an attribute is predictive of the criteria only for certain demographic groups, then the attribute will both have differential validity between demographic groups and not be statistically independent from membership in those groups. In theory, an adversarial learning process should allow the algorithm to tune the model's use of those attributes so that they are no longer dependent on the sensitive characteristic. That, in turn, should help ensure that the algorithmic tool is assessing candidates on the basis of characteristics that relate to job

277. *See* BUOLAMWINI & GEBRU, *supra* note 217, at 5–10.

278. Differential validity and differential prediction are both forms of test bias that create differences in the meaning of test results for different subgroups. Christopher M. Berry, *Differential Validity and Differential Prediction of Cognitive Ability Tests: Understanding Test Bias in the Employment Context*, 2 ANN. REV. ORGANIZATIONAL PSYCHOL. & ORGANIZATIONAL BEHAV. 435, 436 ("[T]wo forms of test bias especially relevant to personnel selection [are] differential validity (subgroup differences in test validities) and differential prediction (subgroup differences in test–criterion regression equations)."). The formal difference between differential validity and prediction is that the former refers to differences in correlation coefficients between subgroups, while the latter refers to differences in the regression line equations between subgroups. *See* JOHN W. YOUNG, DIFFERENTIAL VALIDITY, DIFFERENTIAL PREDICTION, AND COLLEGE ADMISSION TESTING: A COMPREHENSIVE REVIEW AND ANALYSIS 4 (2001); YOUNG, *supra* note 276, at 4 ("[D]ifferential validity refers to differences in the magnitude of the correlation coefficients for different groups of test-takers, and differential prediction refers to differences in the best-fitting regression lines or in the standard errors of estimate between groups of examinees."). Less formally, differential validity can be thought of as differences in the *magnitude* of the relationship between predictor and criterion, whereas differential prediction refers to differences in the *nature* of that relationship.

279. *See also* PRINCIPLES, *supra* note 81, at 24 ("[P]redictive bias analysis should be undertaken when there are compelling reasons to question whether a predictor and a criterion are related in a comparable fashion for specific subgroups, given the availability of appropriate data."). *See generally* YOUNG, *supra* note 276, at 4–5 (discussing how differential validity and differential prediction affect validation).

280. In the terminology of information theory, independence means that there is no mutual information between the variables. Peter E. Latham & Yasser Roudi, *Mutual Information*, SCHOLARPEDIA (2009), http://www.scholarpedia.org/article/Mutual_information [https://perma.cc/RJ89-YJHD].

performance and not to membership in a protected group. These techniques are discussed in greater detail below in section IV.C.

In sum, an employer using a algorithmic selection procedure that adversely impacts one or more protected groups would bear the burden of showing (1) that the chosen criteria are representative of essential and important job functions identified through an adequate job analysis; (2) that criteria reflecting nonessential job functions were not used to screen candidates; (3) that the employer assigned reasonable weights to the identified criteria in constructing the selection tool's ultimate target variable; (4) that the output of the selection procedure are correlated with performance on the chosen criteria; and (5) that the employer made reasonable efforts to ensure that predictors and criteria are not contaminated by construct-irrelevant factors that are correlated with protected-class status.

3. Alternative Selection Procedures

Under the longstanding framework codified in the 1991 amendments to Title VII, the third and final stage of the disparate impact analysis is the employee's effort to rebut the employer's showing of job relatedness by demonstrating the existence of a less discriminatory alternative selection procedure.[281] For at least two reasons—one affecting plaintiffs and the other affecting employers—this framework will prove a misfit for algorithmic selection procedures. The challenge for plaintiffs relates to the black box problem: if a deep learning algorithm is particularly opaque or complex, a plaintiff may not be able to gain the level of understanding necessary to mount an effective rebuttal.

