RESEARCH BRIEF 18 • JANUARY 2021

Cognitive Science as a New People Science for the Future of Work

Frida Polli, CEO and Founder, pymetrics

Sara Kassir, Senior Policy and Research Analyst, pymetrics

Jackson Dolphin, Data Science Research Associate, pymetrics

Lewis Baker, Director of Data Science, pymetrics

John Gabrieli, Grover Hermann Professor, Health Sciences and Technology, Department of Brain and Cognitive Sciences

Director, Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, MIT

Member, MIT Task Force on the Work of the Future

Cognitive Science as a New People Science for the Future of Work

Frida Polli, Sara Kassir, Jackson Dolphin, Lewis Baker, John Gabrieli

"When we measure something, we are forcing an undetermined, undefined world to assume an experimental value. We are not measuring the world, we are creating it."

– *Niels Bohr, Danish physicist and winner of the 1922 Nobel Prize in Physics*

Introduction

The notion of studying people in jobs as a science—in fields such as human resource management, people analytics, and industrial-organizational psychology—dates back to at least the early 20th century. In 1919, Yale psychologist Henry Charles Link wrote, "The application of science to the problem of employment is just beginning to receive serious attention," at last providing an alternative to the "hire and fire" methods of 19th-century employers. A year later, prominent organizational theorists Ordway Teal and Henry C. Metcalf claimed, "The new focus in administration is to be the human element. The new center of attention and solicitude is the individual person, the worker." The overall conclusion at the time was that various social and psychological factors governed differences in employee productivity and satisfaction.

In some ways, the basics of modern people science remain closely aligned with the tenets first established more than 100 years ago. Namely, around the turn of the 20th century, psychologists became particularly focused on studying constructs that measured both group and individual differences, devising tests to measure them in people, and demonstrating correlations between tests (i.e., predictors) and metrics of job success (i.e., criteria). With respect to individual differences, psychologist E.L. Thorndike notably explained the concept in 1918: "We may study a human being in respect to his common humanity, or in respect to his individuality. In other words, we may study the features of intellect and character which are common to all men, or we may study the differences in intellect and character which distinguish individual men." By the 1920s, there was a basic consensus that the scientific method could facilitate employment selection if a measurement tool could clearly demonstrate a relationship with worker efficiency.

But two primary factors have changed significantly since the establishment of the first employment selection tools: the needs of employers and the needs of society. Because hiring assessments must be developed with

a particular set of priorities and circumstances in mind, they tend to become obsolete in the face of dramatic social, cultural, and economic shifts. Consider, for example, the following questions: "In the year 1900, what does the industrial factory worker need to be able to do well?"; "In the year 1950, what does the car mechanic need to be able to do well?"; and "In the year 2000, what does the truck driver need to be able to do well?" All have very different answers, meaning an assessment developed with one context in mind will be less useful for others. Notably, this idea is not unique to people science: Philosopher of science Thomas Kuhn famously coined the term *paradigm shift* in 1962 to describe a fundamental shift in the underlying assumptions governing a field. To quote an adage often used to explain the spirit of his work: "The answer you get depends on the question you ask."

The nature of jobs, firms, workers, and society has transformed in innumerable ways over the past few decades; to understand the shortcomings of traditional approaches to people science, it is crucial to identify the aspects of today's people science paradigm that were absent in earlier iterations. At a high level, these can be summarized as four considerations. First, while the concept of employee satisfaction was fairly novel before the 1930s, with the increased competitiveness of labor markets, modern employers view ***job fit*** as critical to reducing employee turnover. Second, particularly since the widespread adoption of computers, today's employers have fewer needs for skills like rote memorization or task repetition, instead emphasizing the importance of ***soft skills*** (also known as *aptitudes*) in the workplace. Third, contemporary organizations are legally required to consider the ***fairness*** of their hiring strategies, and are socially pressured to prioritize demographic diversity. Fourth, in light of the potential for modern technology to both create and eliminate new types of jobs, modern employers seek more ***flexible*** approaches to evaluating talent than did their predecessors.

Practitioners of traditional approaches to employment selection have undertaken a variety of efforts to better account for the 21st century's talent needs. The simple reality is that significant room for improvement remains, highlighting the need for a fundamental rethinking of people science strategies. Fortunately, entirely new areas of science dedicated to studying human brains, behaviors, and thought processes—fields such as cognitive science, neuropsychology, cognitive psychology, and behavioral neuroscience—have emerged since employment selection first became a research discipline. These advancements allow for the evaluation of job candidates in a manner that is individualized, nuanced, equitable, and dynamic. The result can be massive benefits to the efficiency of employers, the well-being of employees, and the cohesion of society.

Regarding terminology, this brief will often make a distinction between *traditional people science* and *new people science*. Though not formal terms, the goal here is to differentiate between methods that rely on data inputs that are heavily correlated with demographic identity and social position and methods that incorporate modern technology for evaluating human potential. *Traditional people science* therefore encompasses tools such as résumés and CVs, standardized educational tests, IQ (or general mental ability)

tests, and personality inventories based on self-report. *The new people science* refers to the evaluation of behavioral data collected with digital assessments, specifically to measure the underlying cognitive, social, and emotional traits of individuals without self-reports. Best practices for traditional people science are largely captured by the professional standards put forth by Division 14 of the American Psychological Association, also known as the Society for Industrial and Organizational Psychology (SIOP). Best practices for the new people science, on the other hand, are only beginning to coalesce as insights from behavioral and neuroscience labs are applied at scale in the context of employment selection.

This Brief Proceeds in Five Sections:

- First, we review the limitations of traditional approaches to people science. In particular, we focus on four needs of the modern employer that are not satisfied by the status quo: job fit, soft skills, fairness, and flexibility.
- Second, we present the foundations of a new people science by explaining how advancements in fields like cognitive science and neuroscience can be used to understand the individual differences between humans.
- Third, we describe four best practices that should govern the application of the new people science theories to real-world employment contexts.
- Fourth, we present a case study of how one platform company has used the new people science to create hiring models for five high-growth roles.
- Finally, we explain how the type of insights presented in Section IV can be made actionable in the context of retraining employees for the future of work.

I. Limitations of Traditional Approaches to People Science

The first section of this brief aims to explain how the employment paradigm has shifted in recent years to incorporate dimensions of **job fit, soft skills, fairness, and flexibility**, and how conventional approaches to talent selection have failed to keep pace. In detailing the shortcomings of traditional employment tools, which may include résumé reviews, manual sorting procedures, personality inventories, intelligence tests (e.g., *IQ*, *general mental ability*, or *cognitive assessments*), it is clear that modern employers' needs demand a new employment science.

SOME HISTORICAL CONTEXT

Before beginning this review, it is worth emphasizing that the constancy of traditional people science is no coincidence. Many of the deficiencies of traditional hiring tools—such as inconsistent results across different job types—have been recognized essentially since their invention. However, employment scientists diverted attention away from correcting such shortcomings in the 1980s, following the broad acceptance of a few key theories within mainstream people science. Around this time, American psychologists John E. Hunter and Frank L. Schmidt used a novel meta-analytical method to disrupt the prevailing belief that the drivers of success in a job vary depending on the role (e.g., a focus on job fit). Instead, they argued that an unambiguous relationship between job performance and general mental ability (GMA) held across all contexts; they claimed the strong correlation had simply been masked by "noise" in data that could be removed with statistical corrections (which notably deviated from standard meta-analysis practices). Nearly 40 years later, Hunter and Schmidt's theory of *validity generalization* continues to be cited as evidence that cognitive ability is the single best predictor of job success.

It would be difficult to overstate how much Hunter and Schmidt's conclusions changed the nature of employment science. Alegra et al. (1984) summarize the prevailing theory of *situational specificity* that had existed prior to their research: "Especially from the studies of Ghiselli (1959, 1966, 1973), we know that for a specific type of test applied to a particular type of job the validity coefficients from different samples show considerable variation: in one sample the validity coefficient many be high and positive and in another it may even be negative." As Kozlowski writes in *The Oxford Handbook of Organizational Psychology, Volume 1* (2012), "Validity generalization ended the situational specificity hypothesis to the point that professional guidelines such as the SIOP (Society for Industrial and Organizational Psychology) Principles now explicitly recognize the role of meta-analysis as a validation strategy."

While Hunter and Schmidt's conclusions had a dramatic effect on the trajectory of people science, the most remarkable thing about their research is how uncritically it was accepted by other people scientists. Meta-analyses are known to involve many judgment calls on the part of the authors (e.g., which studies to include, how to approximate sample size variance, how to categorize moderators, how to present results, etc.), with different assumptions inevitably yielding different results. As Richardson and Norgate (2015) note, while traditional people scientists commonly attribute the predictive validity of GMA to a "large and compelling literature," this conviction "seems to have developed from a relatively small number of meta-analyses over a cumulative trail of secondary citations." In fact, among those researchers who have endeavored to replicate Hunter and Schmidt's analyses with slight adjustments, many have found that the top-line conclusions regarding cognitive ability are easily undermined.

It is beyond the scope of this brief to exhaustively review every response to Hunter and Schmidt's work, but some aspects of their research particularly cast doubt on the idea that IQ tests are effective tools for

modern employers. For example, Hunter and Schmidt strongly advocate that job performance should be conceptualized as a single measure, typically in the form of supervisor ratings or an average of multiple criteria. This is no small assumption in the modern world of work, where employers can have drastically different priorities and expectations of employees. According to one study by Murphy and Shiarella (1997), the validity of a given selection device can vary significantly depending on how different components of job performance are weighted to calculate an employee's overall score. Without knowing the actual organizational goals of the employer then, it is impossible to support the claim that GMA (or any single talent selection procedure) can universally predict success.

Finally, the nature of Hunter and Schmidt's meta-analytical procedure *should* have been of little consequence to the practice of employment selection, because a corrected validity coefficient is meant to "furnish a theoretical estimate of what the effect size might be if everything in the situation was perfect." However, the distinction between hypothetical approximations and real-world correlations has been lost by many traditional employment scientists; as industrial and organizational (I-O) psychologist Murphy (2013) notes, "Given our awareness of the potential influence of meta-analyses on personnel selection practice and on public policies in this area, we have been surprisingly casual about how our results are presented." One consequence of this conflation is the presentation of GMA as the "final answer" for employment selection, explaining why the science has remained fixed on the construct for the last several decades. As this section will detail, while some researchers have made efforts to deviate from Hunter and Schmidt's mainstream ideology, progress has been limited by the absence of a true paradigm shift.

JOB FIT

The concept of job-matching is rooted in two ideas that may feel obvious in today's labor market: that jobs vary by organization, and that employers should care about workers' happiness. Researchers have found that people who are in roles that align with their personality, preferences, and skills are more satisfied in their roles, and job satisfaction is further related to job performance. Additional studies have indicated that employee satisfaction is a driver of important business goals such as organizational effectiveness, profit, innovation, and customer satisfaction. The basic rationale holds that content employees are more likely to demonstrate work ethic and remain with the organization, reducing costs due to turnover. Surveys also have shown that workers are often willing to take pay cuts to accept jobs that they believe better suit them. It is therefore unsurprising that the modern employer has largely come to view matching people to roles that fit them well as a crucial part of the hiring process.