From the employers' perspective, the major problem with the current framework is uncertainty surrounding the legal standards. Courts have generally avoided deciding Title VII disparate impact cases at the third stage of the analysis.[282] This has resulted in Title VII jurisprudence that lacks clear standards on how a proffered less discriminatory alternative should be judged, particularly on the key point of how available and effective a proposed alternative selection procedure must be to satisfy the plaintiff's burden. Must the employer provide its algorithm to the plaintiff, who then might attempt to reengineer it to reduce the adverse impact? That prospect will deter many employers from using algorithmic selection tools, perhaps even more so than for prior generations of employee selection procedures.

281. *See* Ricci v. DeStefano, 557 U.S. 557, 578 (2009) (citing 42 U.S.C. §§ 2000e–2(k)(1)(A)(ii), (C) (2012)).

282. *See id.* at 589–91 (rejecting three arguments of an "equally valid, less discriminatory testing alternative" brought by plaintiffs that defendant would necessarily have refused to adopt).

For example, many deep learning algorithms in use today rely on mathematical techniques that are guaranteed only to find a locally optimal model rather than the most accurate and effective model possible. That is, from a given set of initial conditions and parameters, the algorithm makes small adjustments until it reaches a point where further small adjustments will reduce rather than improve the accuracy of the model. Algorithm designers use this approach because performing a comprehensive search for a globally optimal model is computationally complex for even a modestly large data set, and wholly impractical for the high-dimensionality data sets that are available in the age of Big Data.

This has two important consequences. First, the process is not guaranteed to find the globally optimal set of parameters for a particular model. Second, two neural networks using the same data may generate different sets of locally optimal parameters, depending on the starting points specified for the parameters at the beginning of the training process.

Because absolute optimization cannot be guaranteed, there is always a risk that a plaintiff will be able to generate a model with equal or better accuracy that has less of a disparate impact. More generally, it is not feasible for employers to know in advance which machine learning algorithm will be most effective in identifying a globally optimal, nondiscriminatory model, or to test every conceivable type of algorithm to discover which one provides the most accurate predictions. If a plaintiff develops or identifies during litigation a better performing algorithm that employer had not previously considered, it seems reasonable for a court to order the tool to be modified, going forward; but it would be punitive for the court to provide retrospective relief based on an algorithm of which the employer had not previously been aware.

The legal system could use the fact that the outputs of algorithmic tools are the result of the mechanistic application of mathematical optimization techniques to greatly simplify the less discriminatory alternative legal standards. The essential decisions in designing an algorithmic employee selection tool are the selection and weighting of the criteria identified in the job analysis. If the criteria are properly identified and incorporated into a single target variable on which to optimize, the algorithmic procedure will find a model that is at least locally optimal. Consequently, and assuming the employer selected and weighted the target criteria properly,[283] a plaintiff's later identification of a model with equal or better accuracy and less disparate impact should not, by itself, suffice to defeat an employer's assertion of business necessity. Instead, the employee should only be able to prevail in the face of an otherwise-valid selection procedure if the employer actually

283. However, if the criteria were not weighted or selected properly, then the test itself is invalid—i.e., the employer would have failed to establish the business necessity defense, rendering the third stage of the analysis moot.

considered and rejected an alternative algorithm that would have resulted in a reasonable reduction in adverse impact and, to paraphrase *Albemarle Paper*, would still have served the employer's legitimate objectives in selecting well-qualified candidates for a particular position.[284] Otherwise, the discovery of a more optimal and less discriminatory model should only bind the employer prospectively.

To defeat an employer's job relatedness defense, then, a plaintiff should be required to demonstrate that the following:

(1) The employer considered and rejected an alternative modeling methodology;

(2) The alternative modeling methodology would have served the employer's legitimate interests in selecting suitable candidates for a particular position;

(3) The alternative methodology would have resulted in a reasonable improvement in the selection rate for the plaintiff's protected class; and

(4) The alternative modeling methodology would not have unreasonably lowered the employer's ability to select the best candidates for the particular position in comparison to the modeling methodology that the employer ultimately chose.