Traditional hiring assessments were not optimized for dimensions of job fit because, at the time of their development, employers faced minimal incentives to be concerned with employees' well-being. Instead, early selection tools seemed to take the position that employers were in positions of absolute control over job candidates; they simply had to identify the "best" candidates to perform the work productively, with

little regard for consequences like job-switching. In the context of traditional assessments, "best" is a notably monolithic concept; certain traits, such as intelligence and conscientiousness, are deemed universally preferable and are assumed to drive job performance across all contexts. An influential challenge to this perspective came from developmental psychologist Howard Gardner in 1983, who presented eight (now, nine)[1] variants of intelligence. Today, the idea that human potential should be evaluated with a multifaceted approach is captured by the concept of neurodiversity, or the idea that variations in mental functions—such as sociability, learning, attention, and mood—are adaptive and nonpathological. The failure to account for neurodiversity in hiring has considerable implications on fairness, which will be discussed later in this section.

The ability of employers to largely ignore the satisfaction of their workers dramatically changed with the advent of labor unions. Mayo (1923) first introduced the concept of emotion into mainstream American I-O psychology with his discussion of factory work as fomenting anger, fear, suspicion, lowered performance, and increased illness, which subsequently created the conditions for worker unrest. In particular, after WWII, businesses realized their error in judgment. As one manager of a nonunion company noted in 1951: "Unions would have far less control and influence in this country if industry had been listening, and once having developed the art of listening, reacted as it should have to what it heard." Jacoby (2008) cites one 1949 report from the Research Institute of America: "The whole question of efficiency and productivity boils down to one thing: understanding the MOTIVATIONS of your employees and taking steps to SATISFY them." Around this time, attitudinal employee research became significantly more commonplace. Subsequently, psychologists in the 1960s and 1970s became more interested in how variations in leadership styles and organizational structures affected employee satisfaction; by the 1980s, researchers began exploring employee personality as a mediating factor.

In its modern conception, the idea of *job fit* has two main variations: *person-organization* (P-O) and *person-job* (P-J). While the former concept is rooted in the notion that behaviors are a product of one's situation, the latter relies on an individual's innate traits as an explanation. During the 20th century, psychologists tended to position P-J and P-O as alternative frameworks for understanding individual variations in job performance. Still, it has become increasingly apparent in recent years that both must be considered in the modern workforce. For example, researchers have found that P-J fit can yield benefits such as reduced employee stress and increased job satisfaction, and that P-O fit is important for maintaining a flexible and committed workforce in a competitive labor market.

The theoretical ideas behind job fit are relatively straightforward, but the prospect of measuring these constructs in the context of employment selection has historically been less so, particularly when subject to the limitations of traditional employment tools. In operationalizing evaluations of person-job fit, an employer must rely on a formal job analysis to determine the *knowledge, skills, abilities, and other characteristics* (KSAOs) required to complete a role. From there, they must identify congruous measures of

an applicant's KSAOs, using information such as prior work experience, résumé data, standardized tests, and reference checks. Given that the majority of the data used to evaluate P-J fit is retrospective (e.g., work history), this approach can overlook candidates who have simply never had the opportunity to demonstrate a particular competency. Additionally, while job analysis is an essential tenet of management science, the procedure largely ignores personality-related characteristics and team dynamics.

In operationalizing person-organization fit, Kristof (1996) suggests that the "constructs of interest are often values, goals, climate, or culture—variables that are most frequently measured by perceptions. Therefore, the aggregation of individual perceptions should be used in the measurement of actual P-O fit." The underlying assumption with this approach is that the majority's stated perception of an organization's culture (e.g., collected through surveys or interviews) effectively represents the organization's culture. This view clearly does not account for any social pressure that employees may feel to regurgitate an employer's stated mission, even if it does not align with reality. Additionally, Cable and Judge (1995) find that hiring outcomes can be predicted based only on an interviewer's perceptions of a candidate's values, even when they do not align with a candidate's self-reported values.

SOFT SKILLS

In the words of one economist: "It is worth stressing that 'soft skills' represents a term that is generally known and understood, but yet not precisely defined." The term was first coined by researchers studying leadership for the U.S. Army in 1972 as "important job-related skills that involve little or no interaction with machines and whose applications on the job are quite generalizable." Heckman and Kautz (2012) describe soft skills as traits such as personality, goals, motivations, and preferences that are valued in the labor market, but not adequately captured in achievement tests. Cimatti (2016) writes that "the term soft skills is used to indicate all the competences that are not directly connected to a specific task; they are necessary in any position as they mainly refer to the relationships with other people involved in the organization. Hard skills, on the other hand, indicate the specific capabilities to perform a particular job." More recently, some thought leaders on the future of work—like President of Dartmouth College Philip J. Hanlon and HR analyst Josh Bersin—insist that the term should be updated to *power skills* to reflect its universal importance. Here, we use the term *soft skills* to refer to cognitive and noncognitive characteristics that tend to be demonstrated across disparate environments and contexts.[2]

While some ambiguity regarding the definition of soft skills persists, the broad consensus is that competencies like communication, teamwork, and people skills are crucial for a successful workforce. According to a 2019 survey of recruiters by LinkedIn, 89% of failed hires lack soft skills. Another 2019 report by the Society for Human Resource Management indicates that 3 out of 4 employers are having difficulty identifying recent college graduates with the soft skills their companies need. Evidence also indicates that such competencies are only growing in importance: Deming (2017) finds that between 1980

and 2012, "social-skill intensive" occupations grew by nearly 12 percentage points as a share of all U.S. jobs and that wages grew more rapidly for these occupations than for any others over the same period.

Intuitively, soft skills are closely related to the idea of job fit—people with a certain personality type or cognitive style will naturally flourish in certain environments more than others will. These underlying traits are largely not accounted for in traditional hiring assessments that seek to place all job candidates on a single spectrum of employability, such as tests of general mental ability (GMA) or IQ. Much of the appeal of GMA tests for employment selection is based on the assumption that they can be applied to virtually any hiring situation, with a higher score almost always being deemed preferable than a lower score. (There are a few notable exceptions, such as police officers in some municipalities in the United States.) However, since the 1980s, newer disciplines like cognitive neuroscience and neuropsychology have indicated that IQ and cognition are not unitary concepts, but rather are comprised of many subcomponents such as verbal ability, visuo-spatial abilities, memory, attention, executive control, task switching, and planning, to name but a few. Most noncognitive traits must also be evaluated in a more context-specific manner; according to Tett et al. (1991), "In the absence of conceptual analyses or personality-oriented job analyses, it is difficult, if not impossible, to determine the extent to which a given personality dimension is relevant to work performance."

Beyond the fact that traditional assessments of soft skills require a specific type of job analysis, there are a variety of limitations around the measurement of these traits. In the typical hiring process, soft skills can only be evaluated through situational judgment tests (SJTs), behavioral simulations like interviews, or self-report instruments, but each of these tools suffers from validity problems. SJTs are defined by Cabrera and Nguyen (2001) as "assessments designed to measure judgment in work settings" that "present the respondent with a situation and a list of possible responses to the situation." However, researchers have indicated that the relationship between a candidate's SJT response and their real-world behavior can vary across different personality types. For example, Slaughter et al. (2014) find that situations designed to test interpersonal skills are less strongly correlated with job performance in people with higher levels of anger hostility (AH). Behavioral interviews are an extremely common form of hiring assessment, but evidence indicates that they generally fail to measure constructs of interest like integrity and customer service orientation. One recent study found that technical interviews for software engineering positions actually measure anxiety, not technical skills. Regarding self-report assessments of soft skills, these tools are affected by a variety of biases, which will be discussed later in this brief. One high-level critique worth noting is that they rely on an individual's ability to accurately understand one's own personality; as Vazire and Carlson (2010) find, "Self-knowledge exists but leaves something to be desired."

FAIRNESS

The importance of fairness in the modern hiring process can be described in two parts: first, the legal requirement to not discriminate against job candidates; and second, the societal desire to promote diversity in the workforce. Students of U.S. history would not be surprised by the fact that traditional hiring assessments were not designed to address either of these dimensions; but today, their strategic significance for employers is clear. Regarding discrimination, between 2010 and 2019, the U.S. Equal Employment Opportunity Commission received nearly 900,000 individual charges of employer impropriety. In addition to the direct financial consequences of litigation, unfair hiring practices yield indirect costs by creating a homogeneous workforce. Researchers have demonstrated that diversity and inclusion in the workplace can drive business outcomes like revenue, innovation, and profit. Konradt et al. (2016) also find that job applicants' perceptions of the fairness of a hiring process can affect both their odds of accepting an offer and their job performance 18 months later.

When traditional employment processes were first developed, the simple reality was that organizations were primarily concerned with evaluating a homogeneous population for jobs. But as GMA tests grew in popularity with employers, so too did problematic evidence indicating that scores were strongly correlated with demographic features, such as educational attainment and socioeconomic status. For example, in 1932, Black psychologist Robert P. Daniel wrote that efforts to use GMAs to measure racial differences in intelligence were "worthless" because "present techniques give measures of differences due to weaknesses in educational opportunities rather than of differences in mental ability." However, despite the strong evidence of racial and class bias in IQ tests, 20th-century I-O psychologists undertook significant efforts to demonstrate the clear relationship between these assessments and job performance. These efforts culminated in the previously mentioned meta-analysis conducted by American psychologists John E. Hunter and Frank L. Schmidt in the 1980s, which led to the conclusion among many in the I-O field that GMAs are the strongest predictors of job performance. Traditional practitioners remain wedded to the use of general cognitive assessments because of this body of work, despite the fact that the racial bias yielded by such tests is significantly more severe than that yielded by any other selection tool. According to one estimate, a GMA test that selects 50% of white candidates will only select 16% of Black candidates from the same applicant pool.

It is important to underscore the fact that efforts to place humans on a single spectrum of cognitive ability are not unique to employment selection: Theories about IQ and its applications have existed in virtually every social science, as have their critics. To fully understand the attachment traditional people scientists developed with GMA tests though, it is important to note how hiring procedures are regulated in the United States. Since the Civil Rights Act of 1964, U.S. employers have effectively been prohibited from engaging in two types of hiring discrimination: first, refusing to employ a person on the basis of race, ethnicity, gender, religion, or national origin; and second, evaluating or limiting applications in a way that

would adversely impact a person's employment prospects due to their race, ethnicity, gender, religion, or national origin. In legal terms, these actions are defined as *disparate treatment* and *disparate impact*, respectively. The former of these concepts addresses what might be otherwise termed direct instances of discrimination (e.g., "I do not want to hire this person because they are Black"), while the latter refers to indirect discrimination (e.g., the hiring process systematically disadvantages Black people, whether intentionally or unintentionally). Technically, the racial disparities in cognitive assessment scores should render such tools illegal since they yield disparate impact.

However, a significant loophole exists in the regulation of hiring procedures, which allows for a biased assessment to be used so long as it aligns with the employer's business necessity. For example, if an employer needs to hire candidates to move heavy boxes, a strength test might be legally permissible, even if it would lead to the disproportionate selection of men over women. In establishing the strong relationship between GMAs and job performance across a variety of professions, traditional people science provided fodder for employers who want to use these tools without concern for the bias they yield against racial minorities. The implied rationale here is that IQ tests are so predictive of job success that an employer cannot afford to sacrifice the certainty of a good hire for concerns regarding racial equity—giving them a business necessity defense. Hunter and Schmidt (1982) attempt to make this case in explicit financial terms with statements such as: "For an organization as small as the Philadelphia police department (5,000 persons), the labor savings stemming from the use of a cognitive ability test to select officers has been calculated to be $18 million for each year's hires." In addition to the ethically questionable prioritization of business profitability over broader societal goals, this position also fails to account for the economic benefits of workforce diversity, as described above.

In more recent years, some I-O psychologists have sought to reduce the bias that occurs when GMAs are used in employment selection, typically not by abandoning such tools, but by suggesting additional metrics to add to the evaluation process. Ployhart and Holtz (2008) conduct a review of 16 of these methods for grappling with what they term "the diversity-validity dilemma," though they find only one approach to be effective. Where traditional employment scientists do advocate for alternatives to cognitive tests, options like the Big 5 personality inventory are common. But these, too, are not immune from fairness concerns in practice: Issues arise because of a desire by employers to place all applicants on a single spectrum in terms of quality. When provided with measures of a candidate's conscientiousness, agreeableness, openness, extroversion, and neuroticism, the suggestion is that high scorers on the first two of these traits make the best employees across most contexts. In contrast, candidates with high neuroticism scores—such as many women and people with mood disorders—are generally deemed less desirable. The consequence of this one-size-fits-all approach to evaluation, whether driven by GMAs or personality tests, is that some "types" of people are effectively the "winners" of the hiring process while other "types" are the "losers" across all contexts.

Formal assessments are, of course, far from the only part of the conventional employment selection process that yields discrimination against demographic groups. With the average résumé review lasting only seven seconds, human recruiters rely heavily on intrinsically biased intuitions, including personal prejudices and social stereotypes, to make rapid judgments about candidates. These effects are well documented by call-back studies, which involve submitting two identical résumés to an employer, with the only change being the applicant's gender or racial identity, as signaled by their name. Using a meta-analysis of these experiments conducted over 30 years, Quillian et al. (2017) find that the average "white" résumé receives 36% more invitations to interview than "Black" résumés, and 24% more than "Hispanic" résumés. Unfortunately, researchers have found that mitigation efforts, such as anonymizing résumés and implicit bias training, are largely ineffective.

FLEXIBILITY

The proliferation of artificial intelligence has intensified conversations regarding workforce flexibility in recent years, but this is far from the first point in history when workers needed to adapt. A 1963 report by the U.S. Department of Labor notes that "occupational and industrial changes have been taking place which have increased the reemployment problems of displaced workers," including "the long-term shift away from the output of goods and toward more services." In the context of the spread of computers, Magrass and Upchurch (1988) write that "new technologies alter the forms of knowledge and productivity that are important to society." More recently, a 2019 article from Forrester, titled "The Future of Work Is an Adaptive Workforce," advises business leaders, "The future of work involves human employees working side by side with robots, intelligent machines from AI, automation, and robotics." According to a survey conducted by major labor law firm Seyfarth Shaw, 72% of employers believe that the future of work will reshape their workforce in the next five years.

Historical efforts to place people into novel types of jobs have followed a basic template. The public sector provides funds for displaced workers to receive formal instruction in skills that the generic "modern" employer needs. The relevant metrics for success are whether the trainees find a job and whether they are making wages equivalent to their previous role. The details of these programs have been updated periodically throughout the mid-20th century via acts such as the Manpower Development and Training Act (1962), the Job Training Partnership Act (1982), and the Workforce Investment Act (1998). However, program evaluations have generally produced pessimistic conclusions about their efficacy. According to Kodrzycki (1997), "Research on existing training programs—such as courses to provide displaced workers with specific occupational skills or advances in general knowledge—fails to show that they enable workers to achieve higher pay at their new jobs." More recently, Muhlhausen (2017) writes, "On Election Day, November 8, 2016...the U.S. Department of Labor slyly released a major experimental impact evaluation that found the federal government's primary job-training programs to be ineffective."

Of course, it is altogether unsurprising that the traditional approach to workforce flexibility falls short because the strategy ignores two important factors: the uniqueness of each worker and the specific needs and evaluation processes of employers. Regarding the former, Leigh (1990) plainly summarizes his survey of existing research on reskilling: "This conclusion is simply that training curriculums offered must match the interests and backgrounds of targeted workers to be effective." Regarding the latter, Fadulu (2018) writes that "federal policy has consistently failed at training" because "it's paid little attention to employers and the question of how they can change to better recruit and retain employees."

In acknowledging that a one-size-fits-all approach to retraining does not yield results, experts have increasingly called for strategies that prioritize the alignment of certain workers with certain opportunities, particularly via job search assistance. For example, in 2002, the Toledo Dislocated Worker Consortium sought to develop a methodology to compare dislocated workers' knowledge and skills with the knowledge and skills required in the occupations related to the training offered. However, this approach effectively assumed that an individual's previous job was the best reflection of what their future job should be. Other efforts have focused on the importance of interest inventories in both directing displaced workers to new roles and unhappy employees seeking an internal change. Nye et al. (2012) find that, while interests are one part of job performance, they are less relevant when an individual's person-environment fit (i.e., their personality, priorities, and motivations) is not also taken into account.

Efforts to account for the specific needs of employers in reskilling employees have largely been couched as "market-oriented training systems." A 1998 report produced by the International Labour Office notes that "alliances between the interested parties have become the key strategy to improve the relevance, efficiency, effectiveness, equity and sustainability of training policies and systems." Lee (2009) describes such alliances as efforts to "promote better matching between supply and demand in the labor market" by ensuring agreement on skill demands and infrastructural capacity to meet them. While it is logical that clearer information from industry can facilitate the more effective design of training curricula, the assumption with this strategy is that employers already know what their workforce needs are and what types of abilities will best meet them. Given that the World Economic Forum estimates that 65% of today's preschoolers will eventually be working in roles that do not currently exist, it is clear that modern employers are in need of tools that can help them craft their workforce in the face of ambiguity.

II. The Foundations of a New People Science

The idea of the new people science does not rely on any particular academic discipline. Rather, the basic concept is to use advancements in our understanding of and ability to measure people, behaviors, and thought processes to align people with opportunities that suit them well. In doing so, the objective is not only to increase the efficiency and productivity of organizations, but also to disrupt patterns of bias and

discrimination in the allocation of job opportunities. The goal of this section is to explain the basics behind the new scientific disciplines that have yielded innovations in measuring people, including the underlying research, the mechanics of measurement, and the benefits relative to traditional methods.

Before beginning this review, it is important to emphasize that no people science strategy should be thought of as a panacea for economic inequality. Disparities in hiring, pay, and promotion stem from various sources, including gaps in education and non-inclusive workplace environments. However, even in cases where candidates are equally capable of performing a job, traditional approaches to employment selection fail to provide an even playing field. The goal of this section is to demonstrate how newer scientific fields have expanded the possibilities for evaluating people in terms of their true potential, rather than in terms of their societal position.

Much of the new people science draws from practical applications of *cognitive science*, *neuropsychology*, and *cognitive/affective/social neuroscience*. Cognitive science is an integrative field that was established in the mid-20th century from related studies in psychology, neuropsychology, neuroscience, computer science, sociology, anthropology, and philosophy. The birth of cognitive science has also been attributed to advances in computer technology. The invention of computers that could perform the same kinds of tasks as humans led to a realization that underlying mental processes govern much of human behavior: If the human mind could be analogized to a computer, then human abilities could be likened to modular components on a motherboard or software package. Neuropsychology and neuroscience also led to similar insights about the human brain—that it was modular, with many different components that could be studied individually rather than only looking at broad, unitary concepts like IQ[3]. Importantly for the new people science, individuals can vary on a spectrum along each modular component, and these can be measured by their parts or in their synergy.

LINK TO SOFT SKILLS

At a high level, modern cognitive scientists produce insights that can be applied to employment selection by studying *constructs* using *behavioral experiments*. A construct is a concept describing an attribute that often cannot be measured directly but can be assessed using behavioral indicators or operational measures. Variations on cognitive, emotional, and social constructs represent variations in soft skills. Cognitive science has produced many important insights about the human brain of unique individuals, such as the neurological differences of bilingual speakers or how athletes and artisans hone their abilities. In the 1970s, experimental investigations of individual differences on these constructs using behavioral paradigms became commonplace. Revelle et al. (2011) explain the method: "We can investigate the relationship between individual differences and the experimentally manipulated conditions to test theories of individual differences."[4] With the advent of digital technology, researchers have looked more at the

individual, collecting data on very large numbers of people as they go about completing real-world tasks in order to make inferences on the cognitive and personality spectrums of humanity.

To summarize the basics of the new people science then, researchers use behavioral assays to conduct experiments that evaluate many domains of soft skills in well-defined and falsifiable terms. These experiments allow for the establishment of clear relationships (or lack thereof) between individual differences in a construct (e.g., cognitive or personality traits) and outcomes of interest (e.g., decision-making speed). When the outcomes of interest are also related to job performance (e.g., being an ER doctor requires a propensity for rapid decision-making), the same experiments can be used to evaluate job candidates in terms of fit for the role. Notably, advancements in technology have certainly allowed researchers to develop increasingly sophisticated tools to measure and observe human behavior, but many of the best-validated assays in use today rely on very simple designs that have existed for several decades.

In contrast to self-report questionnaires and other means of measuring soft skills (e.g., aptitudes), behavioral tools provide many benefits in the context of employment selection. As previously noted in this brief, self-report surveys limit an employer's ability to accurately assess a candidate's aptitudes and overall fit for a particular job. This is due to several biases that are especially likely to emerge in a high-stakes process like a job application. For example, *social desirability bias* reflects the human tendency to present oneself in a positive manner to others, but this tendency is mitigated in contexts where the respondent cannot tell what a test is meant to measure. *Reference bias* relates to the fact that survey questions often require a person to draw a standard of comparison (e.g., are you a hard worker?), and that standard may differ across individuals (e.g., does a hard worker turn in 80% or 90% of their assignments?). Behavioral assessments, on the other hand, do not require the establishment of a point of reference. Even if aware of these biases, individuals may lack the *introspective ability* to provide an accurate response to certain questions.

LINK TO JOB FIT

Additional advantages regarding the use of behavioral assessments in hiring include the *breadth*, *granularity*, and *non-directionality* of the characteristics measured. Regarding breadth, consider the fact that workers in the food services industry might benefit greatly from having a strong short-term memory. This is clearly not an aptitude that is easily incorporated into a questionnaire; however, through behavioral experiments, scientists have established a basic tool to measure it by testing how well a person can recall words presented to them serially. With this information, a restaurant employer could better screen for waiters who will correctly remember customers' orders. As for granularity, a personal attribute like decision-making capacity may be divided into a few different components, such as speed, consistency, and degree of confidence. While conventional wisdom might conflate "good" decision-makers with confidence,

in some jobs confidence might be far less important than speed or consistency. Behavioral assays often capture multiple dimensions of a trait that may be oversimplified on a self-report questionnaire. Finally, with non-directionality, behavioral data can better identify candidates not just in terms of whether they have a particular characteristic like sustained attention, but also whether they have its opposite. For example, while sustained attention might be advantageous for an accountant who needs to focus on a single task for a prolonged period, a very short attention span could be preferable in a fast-paced sales environment.

Despite the obvious benefits of behavioral data over self-reports, early behavioral experiments and their high-quality data have previously been confined to laboratory settings. This provided little opportunity for the behavioral assessment tools used by researchers to be translated to HR departments. However, in the last 20 years, the emergence of the web as a means of gathering laboratory-grade behavioral data has allowed newer people science to hurdle many of the limitations faced by traditional approaches. Many studies have established that online assessments retain the quality of measurement observed in classical, in-person settings. Considering that decades of cognitive science research has produced tasks to measure cognitive, social, and emotional attributes ranging from planning ability to motivation for rewards, the new people science has provided a range of validated tools that are now deployable in the context of employment selection. Examples of job-relevant traits that can be readily measured using web-based behavioral assessments are illustrated in Table 1.