Note that this test does not require the employer to identify the globally optimal model. Given the impracticality of searching for a global optimum for large high-dimensionality data sets, employers should not be penalized for failing to do so. If there is evidence that an employer, with intent to discriminate, selected or manipulated the initial conditions or consciously chose not to search for a global optimum where it would have been reasonable to do so, then the plaintiff would have a claim for disparate treatment. But an employer should not be subject to disparate impact liability for a valid test constructed using well-established optimization methods simply because the plaintiff chances onto a more accurate and less discriminatory model later.

Components (3) and (4) of this analysis are intertwined and, as with the proposed test for prima facie disparate impact, eschew tests of significance in favor of tests of reasonableness. An improvement in adverse impact is more likely to be reasonable if adopting the alternative methodology would have caused little or no reduction in model performance, and a reduction in model

284. Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975) (citing McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801 (1973)).

performance is more likely to be deemed unreasonable if the alternative modeling methodology would have only slightly improved selection rates for an adversely impacted group. The crux of these final two elements of the employee's rebuttal is demonstrating that the employer's rejection of the alternative modeling methodology was objectively unreasonable.[285]

C. Disparate Treatment

The Supreme Court's decision in *Ricci* greatly curtailed employers' ability to use race-conscious methods to mitigate statistical imbalances and, despite Justice Ginsburg's prediction that the *Ricci* decision would lack staying power,[286] the current composition of the Court makes it unlikely that the decision will be overturned in the foreseeable future. Employers seeking to remove potential biases from algorithmic selection tools must do so while complying with *Ricci's* strictures. Fortunately, employers could use machine learning and statistical techniques to prevent disparate impacts from arising while remaining within the boundaries set by *Ricci*. These methods could serve as safe harbors for employers seeking to correct disparate impacts in algorithmic selection procedures without running afoul of the prohibition against disparate treatment.

One strategy would be to engage in new forms of differential validation; that is, ensuring that a selection procedure has validity not only within the

285. Even if a group of plaintiffs makes this showing, however, a largely unresolved question in the context of disparate impact discrimination is how to calculate damages in cases involving competitive selection processes. Say, for example, that 50 white and 50 Hispanic candidates apply for 10 open positions with a company, and the company, based on an algorithmic selection procedure using facially neutral criteria, selects white candidates for all 10 positions. Even if the Hispanic candidates brought suit and established that the selection procedure violated Title VII's disparate impact prohibition, it would have been impossible for all 50 Hispanic candidates to be hired into just 10 open positions. The appropriate method of calculation of damages in such cases remains an open legal question. Judge Posner suggested in *Doll v. Brown* that courts use the probabilistic, tort-based recovery theory of "loss of a chance" as a basis for calculating damages in discrimination cases involving competitive settings. *See* Doll v. Brown, 75 F.3d 1200, 1205–06 (7th Cir. 1996). While *Doll* was a disparate treatment case, its "loss of a chance" logic has been generalized to disparate impact cases where the number of open positions is less than the number of affected plaintiffs. *See, e.g.*, Howe v. City of Akron, 801 F.3d 718, 752 (6th Cir. 2015) (remanding disparate impact for further proceedings to determine whether loss of a chance is an appropriate method of calculating back pay); Biondo v. Chicago, 382 F.3d 680, 688 (7th Cir. 2004) (approving use of lost chance calculation in a disparate impact case, and stating "[i]f four people competing for one position lost an equal chance to get it, then each should receive 25% of the benefits available"). A full exploration of this issue is beyond the scope of this Article, but the applicability of "loss of a chance" theory in disparate impact cases appears to remain underexplored by courts.

286. *See* Ricci v. DeStefano, 557 U.S. 557, 609 (2009) (Ginsburg, J., dissenting).

dominant demographic groups, but also across all demographic groups. Deep learning gives employers the ability to differentially validate selection procedures without resorting to the blunt instrument of race norming, a procedure outlawed by the Civil Rights Act of 1991 that previously had been the primary method that employers used for differential validation. Another de-biasing technique involves using adversarial learning to ensure that the outputs of a selection procedure are statistically independent from protected class status.