Table 1: Some Aspects of People That can be Measured Using Behavioral Assessments Derived from Cognitive Science Literature

CONSTRUCT	DEFINITION
Planning Ability	The capacity to coordinate actions to achieve a specific goal, which involves representing and evaluating several possible actions along with their consequences.
Inhibitory Control	The ability to suppress inappropriate, unwanted actions and to resist and modulate impulses, particularly in the presence of conflicting signals to act.
Reinforcement Learning	The capacity to self-monitor one's responses and to evaluate external information in order to learn about the appropriateness of those responses.
Sustained Attention	A process that enables protracted performance on tasks—particularly those which require detection of rare and unpredictable signals—over extended periods of time.
Trust	The extent to which a person is willing to depend on others, specifically by choosing to give another individual the right to make a decision which affects both the persons in question even though negative consequences are possible.

While it is clear that the above constructs represent information that could be very useful for identifying strong employees, as presented in Section I of this brief, the modern hiring paradigm requires more than attention to job performance. The new people science can help account for dimensions of fit and soft skills

by mitigating the shortcomings of self-report methods, but it can also improve on the fairness and flexibility of the traditional hiring assessments.

LINK TO FAIRNESS

Regarding fairness, as previously discussed, certain demographic groups perform systematically worse on traditional GMAs, in part because a person's educational background can significantly affect scores. Consider an assessment like the SAT Reading test, which might ask respondents to read a passage about DNA and subsequently answer a series of questions to gauge their reading comprehension. If respondents have previously been exposed to the concepts presented in the passage, they will obviously have an advantage over respondents who have never learned the material before; put differently, the assessment makes it very difficult to separate a person's true reading comprehension abilities from their exposure to a high-quality education.[5] The consequences of this conflation are evident in test results: The SAT is just one example of a traditional assessment that is demonstrably biased against Black and Hispanic students. Behavioral assessments avoid such problems by measuring traits in a manner that does not require reference to a particular context, such as educational or cultural knowledge.

LINK TO FLEXIBILITY

Flexibility in employment selection can also be improved by the new people science's context-independent approach to measuring personal characteristics. In thinking about the future of work, it is clear that certain jobs will be rendered obsolete by forces like automation, posing the question of how workers who once held these jobs can be most effectively redeployed in the economy. As noted in Section I of this brief, prior efforts to retrain displaced workers have largely been ineffective, in part because they have treated workers as uniform in terms of skills, interests, and abilities. This one-size-fits-all approach may seem reasonable when the only information available about a person is their résumé, since many displaced workers may look the same on paper, but the reality is that prior work experience is an ineffective predictor of future job performance. With the new people science, workers can be evaluated in terms of their aptitudes, providing the opportunity to optimize the alignment of trainees with reskilling initiatives. Additionally, as new types of jobs emerge, behavioral assessments allow for a more granular evaluation of the cognitive, social, and emotional traits that may position a person to perform well in the role.

III. Theory to Application: Best Practices for Employment Selection

While Section II of this brief presented the theoretical basis for the new people science, the goal of this section is to explain how cognitive science insights can be made actionable in the workplace. Variants of the four principles presented here—*data integrity*, *criteria for success*, *model selection*, and *auditing*—are commonly discussed in the ethical AI literature. Particularly in light of the direct implications that hiring

decisions may have on job candidates' lives, careful thought on each of these issues is crucial to ensuring that the new people science positively impacts its constituents.

PRINCIPLE 1: DATA INTEGRITY

The data science community often discusses data integrity with the adage, "Garbage in, garbage out." Appropriate data is vital to making accurate, reliable judgments about people, and selecting inappropriate data can have dramatic consequences. Zip codes, for example, are commonly collected on a job application and easily processed by statistical or machine learning models. However, a person's zip code is also a strong predictor of educational attainment and wealth. This makes it a strong proxy for socioeconomic status and, due to historical grievances in the United States, a predictor of systemic racial bias. In using information like zip codes to predict workforce trends, some relationships with success might emerge, but only because of the underlying information about how privileged employees are.

The two signals of high-quality, predictive data are *validity* and *reliability*. Validity refers to how effectively a data input actually maps to the outcome of interest; because zip codes are measures of wealth rather than ability, they are not valid in the context of hiring. Conversely, measures *of job-relevant aptitudes*, such as focus or emotional intelligence, are far more valid metrics of success. Reliability refers to how stable a data input is over time. Descriptions of work history, for example, likely have low reliability, since the information provided by a résumé is subject to change depending on a candidate's judgment. Importantly, data must be both reliable and valid to be appropriate for use in employment selection.

PRINCIPLE 2: CRITERIA FOR SUCCESS

Success criteria relate to the question, "What does good look like in this situation?" In trying to use data to predict job performance, it is necessary to first establish a definition of performance. This tenet may stand in contrast to the gut reaction of many data analysts who are often inclined to immediately begin the process of looking for interesting patterns, but predefined success criteria are imperative for success. Without validating the definition of "good," an analytics team can find themselves running toward the wrong goalpost.

Consider the problem faced by the talent development group at the hypothetical Acme Corp. This analytics team wants to use people science data to select junior software developers who are struggling for placement in a special training program. However, the definition of "struggling" might have several possible answers, such as "relatively few tickets closed" or "high proportion of downstream tests failed." Upon gathering the data, the team is unsure how to interpret it and what values to benchmark against (e.g., are three, four, or five code commits per week considered "struggling?"). The group decides to collect the same data from senior developers to provide a standard for comparison; but this is an inappropriate strategy, because the two roles entail different responsibilities and allocations of time. If the

goal of reviewing the junior developers' performance data was to identify individuals who could benefit from additional training, the talent development team should have started by developing a better understanding of the program. For example, they might have identified the skills covered by their training course, identified work-related activities linked to those skills, and then compared junior developers to other people in their same role to benchmark performance.

Decisions regarding success criteria can only be made in the context of the analysis goals. Objectives such as reducing employee turnover, improving employee satisfaction, and increasing average sales all require different conceptions of what "good" looks like. In the absence of thoughtful decisions, inappropriately selected success criteria can actually undermine strategic goals. For example, an employer that wants to increase sales might change the organization's compensation structure to reward top sellers; however, this could have the effect of reducing employee satisfaction for most workers.

Proper definition for success criteria is necessary for establishing the criterion validity of a measure or for determining how well a measure accomplishes the desired outcome. The field of industrial-organizational psychology has codified two gold-standard forms of *criterion validity*. *Concurrent validity* is a metric of how well the measure correlates with success at the time of measurement. In machine learning, concurrent validity is satisfied through a process known as *cross-validation*, where a model is trained on some proportion of the data and tested on a held-out set. Concurrent validity satisfies the question, "Does this measure of success work on the data I have now?" As a complement, *predictive validity* is a metric of how well a measure predicts future success. Predictive validity is definitionally identical to concurrent validity, except that it is evaluated from model performance over time. Concurrent and predictive criterion-related validity are assessed at the model selection stage.

PRINCIPLE 3: MODEL SELECTION

Following the identification of appropriate data and a meaningful definition of success, the third principle for the new people science relates to the *how* of mapping these components to one another. While the term *model* might prompt discussions of sophisticated machine learning procedures, it is important to emphasize that a model is simply a framework for representing an idea, object, or process. A globe is a model of the Earth, a manager's roadmap is a model of project difficulty, and a stock projection is a model of financial success. Regardless of whether a model is identified using a machine learning algorithm or a human process, three aspects of the selection process are key to keep in mind: *performance*, *explainability*, and *fairness*.

Performance

First, a model must be performant, meaning it can be used to make accurate determinations about the real world. If a people analytics model is meant to predict which employees have leadership potential, a

performant model would be able to separate good managers from bad managers. Performance is most often assessed through cross-validation at the time of model building, as a means of establishing concurrent validity, but the same measures are used during the monitoring stage of model deployment to evaluate predictive validity. The degree of a model's performance is easily captured in a basic 2x2 table, often called a c*onfusion matrix* in the machine learning literature. A confusion matrix (see Figure 1) is a type of contingency table that represents how well a model performs on classification problems: given a set of positively and negatively labeled data, a perfect algorithm would result in only true positives (TP) and true negatives (TN), while any classification errors would be indicated by false positives (FP) or false negatives (FN). This is similar to how one could understand the efficacy of, say, a test for COVID-19, where high rates of false positives and false negatives suggest a low-quality test.

Figure 1: The Confusion Matrix for Machine Learning Classification

		REALITY	
		TRUE	**FALSE**
MODEL PREDICTION	**GUESS TRUE**	True Positive (TP)	False Positive (FP)
	GUESS FALSE	False Negative (FN)	True Negative (TN)

A model might be optimized for different performance metrics depending on the success criteria of the model. Here, we review three common, intuitive performance metrics.

- *Accuracy* is a fairly intuitive metric for a model's performance, representing the sum of True Positives and True Negatives over the total number of cases ([TP+TN]/N). In many cases when dealing with noisy data (as is often the case with human behavior), accuracy even as low as 60% could be considered acceptable, especially if the alternative is effectively chance.

- *Precision* is the sum of True Positives over the number of total positive guesses (TP/[TP+FP]). Precision is an appropriate metric for a model that is meant to predict an individual's likelihood of defaulting on a loan, since the success criteria is for the company to minimize losses on bad investments. Stated another way, the model should avoid False Positives (giving loans to clients that default on loans), but is less concerned with False Negatives (not giving loans to people who would have paid in full over time).

- *Recall* is the ratio of True Positives over all true examples (TP/[TP+FN]). As one example of its relevance, if a model is meant to identify at-risk youth for a drug intervention program, the cost of selecting a teenager with low risk of drug use (False Positives) to participate is relatively low. In

contrast, the cost of failing to select someone who is high risk (False Negatives) is very high. In this case, optimizing for recall would be important.

It should be noted that these three metrics, while related, often come at a tradeoff. As precision rises, recall often suffers. Accuracy may improve due to large True Negative rates, even as recall and precision drop. Defining performance metrics upfront will save much frustration at the time of analysis.

Explainability

Second, the selection process should yield a model that is explainable, meaning those who use it are able to understand why one conclusion was reached over another. Consider a model that is meant to predict academic achievement in college from a variety of metrics collected from high school students. Of course, a guidance counselor who sees that a student has low odds of success will want to know why the model made that choice and what can be done to improve their prospects. In short, a useful model is an actionable model.

Model explanations are also important to establish enough trust with users that they are willing to act on the recommendations produced. Users are especially likely to distrust models that cannot explain their conclusions when those conclusions contradict human instincts. Consider what might happen if a retention model indicates that a given employee is highly likely to leave their employer. The employee's manager might find it difficult to take this prediction seriously in the absence of an explanation because that employee has a long track record of being hardworking, likable, and consistent. A clear explanation could help the employer understand if the problem lies in the algorithm (e.g., homogeneous training data failed to capture this type of employee) or in the manager's biased perceptions.

Notably, the issue of model explainability has become more prominent with the increased use of black-box models, which often interpret data using complex, nonlinear structures that cannot be made interpretable for a human. However, even black-box algorithms can and should be explained to a limited degree.

Fairness

Finally, the model selection process must account for fairness. In truth, fairness is the desired outcome of a model that is both accurate and explainable. While the great promise of machine learning is to automate human decision-making to drive efficiency, precision, and objectivity, the reality is that models are not infallible. Algorithms are trained using data collected in the real world; therefore, if not carefully considered, systemic biases in society can be replicated in scaled technology. Well-known examples include high-profile cases like Amazon's discarded "sexist" résumé screening program and Google's hate speech classification algorithm that contained bias against Black vernacular English. Because of these incidents, public concern regarding the issue of biased AI has increased in recent years.