The final segment of this section addresses what often is developers' first instinct when seeking to design an algorithm that avoids disparate impacts: imposing explicit constraints on the model to ensure that selection rates are roughly equal across protected classes. Imposing such constraints sits less easily, however, with the spirit and text of antidiscrimination laws than eliminating differential validity or enforcing statistical independence.

1. Safe Harbor 1: Differential Validation

In the Title VII case law, differential validation—that is, eliminating differential validity from a test[287]—has been conflated with the practice of race norming, which was the most common method that employers used to correct for differential validity between whites and nonwhites.[288] Race norming ordinarily involves using different cutoff scores for members of different subgroups or adjusting test scores so that the highest performing members of one subgroup receive the same final scores as other subgroups. The 1991 amendments to Title VII explicitly prohibited those practices.[289]

But differential validation is not the same as—and need not involve—norming. At its root, differential validation simply means making sure that a procedure can distinguish between higher and lower performers not only in the majority groups but also within and across all demographic groups of interest. Some methods for correcting differential validity, such as the race norming of test scores, may constitute disparate treatment, but that does not mean that differential validation itself is inconsistent with Title VII. On the contrary, ensuring that a selection procedure is useful for all applicants, not just for applicants in certain groups, is precisely the sort of "removal of

287. *See supra* notes 276-279, and accompanying text.

288. LARSON, *supra* note 84, at § 27.12 ("[T]he concept of 'differential validation,' leading to the practice of adding points to or otherwise adjusting the scores of protected group members, has come under attack by scholars.").

289. 42 U.S.C. § 2000e-2(l) (2012).

artificial, arbitrary, and unnecessary barriers to employment" that *Griggs* recognized as the very crux of Title VII.[290]

Any employment test that truly measures job performance should be able to clear that bar. The Guidelines require that subjects of a criterion-related validation study be representative of the relevant labor market precisely because unrepresentative samples can result in a test that does not accurately measure competence for the actual applicant pool.[291] And the Guidelines specifically warn of the need to check for bias and relevance when a criterion results in "significant differences in measures of job performance for different groups."[292]

Differential validation need not involve norming if it is done as part of the test-design process, rather than as a post hoc adjustment to test scores. But incorporating differential validation into the test-design process would have been impractical for written employment tests and other traditional selection procedures. Eliminating differential validity requires complex analyses of how different groups performed on different proposed components of the test to determine which components should be selected and weighted so the test as a whole is comparably accurate within and across groups. That process would have been exceedingly time consuming and costly with traditional examination-based selection procedures. Thus, employers' only practical option for eliminating differential validity was to use the blunt instrument of norming at the back end. When the 1991 amendments to Title VII prohibited that practice, it had the practical effect of eliminating employer efforts to engage in differential validation; only three cases even mentioned the terms *differential validation* or *differential validity* since 1991 and none have done so since 2005.[293]

But in the context of data-driven selection procedures and with the advent of deep learning, the complexity of differential validation is manageable. Using well-established machine learning techniques, an algorithmic tool could be designed to check different combinations of attributes, test them for

290. Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

291. *See* 29 C.F.R. § 1607.14(B)(4) (2019) ("[T]he sample subjects should insofar as feasible be representative of the candidates normally available in the relevant labor market for the job or group of jobs in question, and should insofar as feasible include the races, sexes, and ethnic groups normally available in the relevant job market."); *see also* Albemarle Paper Co., v. Moody, 422 U.S. 405, 435 (1975) (discussing validation study). The Court rejected employer's validation study in part because "Albemarle's validation study dealt only with job-experienced, white workers; but the tests themselves are given to new job applicants, who are younger, largely inexperienced, and in many instances nonwhite." *Id.*

292. 29 C.F.R. § 1607.14(B)(2).