Another way to think about the parameter of fairness for a model is in terms of its *generalizability*, or the extent to which it can be applied to all people and all contexts where it will be implemented. In a 2017 review, Sendhil Mullainathan and Ziad Obermeyer argue that the prevalence of machine learning models without attention to the systemic biases they may contain is a moral hazard. They further offer that a model that appears accurate when using biased performance metrics may, in fact, exacerbate societal problems. Take the example of a model that predicts healthcare outcomes based on mostly biographical datasets, with items like marriage status, weight, insurance claims, and recent bloodwork. The model is statistically more accurate than medical professionals at predicting health outcomes of 1,000 participants in a clinical trial. Now, consider again the datapoint of insurance claims. Due to systemic bias, participants from privilege are more likely to have better insurance, which in turn means that they are more likely to use that insurance for frequent medical checks, and in turn have better health outcomes. Conversely, someone without that privilege, with poorer insurance and less income, will likely have fewer health checks due to the expense and therefore is likely in worse health. A model trained and tested only on white men between the ages of 35 and 60 may appear accurate, but may prove to be much less performant when applied to a diverse dataset.

Issues of fairness in human and algorithmic usage of data is an incredibly heated topic at present. Fundamental questions remain unresolved in the academic and policymaking community, including: "What constitutes bias?" and "How is bias ethically resolved?" These challenging debates are further complicated by the broad range of legal definition of fairness across sectors and geographies; for example, in the United States, standards are inconsistent among employment, real estate, and financial services. Governmental and nongovernmental organizations around the world are also developing their own standards. The exact standards for success in each category depend on the type of model, its use case, and the existing benchmarks of success. A model with 90% accuracy may seem like a high standard, but in the domain of handwriting recognition it is remarkably poor. Likewise, a model that is 20% less accurate at evaluating women than men may seem inequitable; but in the realm of workplace performance evaluation, human evaluators may be far worse offenders. Biased algorithms ultimately derive from biased humans; it is inevitably easier to fix the algorithms than the people.

It is beyond the scope of this brief to adjudicate the countless proposals for analytical definitions of fairness that have emerged in academic discourse. In the context of people science, the relevant definition is *group-level statistical parity*. While alternatives such as "fairness by blindness" and "counterfactual individual fairness" might be compelling on epistemological grounds, they are not appropriate in the context of employment selection for three key reasons. First, the primary pro-social benefit of AI for hiring is to mitigate human bias as a means of improving workforce diversity; the only way to ensure this goal is by actively optimizing models to overcome the demographic trends of the status quo. Second, the simple omission of demographic information in training data cannot actually guarantee a "race-blind" or

"gender-blind" process, since proxies for these variables are extremely common in employment data. Third, and perhaps most importantly, employment tools are currently regulated by a definition of group-level fairness that compares the selection rates of different demographic groups; in this view, "fairness" is synonymous with a lack of *disparate impact* against a protected class. As it is certainly not the intention of this brief to contest the wisdom of Title VII of the Civil Rights Act of 1964, our assumption is that talent selection systems should be built in accordance with the law.

A model is ready for production if it passes these three broad standards of performance, explainability, and fairness. The final judgment for using people analytics for any model, be it an advanced neural network model or the mental model of a regional manager, ultimately depends on the context in which it is used.

PRINCIPLE 4: AUDITING

Upon building an accurate, explainable, and fair model, the final tenet of the new people science is auditing. Auditing is how responsible practitioners of people science ensure that well-designed systems are living up to their true promise. In this way, development teams should not view it as a "nice to have" for the model building process, but as the only way to establish confidence in the decisions generated by machine learning.

As an example of how auditing might work, consider a people analytics model that looks for indicators of illegal financial activity in employee emails. The training data for this model is a large repository of old emails, collected from previous employees convicted of wrongdoing. The development team's relevant success criterion is fairly straightforward: The model should be able to detect fraud in the future. To quantify this criterion, they decide to optimize the system to reduce false negatives, since the cost of failing to catch a suspicious email is higher than the costs of needing to manually review an extra email that ends up showing no wrongdoing. Per the tenets described earlier in this section, the model they build appears accurate, explainable, and fair.

The auditing process begins with the *monitoring* stage, where an analyst will track the usage of the model over time. The first few cases will be investigated by hand to see if the model is picking up on an anomaly in email traffic, such as an uptick in conversations about a company called Fraud Inc. driven by a public scandal. In addition to looking for such false alarms, this part of auditing should entail simple and intelligible metrics, such as the proportion of alerts raised, the departments those alerts came from, and the specific email addresses that raised suspicion.

There is no one-size-fits-all approach to what analysts should look for in monitoring. In this example, the team may come to understand typical versus atypical patterns, such as alerts frequently being raised by employees in IT Security, due to the language they commonly use in emails. One best practice for

monitoring also involves examining the demographic backgrounds of people flagged by the model across dimensions like race, gender, culture, or national origin, to determine whether any proxy variables are resulting in disparate treatment against a given group. Initial insights formulated during the monitoring process may prove useful in establishing the scope of a formal audit.

If no obvious red flags appear, monitoring should nevertheless proceed for a set time before conducting an actual audit. Audits provide feedback on the performance and fairness of a system in the real world. Well-known examples include explorations that revealed inequities in Uber's opaque surge-pricing algorithms and leaks of personal information in Facebook's ad recommendations. Because statistical tests are used to conduct such assessments, audits can only happen when a reasonably large amount of data is collected, otherwise the process can be undermined by sampling bias. Audits should also not be conducted too frequently in order to avoid an issue known as the problem of multiple comparisons. This dilemma refers to a fairly intuitive situation: If a test to look for an issue is performed enough times, the test will eventually come up positive at least one time due simply to chance. By overcorrecting for a chance result, people science teams can unnecessarily reduce the performance of a useful model.

If a model is found to be behaving unsatisfactorily after the period monitoring and auditing, the development team must undertake *remediation*. The definition of remediation also changes across context, but in general the model designers should take the data that has thus far been collected into account when constructing a new model. Strong model explanations can greatly improve remediation efforts. For example, if the system is falsely flagging women more often than men, analysts might explore the keywords that women use more often than men and de-weight these terms to reduce the number of false alarms. Notably, this process only works as described with explainable models. Black-box models can also be audited using penalty parameters; while no one will know exactly why the model misbehaved, it can still be subjected to repeated testing and monitoring.

SOME COMMENTS ON TRAINING DATA

People science models that are built with careful consideration of *data integrity*, *criteria for success*, *model selection*, and *auditing* are very likely to positively affect any talent process. Given the degree of public anxiety over poorly designed machine learning models in recent years, it is also worth commenting on one issue that has preoccupied many people thinking about the ethical implications of AI for hiring: the inevitable imperfection of a training dataset.

Concerns regarding training data stem from a key tenet of employment science: In order to select job candidates who will be successful in the future, it is necessary to look to previous candidates who performed well in the role. However, because those previous candidates were selected and evaluated through biased processes (e.g., using data or procedures that disproportionately benefit white men), it is

impossible to know whether they would have still been identified as top performers if the process had been fair. It could be the case, for example, that many women of color would have outperformed their peers if they had been assessed in an objective manner, but the status quo excludes them from the training dataset, suggesting that the model may fail to identify similar high-potential candidates when deployed. In other words, the *ground truth* of which people are truly the best employees is not perfectly known.

Unrepresentative training data can certainly be a problem for effective employment selection models; but for employers who are genuinely invested in disrupting traditional hiring processes, this issue can be mitigated in a few ways.

First, in identifying the set of incumbent candidates to train a model, organizations should be cognizant of the bias that may exist in their internal performance evaluation systems. Research has shown that subjective supervisor ratings are easily affected by managers' cognitive biases, making them very poor measures of real success. As such, instead of simply opting to build a model on the incumbents an employer *believes* are effective, an organization should closely examine available "hard" metrics, such as revenue for a sales role. While an employer's internal procedures may fail to recognize a strong performer because of cultural or other biases, such a strategy ensures that the training dataset will still capture their unique competencies.

Second, as mentioned in Section II of this brief, advances in cognitive and behavioral science have dramatically expanded the types of information that can be easily collected about a person. While traditional hiring criteria like prestigious degrees and standardized test scores are strongly correlated with demographic identity, many cognitive science measures that relate to job performance—personality traits, decision-making styles, etc.—are *not* biased along gender or racial lines. Accordingly, even if incumbent employees are homogeneous in terms of race or ethnicity, a model trained to look for certain cognitive science measures among candidates will identify strong fits across all demographic groups.

Of course, both of these training data strategies must be underpinned by a model selection and auditing process that prioritizes fairness in order for the new people science to truly achieve its potential.

IV. Case Study: Using the New People Science to Assess Candidate Fit to High-Growth Jobs

To demonstrate the nature of the insights that can be produced by adhering to the tenets identified in Section III of this brief, this section serves as a case study on one organization that is practicing the new people science. The data presented here comes from a New York-based startup called pymetrics, which has been developing job models since 2016. Four of this brief's authors, are affiliated with pymetrics, which permits us to use the company's data and examples for this case study. These models are built on

behavioral assessment data from top-performing incumbents in a particular role. They are then used to evaluate candidates (who take the same behavioral assessment) in terms of their fit to various roles. This process of assessment notably does not rely on data inputs like education, hard skills, or work experience.

The pymetrics models are designed with a particular part of the hiring pipeline in mind. After receiving applications, employers need a means of efficiently deciding which candidates should receive further consideration (e.g., first-round interviews, timed work sample submissions, etc.). As discussed in Section I of this brief, traditional filtering processes often involve scanning résumés (either manually or automatically) and sorting them into "yes" and "no" piles, having candidates complete an IQ or personality test, and removing everyone who does not meet a particular cutoff score or profile, or electing to only seriously review résumés from particular universities. One implication of such practices is that much of the diversity in a hiring pool is eliminated, without any specific consideration for the particular needs of the role. The pymetrics assessment serves as an alternative to this "filtering" part of the hiring process; once candidates complete the platform's assessment, a custom job model will be used to recommend a subset of people for further consideration based on their fit to the job. Crucially, the suggestion of model recommendations is not that other candidates *cannot* succeed in a role. Rather, in cases of high-volume hiring, narrowing the candidate pool is an operational necessity for employers; the goal is simply to provide them with a means of filtering that is fair and effective.

At a high level, pymetrics produces two types of job models: *industry-level* and *employer-specific*. Regarding industry-level models, one common use case is an organization seeking to hire for a role they have never had before, such as a new digital marketing analyst position. To help the employer evaluate candidates, a general digital marketing analyst model can be built by using training data from individuals in very similar roles, aggregated across different employers[6]. For employer-specific models, the goal is generally to assess a large number of applicants for a particular role in terms of their fit—for example, sorting candidates into *high-fit* versus *low-fit* categories to determine whom to interview. In this case, the training data comes from top-performing individuals who are currently in the target role at the relevant organization.

The first part of this case study (Section IV) will focus on industry-level models, demonstrating how the new people science can be deployed to develop success profiles for five rapidly growing roles: data science, systems engineering, front-end engineering, digital marketing, and software development (summarized in Table 2). The primary question answered by these models is at the group level, such as an employer who has just started engaging with a new technology and is in need of an effective way of dividing an extremely large applicant pool into high-fits and low-fits for the associated role. The second part of this case study (Section V) will shift attention to the employer-specific models, which have greater utility in the context of optimizing role transitions for displaced workers.