293. The Authors conducted a Westlaw search on January 17, 2019, of all federal cases ("All Federal") and all state cases ("All States") for the terms "differential validation" or "differential validity."

validity within each subgroup, and make minute adjustments to the weights on the components until the model has comparable predictive validity across different protected classes. The resulting model then can serve as the basis for the selection procedure.

For at least two reasons—one temporal and one teleological—this type of differential validation would not run afoul of either the text of Title VII or the Supreme Court's holding in *Ricci*. From a temporal standpoint, § 703(*l*) only prohibits "adjust[ing]" scores, "us[ing] different cutoff scores," or "otherwise alter[ing]" the scores of a test.[294] By its own terms, and consistent with *Ricci's* design versus post-design distinction, this prohibition against norming in § 703(*l*) applies only to selection procedures whose content has already been determined. Because the whole purpose of algorithmic differential validation is to decide on a scoring system in the first instance, there simply are no scores to adjust.

But perhaps more importantly, the differential validation process has as its objective not achieving equal score performance across protected groups, but rather equal predictive performance—that is, an equal ability to distinguish between high- and low-performing future employees within and across protected groups. The resulting model would not necessarily—or even usually—achieve roughly equal selection rates or test performance across protected groups. It would instead ensure that the model does not give undue weight to characteristics that are only associated with good job performance among certain subgroups. This helps the model focus on characteristics that are tied to the underlying job-related constructs, rather than construct-irrelevant attributes that happen to be more prevalent within specific groups. When conducted using deep learning, differential validation thus is both different in time and in kind from the types of protected class-driven adjustments that Title VII's race norming prohibition targets.

United States v. City of Erie provides some legal precedent for this distinction.[295] At issue was the validity of a physical agility test administered to entry-level candidates to the police department.[296] The City had required all candidates to complete seventeen push-ups as part of a broader physical fitness test.[297] One of the expert witnesses for the U.S. Department of Justice, which brought the suit, testified that this test suffered from differential validity because "if a man and a woman obtained the same score on the push-ups test,

294. 42 U.S.C. § 2000e-2(*l*) (2012).

295. *See generally* United States v. City of Erie, 411 F.Supp.2d 524 (W.D. Pa. 2005) (noting that different standards when applied to different genders can function as the same standard in terms of predicted success for a particular job).

296. *Id.* at 524.

297. *Id.* at 532–33.

the woman's predicted job performance would be better than the man's."[298] Another expert testified that in physical fitness tests recognized by the American College of Sports Medicine, "the format for women is typically modified, requiring them to push-up from the knees" rather than with their entire body outstretched.[299]

The City argued that allowing women to pass with a lower number of push-ups would constitute unlawful gender norming, but the Court firmly rejected this argument: "[I]n this circumstance, requiring that men and women complete different numbers of push-ups to pass the test is not 'gender-norming,' and it is not using 'different standards' for males and females. Rather it is using the same standard in terms of predicted success on the job task at issue."[300] Machine learning can similarly be used to adjust model parameters so that the output of an algorithmic tool has equal predictive power among different protected groups.

Modern social science recognizes eliminating construct-irrelevant variance between groups—including variance that arises as a result of differential validation or prediction—as an essential part of the validation process.[301] Differential validation or some other comparable measure of bias control may be especially needed in the context of algorithmic selection tools because of well-recognized problems that machine learning systems encounter when they are used on groups that were underrepresented in the training data. Take the example of Beauty.AI, billed as the world's first AI-judged beauty contest.[302] Similar to the gender classification tool used in the Gender Shades study,[303] the Beauty.AI judge was trained on a data set where darker skinned individuals were underrepresented.[304] As a result, the system was less accurate in rating photographs of nonwhites and ended up picking winners that were mostly white and, to a lesser extent, Asian[305]—groups for which the algorithm was more confident in its rating because of their greater representation in the training data.[306] Given that at least some protected class groups are certain to be underrepresented in any given data set, differential

298. *Id.* at 560.

299. *Id.* at 549.

300. *Id.* at 560.

301. *See* PRINCIPLES, *supra* note 81, at 4, 24.

302. Sam Levin, *A Beauty Contest Was Judged by AI & the Robots Didn't Like Dark Skin*, GUARDIAN (Sept. 8, 2016), https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people [https://perma.cc/9SAC-79EJ].

303. *See supra* text accompanying notes 234–40.

304. Levin, *supra* note 302.

305. *See id.* (noting that out of forty-four winners, nearly all were white, a few were Asian, and only one had dark skin).

306. *See id.* (noting that when a minority group is underrepresented in a data set, "algorithms can reach inaccurate conclusions for those populations").

validation is likely to prove essential for employers seeking to build an unbiased and legally compliant algorithmic selection tool.

Unfortunately, differential validation alone will not eliminate disparate impacts in all cases. Due to inequalities in education and socioeconomic status as well as the effects of prior discrimination, members of different protected groups will sometimes differ in construct-relevant attributes as well. For example, in industries where women have been historically underrepresented, men may have had more opportunities to receive job-relevant training and to gain experience performing high-level tasks. Those attributes may be equally predictive of actual job performance for both men and women, but male applicants would be more likely to have those attributes. That is certainly a type of "artificial, arbitrary, and unnecessary barrier[] to employment"[307] that Title VII was intended to eliminate, but differential validation would be of little assistance in eliminating that barrier.

2. *Safe Harbor 2: Statistical Independence*

The concept of statistical independence[308] supplies a basis upon which employers seeking to implement algorithmic selection procedures may be able to greatly reduce disparate impacts—whether from construct-relevant or construct-irrelevant sources—without running afoul of the prohibition against disparate treatment. Of course, statistical independence could conceivably be achieved through means that do nothing to advance the employer's objective of identifying the best candidates for a particular position; randomly assigning test scores to test-takers would result in independence, but it would be useless as the basis for a selection procedure.

The type of statistical independence that could serve as a benchmark for disparate treatment is conditional independence. Two variables x and y are conditionally independent given a third variable z if, once the value of z is known, knowing the value of y provides no additional information about x (and vice versa). For employee selection procedures, therefore, the relevant inquiry would be whether protected class status and the outcome of the selection procedure are independent given the values of the independent variables ultimately used in the procedure.

As with differential validity, designing a traditional examination-based selection procedure to have such conditional independence would be impractical. But modern machine learning techniques provide a potential path through which statistical independence can be achieved algorithmically. Harrison Edwards and Amos Storkey of the University of Edinburgh demonstrated how this can be accomplished through adversarial learning,

307. Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

308. *See supra* note 280 and accompanying text.

which is a machine learning technique in which a digital "adversary"—essentially, a second training algorithm—is programmed to disrupt the performance of the predictor algorithm in some way.[309]

In Edwards and Storkey's technique, the adversary is fed a representation of the data on a particular candidate and attempts to predict a sensitive attribute such as race or gender.[310] If the adversary's prediction is correct, the predictor algorithm is penalized and the adversary is rewarded.[311] Over many iterations, the predictor algorithm reduces the weights of attributes that carry substantial information about a person's protected class status, while increasing the weight of attributes that correlate well with the target variables that do not reveal information about protected class status.[312] Eventually, these adjustments should result in model outcomes that are independent of the sensitive attribute.[313]

Conditional independence is, in many ways, the antithesis of disparate treatment. If, in the words of *Griggs*, the goal of § 703(a) is to make "race, religion, nationality, and sex become irrelevant,"[314] that is precisely what conditional independence ensures. True, if characteristics relevant to the job constructs are, in fact, unequally distributed between protected groups, statistical independence may reduce the selection procedure's overall predictive accuracy. But for many companies looking to leverage the combined power of Big Data and deep learning, that loss in accuracy would likely be a price worth paying if the law recognized statistical independence as providing employers with a safe harbor from disparate treatment liability while still benefiting from the efficiencies and predictive power of algorithmic selection tools.