To frame the discussion of the industry-level models, this section will use the four tenets presented in Section III of this brief: data integrity, success criteria, model selection, and auditing.

Table 2: Target Jobs for Industry-Level Models

HIGH-GROWTH ROLES

ROLE	O*NET CODE	DESCRIPTION	GROWTH
Data Science	**15-2041**	Apply statistical and computational methods to deliver meaningful products and insights. Conduct research to identify the best solutions. Visualize and communicate findings to technical and non-technical audiences. May have several specializations, such as machine learning, statistics, analytics, data engineering, or research.	Much faster than average **(11% or higher)**
Systems Engineering	**15-1199.02**	Design and develop solutions to complex applications problems, system administration issues, or network concerns. Perform systems management and integration functions. Verify stability, interoperability, portability, security, or scalability of system architecture. Communicate with staff or clients to understand specific system requirements.	Faster than average **(7% to 10%)**
Front-End Engineering	**15-1134**	Architect front-end systems that drive complex web applications. Collaborate with Product Designers, Product Managers, and Software Engineers to deliver compelling user-facing products. Experience in creating robust UI automation testing as part of the development process.	Much faster than average **(11% or higher)**
Digital Marketing	**15-1199.10**	Employ search marketing tactics to increase visibility and engagement with content, products, or services in Internet-enabled devices or interfaces. Examine search query behaviors on general or specialty search engines or other Internet-based content. Analyze research, data, or technology to understand user intent and measure outcomes for ongoing optimization.	Faster than average **(7% to 10%)**
Software Development	**15-1132.00**	Develop, create, and modify general computer applications software or specialized utility programs. Analyze user needs and develop software solutions. Design software or customize software for client use with the aim of optimizing operational efficiency. May analyze and design databases within an application area, working individually, or coordinating database development as part of a team. May supervise computer programmers.	Much faster than average **(11% or higher)**

DATA INTEGRITY

The datasets collected by the platform are objective measures of real-time behavior, measured via gamified assessments. These assessments—commonly described as "tasks" or "games"—are based on decades of behavioral science and psychological research, which have been adapted into a single battery. As per the gold standard, the exercises used are all derived from the cognitive science, behavioral science, and behavioral economics literature and are therefore substantiated by decades of scientific research on measurement and construct validity. These measures have also been linked to more traditional personality measures, such as the Big 5 and employment outcomes in previous peer-reviewed studies. For instance, high scores on the *Balloon Analogue Risk Task* (BART)—one measure used by pymetrics to assess the propensity for risk-taking—have been found to be significantly associated with effective workplace *maverickism*, the tendency to engage in bold, creative, and often disruptive behaviors that deviate from the status quo but which are ultimately beneficial to an organization. Each game measures a targeted construct in social, emotional, or cognitive realms, as described in the two examples below. Data is collected over multiple games, each with multiple trials, to improve the reliability of the data above a single measurement.

It is beyond the scope of this brief to summarize all of the constructs measured by this platform. Rather, the goal is to demonstrate the type of data collected by a couple of the assessments, and how this data can then facilitate the employment selection process.

Example Behavioral Assay 1 – Flanker task:

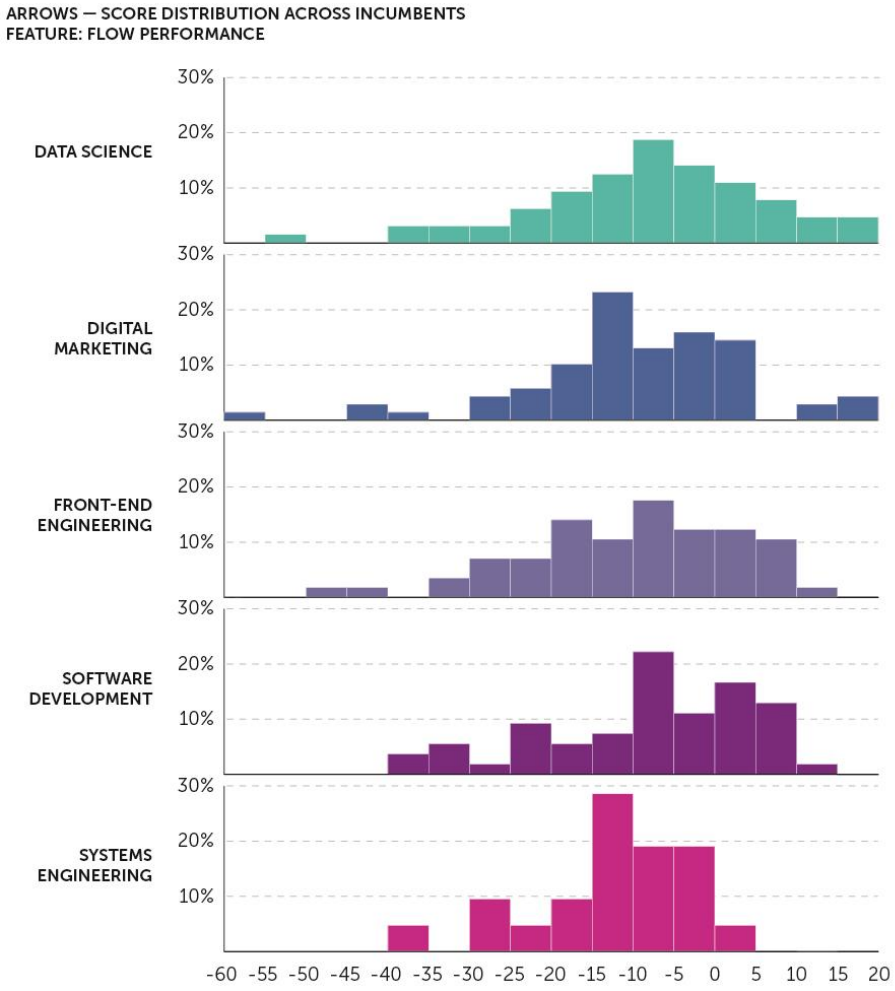
First, consider *the flanker task*. The flanker task is used to measure *attentional control*, *response inhibition,* and *cognitive inhibition*. In the tasking-switching flanker task, participants are presented with stimuli consisting of a central stimulus and flanking stimuli—sets of five arrows (e.g. <<><< or <<<<<). Players are asked to press a button that corresponds with the middle arrow when shaded blue and to respond similarly to the outside arrows when shaded red. Crucially, the rules for responding to the stimuli change, requiring participants to task-switch if the rule for one trial differs from the next. The flanker task is a common and reliable measure of executive attention, the ability to shift attention depending on context, and has been demonstrated to be a stronger predictor of supervisor performance ratings than tests of general mental ability.

Each play through the flanker task produces hundreds of raw datapoints. The raw data is converted to scientifically interesting features that can be used for building a predictive model. It is important to emphasize that particular features are never interpreted as universally "good" or "bad"—rather, they are simply reflective of fit. Even in the case of executive attention, which many would view as a positive trait, there is no good or bad connotation with direction. For example, data science requires uninterrupted work, and it would make sense for top performers to perform better on successive trials and to be slowed by

distractions. Conversely, salespeople often deal with rapid changes in conversation or workflow, and may perform relatively better when distracted rather than focusing on a single task. Depending on the results of the machine learning analysis, features will vary in terms of their weighting across models. In other words, reaction time (whether fast or slow) might be an important variable for assessing someone's potential as a data scientist, but irrelevant for assessing their potential as a systems engineer.

Of course, it would be impossible for a human being to manually interpret trends in the highly granular data collected by an assessment like the flanker task, highlighting the importance of machine learning to the new people science. For the sake of illustration, a histogram of incumbents' scores on a single feature is shown in Figure 2. This feature broadly measures how users respond when they have successive correct trials (e.g., the user gets "in a flow").

Figure 2: Data Collected from a Single Feature of a Single Behavioral Assay



Notably, it is not possible to draw any definite conclusions from the distributions shown above, because even if a feature trend seems obvious across incumbents in a given role, features are assigned weights

during the machine learning process. Once features have been weighted with the assistance of machine learning, however, it is possible to get a better sense of which features are associated with success across the various models. This information is captured by values called *feature importances*, communicating both the directionality and weight associated with an incumbent success profile.

Example Behavioral Assay 2 – Dictator task:

Another assessment, known as the *Dictator Game* (Forsythe et al., 1994), is used to measure altruism. In the Dictator Game, participants are matched with an anonymous partner, and both receive a sum of money.[z] Throughout the game, participants are allowed to share money with their partner, and eventually take money from them. The game measures altruism when allocating finite resources, which may notably be a useful trait in some jobs (e.g., caretaking) and a less useful trait in others (e.g., financial planning). Altruism can be assessed through a number of features in the game, most intuitively by the amount of money transferred by the participant to the partner. Since its conception, more than 100 experiments have been published using the Dictator Game as a measure for altruism. In peer-reviewed literature, the Dictator Game has been related to real-world generosity, as well as real-world shrewdness. Likewise, it has been measured as a predictor of leadership and team performance.

CRITERIA FOR SUCCESS

The success criterion for this model is the ability to select applicants from an applicant pool who are successful in the target role. Success may be defined as increased job satisfaction, productivity, or tenure in the role. Clients identify high-performing employees through the use of a job analysis tool that evaluates the objective skills and abilities required for agreed-upon success criteria. This tool is derived from O*NET (The Occupational Information Network), the U.S. Department of Labor's definitive ontology of occupational requirements. The skills and abilities identified in the job analysis tool are then reviewed for confirmation of a link with the behavioral traits measured by the exercises. This evaluation is used to select the top-performing employees who will be used as a benchmark for success.

After the selection process of high-performing individuals is complete, those individuals are asked to go through the assessment battery. Cross-validated models then allow for the identification of cognitive, social, and emotional features that are unique to high-performing individuals in those roles. In addition to this data-driven methodology of identifying success profiles, a concurrent job analysis also is conducted.

While the pymetrics approach is primarily data-driven, job analysis is utilized to ensure the success of data science methods and to guarantee that the final model is explainable and defensible. pymetrics utilizes a multi-method approach to job analysis for each engagement involving job description review; stakeholder and subject matter expert (SMEs) interviews to understand critical successful behaviors; and a structured, survey-based approach to understand the actual knowledge, skills, abilities, and other skills (KSAOs) and

work activities identified as relevant by incumbents. Job analysis results are used to: (1) ensure that the successful job incumbents identified by each client for model-building belong together as a collection of employees doing similar work; (2) understand the relationship between behavioral measures and categories that emerge as predictive of performance in our models and the actual KSAOs and work activities identified as relevant for the position; (3) document local content validity of pymetrics' success models for legal defensibility; and (4) provide additional insights to clients about the relevant jobs within their organizations, including how they are both similar to and different from one another.

MODEL SELECTION

The model building process stems directly from the success criteria. Traditional machine learning follows a set of data labeled to be *yes* or *no*, (i.e., *good hire* vs. *bad hire*). The criteria for success, however, is not to differentiate good and bad employees, but rather to select a potential good employee from a pool of applicants. There happens to be a field of machine learning dedicated to this problem, known as *semi-supervised learning*, where data from a few known positive examples is used to identify patterns with exponentially more unknown examples (e.g., the general population). If the model building process were to have ignored the success criteria and used a more traditional machine learning approach, potentially the wrong behavioral traits could be selected and applied, leading to worse outcomes. All models are optimized for performance to minimize false negatives (incorrectly rejecting a strong candidate). Models are also optimized to maximize fairness using their open-source software package, audit-AI. All models have an explanatory layer that is used to provide feedback to the data scientists building the models, to the recruiters using the models, and to the applicants who are scored against the models.

Performance

Model performance is evaluated using *criterion-related concurrent validity*, known as *cross-validation* in the machine learning community. Models are trained on 80% of the data and tested on a held out 20%. The data is shuffled and the process is repeated so that every datapoint is held out exactly once. The average performance on this data yields an estimate of the model's behavior. The model's success criterion is the selection of a quality candidate from the applicant pool, where the cost of a false negative (rejecting a good candidate) is much higher than the cost of a false positive (interviewing a poor candidate). As such, models are optimized for *recall* (selection rate of current employees during cross-validation), but also *overall accuracy*. The average recall of the five models discussed here is 81.4%, and the average accuracy is 70.4%. Model sample size across the five models was n=57,858.

Explainability: Factor analysis method

When dealing with numerous data sources—for instance, when pooling individuals' results on multiple games which each collect many features—a common approach to interpretation involves identifying

higher-level factors that describe the data in more interpretable terms. Measures from across assessments can be combined through a process known as C*onfirmatory Factor Analysis* (CFA). CFA uses a data-driven component to group features into factors, each of which consists of multiple related measurements. These factors provide more meaningful interpretations of results that can reliably measure higher-order constructs (e.g., decision-making), while maintaining explainability. CFA is confirmed and named by experts in cognitive and personality sciences, who verify the integrity of the data's interrelations and ensure factors are correctly interpreted.

The nine factors produced by the CFA process are provided in Table 3. Factor scores capture where an individual falls on the spectrum of a given construct; for example, for *Altruism*, the spectrum is *frugal* to *generous*. As is the case with more granular features, a given score cannot be interpreted as universally good or bad; while being generous might be useful for a home health worker, being frugal is likely more appropriate for managing a tight budget. Also, in line with features, factors vary in terms of their weighting in a given model; while Altruism might be a very significant part of the home health worker model, it could be irrelevant in gauging someone's fit for being a copy editor. *Factor importances* for the five high-growth roles described above can be found in Figures 3 through 7.

Table 3: Descriptions and Spectrum Ends of Pymetrics' Factors

◂ Frugal Focused on achieving personal goals and self-sufficiency. A shrewd investor of resources.	Altruism Tendency to prioritize the needs of others above your own.	Generous ▸ Trusts the good intentions of others. Balances personal desires with the needs of others.
◂ Biased to action Not easily flustered by mistakes. Quick to react and open to information outside of the immediate task.	Attention Approach to managing incoming information and distractions.	Methodical ▸ Thorough and restrained. Prefers accuracy over speed to avoid mistakes.
◂ Multi-tasking Quick thinker with a shorter attention span. Tends to handle changes in multiple tasks or environment smoothly, adapting to dynamic circumstances with fast responses.	Cognitive Flexibility Concentration style for one or more tasks.	Focused ▸ Focused and consistent in their work, with above-average memory. Can effectively attend to a single task even in the presence of distracting information.
◂ Instinctive Acts quickly and decisively. Trusts intuition rather than focusing too much on planning.	Decision-Making Approach to making decisions.	Deliberative ▸ Reflects before making decisions rather than following 'gut-instinct'. Thoughtful planner who takes time to deliberate before reacting.
◂ Outcome-driven Works selectively, rationing effort for high reward tasks. Prefers simple alternatives.	Effort Expenditure Effort invested based on size of reward and probability of success	Hard-working ▸ Works hard for all tasks, regardless of expected reward.
◂ Context-oriented Interprets emotions from contextual cues more than facial expressions.	Emotion Focus Strategy for interpreting others' emotions.	Expression-oriented ▸ Interprets emotions from facial expressions rather than contextual cues.
◂ Critical Slower and more critical when judging the fairness of social situations.	Fairness Perceptions Perception of the fairness of social situations.	Accepting ▸ Quick to judge most situations as fair.
◂ Consistent Not deterred by mistakes and prefers taking known approaches over trying new ones. Takes time to deliberate before changing approach to a problem.	Learning Tendency to change behavior based on new information	Adaptive ▸ Learns quickly from mistakes, recognizes patterns in the environment, and changes behavior based on immediate feedback.
◂ Cautious Carefully tests options and chooses the safe alternative. Avoids negative outcomes.	Risk-Aversion Comfort with risk-taking.	Adventurous ▸ Willing to take risks despite potential repercussions. Responds quickly with less concern about the negative outcomes.