3. A Treacherous Harbor: Constrained Optimization

Correcting differential validity and achieving statistical independence are indirect methods of preventing or correcting disparate impacts because they do not alter selection rates in and of themselves. A more direct—but possibly unlawful—approach to correcting disparate impacts would be constrained

309. Harrison Edwards & Amos Storkey, *Censoring Representations with an Adversary* (Int'l Conf. on Learning Representations Conference Paper) (Mar. 4, 2016), https://arxiv.org/pdf/1511.05897.pdf [https://perma.cc/4X8H-ED8M].

310. *See id.* at 1, 5.

311. *See id.* at 1–2.

312. *See id.*

313. Edwards and Storkey use the term "fairness" to describe this statistical independence. *See id.* at 1 ("Here, fairness means that the decision is not-dependent on (i.e. marginally independent of) the sensitive variable.").

314. Griggs v. Duke Power Co., 401 U.S. 424, 436 (1971).

optimization. In this context, constrained optimization means finding a model that maximizes predictive accuracy (optimization) but limits the search space by requiring the model to satisfy certain conditions (constraints).[315] In the context of an algorithm designed to avoid disparate impact liability, the constraints could be catered to the rules governing a prima facie case of disparate impact. Thus, under the Guidelines' four-fifths rule, the algorithm could find an optimal model subject to the constraint that the selection rate for each protected group can be no lower than 80% of the selection rate of any other group in the same protected category.

Enticingly, *Ricci* rejected the proposition that employers may not take disparate impact into account—even where this means being race conscious—when designing a selection procedure to ensure that the procedure provides a fair chance for all individuals.[316] *Ricci* thus suggests that there is a distinction between designing a selection procedure in a way that checks for and mitigates bias on one hand, and post-design test score adjustments and conscious decisions to incorporate protected class preferences into a model on the other.[317] Technically, using constrained optimization during the design phase would be consistent with this principle.

That said, the fact that constrained optimization explicitly examines and makes adjustments based on the selection rates of different groups distinguishes it from the approaches geared toward achieving differential validation and statistical independence. Differential validation aims to ensure that the metrics have comparable accuracy across protected groups, and statistical independence ensures that the selection tool is not encoding protected class information as part of its model. Any effect on selection rates is a beneficial side effect of these techniques, rather than their objective. A constrained optimization approach, by contrast, makes equalizing group selection rates a direct and explicit goal.

If Title VII could be viewed through a purely antisubordination lens, such a direct approach would be unproblematic. But it runs contrary to the law's anticlassification strictures. True, an employer may not know beforehand which precise groups will see their selection rates improve with constrained

315. *See* Introduction to Constrained Optimization in the Wolfram Language, Wolfram Language & System Documentation Center, https://reference.wolfram.com/language/tutorial/ConstrainedOptimizationIntroduction.html (defining constrained optimization as a problem for which a function must be minimized or maximized subject to constraints).

316. Ricci v. DeStefano, 557 U.S. 557, 585 (2009) ("Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race."). Because design-stage modifications were not at issue in *Ricci* and the Court's articulation of the design versus postdesign distinction was not essential to its holding, this statement from *Ricci* is arguably dictum, but at least one court has relied on it as precedent. *See* Maraschiello v. City of Buffalo Police Dep't, 709 F.3d 87, 95 (2d Cir. 2013).

317. *See Ricci*, 557 U.S. at 584–85.

optimization as compared to an unconstrained modeling approach. But that logic would not save an employer that normalizes test scores after administration of a traditional employment test; § 703(*l*)'s prohibition against norming prohibits all score adjustments made "on the basis of" a protected classification without regard to whether the employer knew in advance which groups would benefit from such norming. Even at the design stage, making adjustments explicitly based on protected characteristics sits uneasily with § 703(a)'s broad prohibition against employment decisions made because of such characteristics.

The Authors would encourage courts and agencies to adopt *Ricci's* design versus post-design distinction, thus giving employers the freedom to fashion unbiased algorithmic selection tools without risking disparate treatment liability. Indeed, the only published case interpreting this passage from *Ricci* relied on this distinction to affirm a grant of summary judgment in favor of an employer that may have been "motivated in part by its desire to achieve more racially balanced results" when it adopted a new employment test.[318] This rule is fully consistent with the objectives of Title VII, as elucidated in *Griggs* and *Albemarle Paper*. But a rule permitting constrained optimization will likely face greater resistance than one permitting employers to use more sophisticated machine learning approaches that avoid such explicit reliance on protected class status.

V. CONCLUSION

For now, the above proposal is just that—a proposal, albeit one strongly rooted in the text of Title VII and the case law interpreting it. At this point, employers considering implementing AI-powered recruitment and hiring at scale simply do not know how a court would analyze an algorithmic selection procedure under Title VII. That is one reason the EEOC should act quickly to clarify the legal standards by which it will assess algorithmic selection procedures. Employers will undoubtedly be wary of developing (or at least implementing) such procedures in the meantime.

The framework offers two routes by which employers can avoid liability for the inevitable adverse impacts that algorithmic selection tools will generate: (1) correcting any disparate impacts by using one of the disparate treatment safe harbors; or (2) satisfying the business necessity defense by conducting a proper job analysis followed by criterion validation. That presents a conundrum for employers wishing to use algorithmic selection procedures today. The most efficient and practical way to achieve legal compliance under the above proposal—using one of the proposed disparate

318. *Maraschiello*, 709 F.3d at 95.

treatment safe harbors—is the path that carries the greatest legal uncertainty because using algorithms that use machine learning to eliminate differential validation or achieve statistical independence have never been tested in court. Conversely, the path to compliance that would provide the greatest legal certainty—following the validation standards described in the framework, which adhere closely to the Guidelines and existing case law—may be neither efficient nor practical, and may prove to be a wasted effort in any event if the Guidelines receive a long-overdue overhaul to bring them in line with the modern social science.

The difficulty of validation is partially just a function of the difficulty of validating large high-dimensionality data sets. But perhaps even more fundamentally, employers may find it extremely difficult to build a sufficiently representative set of measurable job behaviors and outcomes to serve as a proper set of validation criteria. Traditional employee selection procedures were actual examinations whose content an employer could cater to the actual skills and knowledge relevant to job performance. Algorithmic selection tools, by contrast, generally rely primarily on passive analysis of historical (and therefore static) data, which often cannot easily be crafted to fit the job functions of a particular position. Many employers have only very general or subjective measures of job performance available, such as tenure or impressionistic supervisor ratings, which courts have historically disfavored for purposes of validation when examining an employer's proposed business necessity defense.

These factors, coupled with the availability of data sets containing hundreds or thousands of attributes, will make it increasingly difficult for employers to validate employee selection tools in accordance with the Guidelines. That underscores the need for policymakers and courts to both adopt new standards for validation and establish clear safe harbors that allow employers to prevent disparate impacts from arising without exposing themselves to disparate treatment liability. Modifying the traditional framework by eliminating tests of statistical significance and replacing them with reasonableness standards is also necessary to avoid missing opportunities to materially improve the diversity and inclusiveness of an employer's workforce with minimal sacrifice in quality.

Companies today are leveraging algorithmic tools powered by machine learning and built on Big Data to enhance every aspect of their business activities. Despite the fact that algorithmic tools offer employers a vehicle for more effective and inclusive HR selection decisions, with less discrimination and more long-term accountability, the use of such tools to improve recruitment, hiring, and other human resources decisions has lagged behind their use in other business operations. The levee eventually must break, and legal standards will have to evolve quickly to stem the tide. For now, courts,

agencies, and employers alike must be attuned to the growing mismatch between the state of technology and existing legal standards so that the promise of these technologies is not squandered.