Figure 3: Factor Importances for Systems Engineering Role

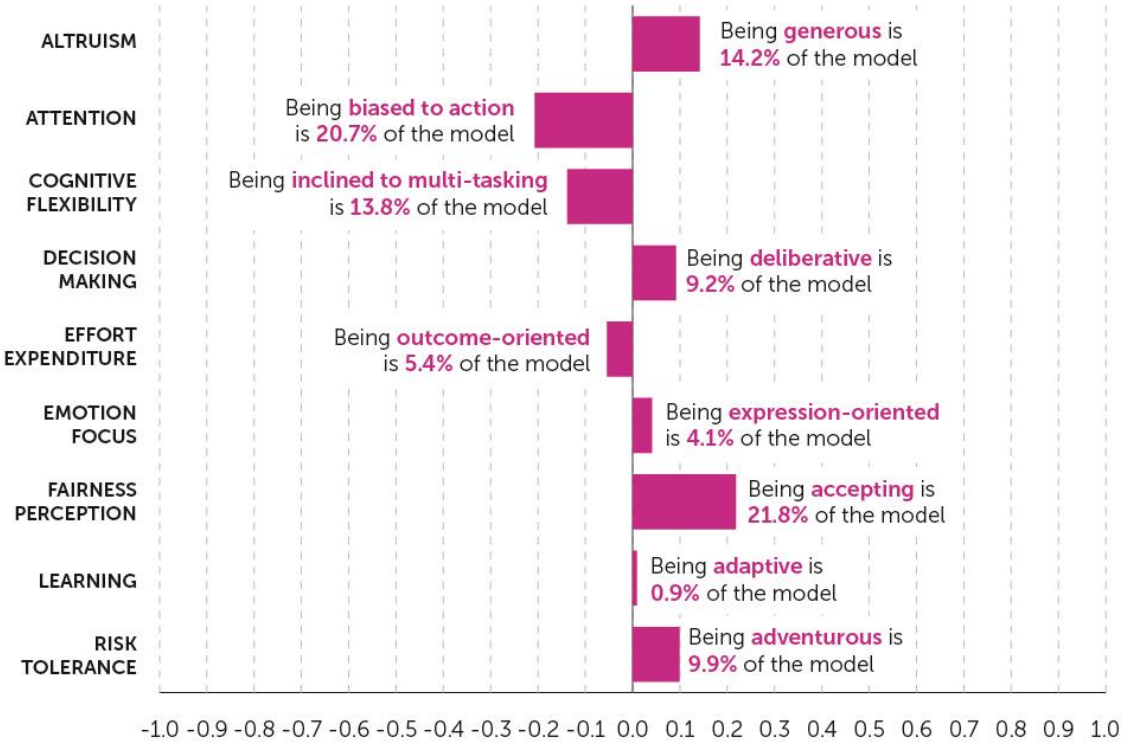


Figure 4: Factor Importances for Digital Marketing Role

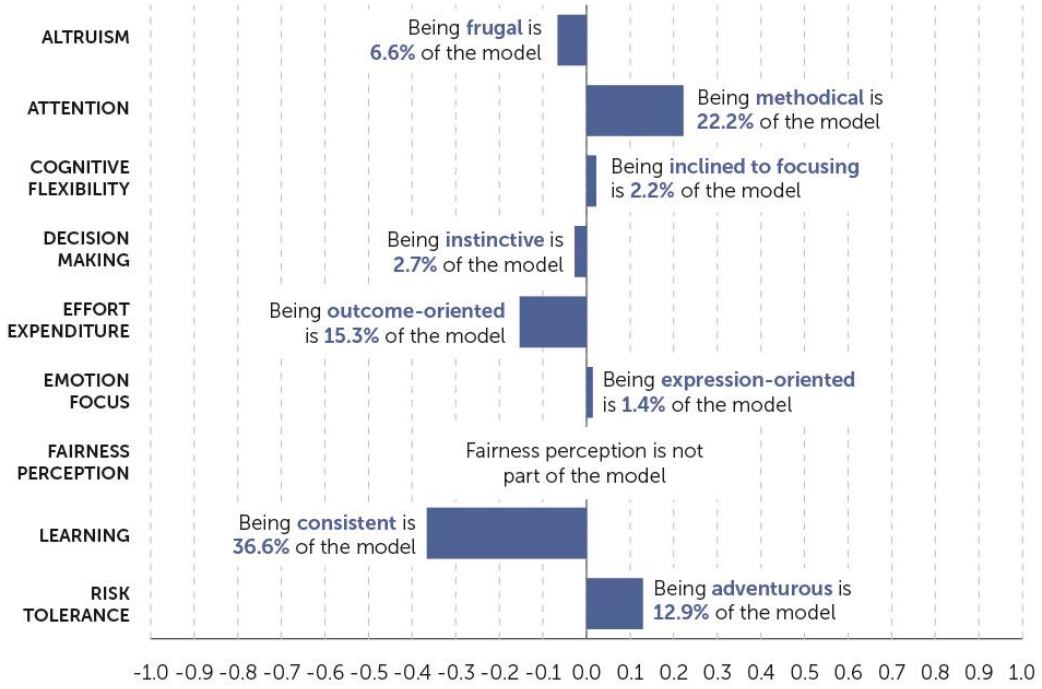


Figure 5: Factor Importances for Data Science Role

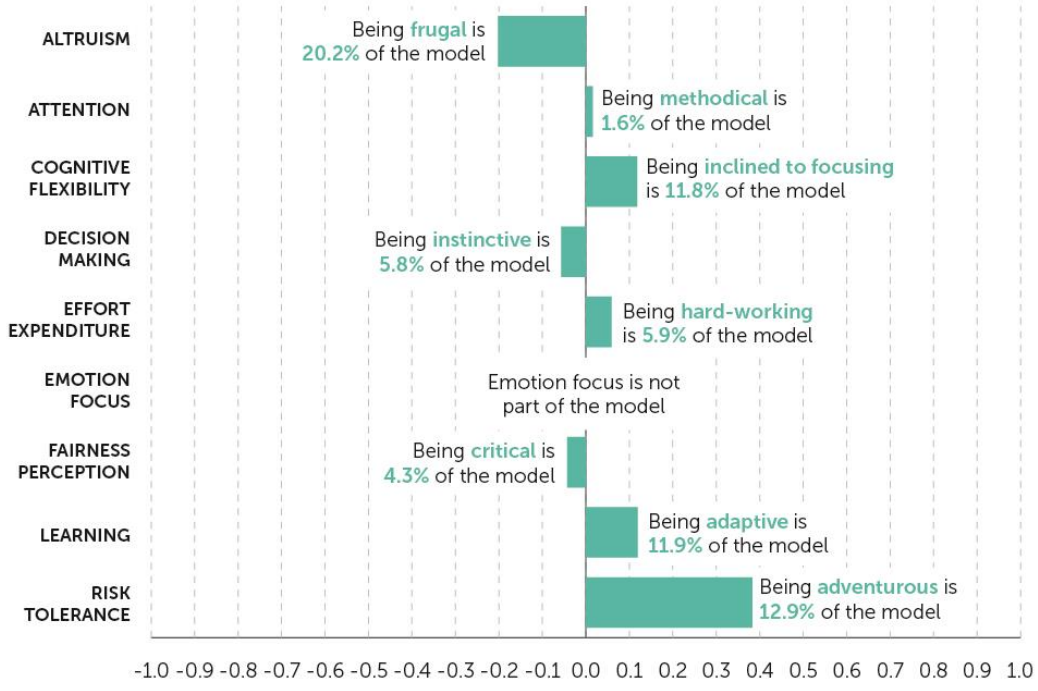


Figure 6: Factor Importances for Software Development Role

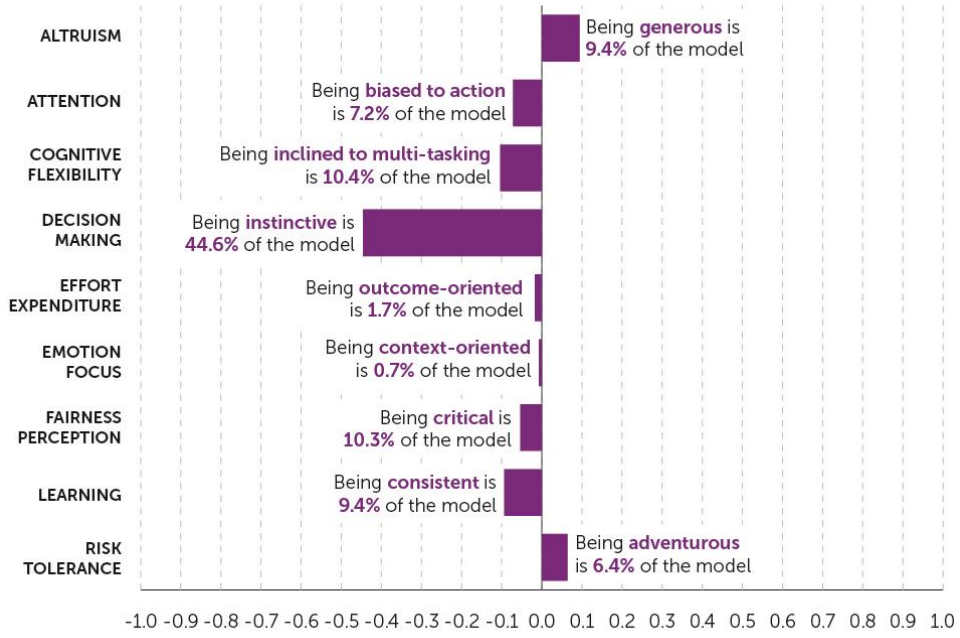
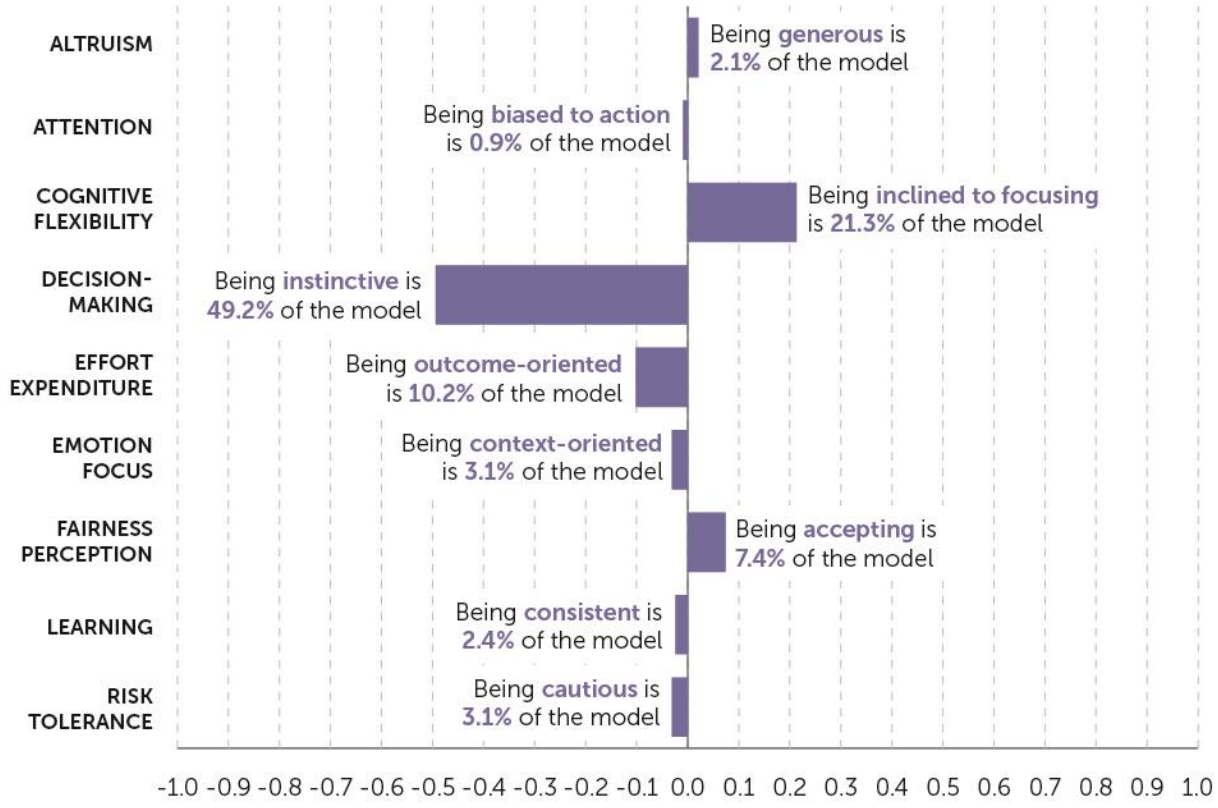


Figure 7: Factor Importances for Front-End Engineering Role



Fairness

All models are tested for fairness prior to deployment. This is done by testing a representative sample of the population against the model, and comparing the relative pass rates of both gender and racial groups. **As these models are used for employment selection, the regulations for fairness fall under the auspices of the U.S. Equal Employment Opportunity Commission (EEOC), which defines fairness using the Four-Fifths Rule.** This states that the lowest passing group must be recommended by the model at no less than four-fifths (80%) of the highest passing group. As shown in Table 4, all models meet this threshold, with an average pre-deployment estimated bias ratio of 84.1%.

Table 4: Model Fairness – Pass Rates by Gender and Ethnicity

	GENDER PASS RATE		ETHNICITY PASS RATE				MINIMUM
Model Name	Female	Male	Asian	Black	Hispanic	White	Total
Systems Engineering	0.927	1	0.951	0.968	1	0.857	0.857
Data Science	0.899	1	1	0.854	0.883	0.972	0.854
Digital Marketing	1	0.939	0.938	0.947	0.835	1	0.835
Front-End Engineering	0.877	1	0.956	0.800	1	0.953	0.800
Software Developer	1	0.875	0.890	0.896	0.994	1	0.875

AUDITING

Lastly, the platform uses a strong monitoring and auditing framework. Models are monitored through a series of dashboards, and audits are conducted every 6 to 12 months by an internal but independent audit team. Due to the transparency of the model selected, models can then be refreshed to improve performance and fairness using the data collected during the monitoring phase.

SOME COMMENTS ON MODEL IMPLEMENTATION

The above discussion focused on the use of job models to evaluate candidates for fit to roles. Here, it is worth reiterating that pymetrics' models: (1) do not suggest that a candidate *cannot* succeed in a job, and (2) are not meant to replace the entirety of the hiring process.

On the former point, it is useful to think of filtering job candidates as an inevitability for large employers. The simple reality is that an organization engaged in hiring needs to have a system for deciding which applications to prioritize, but they may either use a fair or an unfair procedure for doing so. The new people science facilitates alternative ways of gauging which candidates are *most likely* to perform well in a job. Moreover, the new people science's attention to *fit* disrupts the longstanding practice of the same

types of people repeatedly being deemed the "safe bets" for employers, thereby increasing the diversity of the workforce.

On the latter point, it is important to note that hiring procedures are virtually always multi-staged. A job model, whether built on new or old understandings of employment science, only functions in one part of the process. In the case of pymetrics, as previously mentioned, the models *do not* evaluate applicants on the basis of educational or technical criteria. One reason for this: The extent to which formal qualifications are actually relevant for job performance varies widely across industries and roles. A law firm, for example, would obviously need to restrict its candidate pipeline to those with a Juris Doctor (JD) degree before using a pymetrics model to assess for fit. In other cases, however, arbitrary degree or skills qualifications are applied to job postings, simply in an effort to decrease the volume of applications. In these instances, the use of an alternative sorting mechanism can lead employers to think more carefully about whether additional filters are truly necessary. Where they are not, removing them can be an additional source of increasing workforce diversity.

V. Case Study: Using New People Science Job Models in the Context of Workforce Redeployment

While the previous section focused on how industry-level models can help sort large groups of applicants on the basis of fit, this section shifts to a discussion of employer-specific models. Specifically, these employer-specific models demonstrate how cognitive, social, and emotional traits can support the redeployment of at-risk or displaced workers, whether caused by automation, globalization, or unprecedented public health events. Here, we particularly focus on the issue of workforce transitions as experienced by some industries in the face of COVID-19.

CONTEXT

It would be difficult to understate the degree of social and economic disruption caused by COVID-19. While the full extent of its impacts remain to be seen, tens of millions of workers around the world have been displaced by a combination of demand for certain services evaporating overnight and businesses accelerating investments in automation as they struggle to find a new normal. Policymakers around the world are faced with unprecedented questions in the face of this "Reallocation Shock." In the short term, governments are responding with unemployment insurance payments in the United States and subsidized wages in Europe. In the long term, however, plans are less clear. As economist and former Governor of the Bank of England Mark Carney writes, "How many once-viable companies will be permanently impaired? And how many people will lose their job and their attachment to the labour force? The answers to these

questions...will be the true measures of the effectiveness of the responses of governments, companies, and banks."

Of course, employment and unemployment are not issues that can be understood by examining a society's top-level numbers. In times of economic crisis, it is well established that certain population segments experience disproportionate hardship. While the 2001 recession saw white unemployment increase from 3.5% to 5.2%; Black unemployment went from 7.6% to 10.8%. Less than a decade later, the Great Recession saw the median net worth of Black households in the United States drop over three times more (53% decrease) than for white households (17% decrease). Today, in the face of COVID-19, the trend persists, with the Department of Labor reporting that the Black unemployment rate went from 5.8% in February to 16.8% by June.

Black workers are not the only demographic group that has been sharply affected by the recent crisis; for example, Adams-Prassl et al. (2020) find that women and those without college degrees are also more likely to experience job losses. However, the comparison of racial groups' experiences has become particularly salient in the United States in recent months. In testimony before the U.S. House of Representatives, Human Rights Watch stated that "all levels of government in the U.S. are failing to protect Black and Brown people's basic rights, in ways that exacerbate their vulnerability to COVID-19." Against the backdrop of Black people in the United States dying at disproportionate rates due to the pandemic, high-profile displays of police violence have further underscored the realities of marginalization and discrimination across the country, fomenting unprecedented support for the Black Lives Matter movement. As a special report from *Scientific American* summarizes, "What began as a call to action in response to police violence and anti-Black racism in the U.S. is now a global initiative to confront racial inequities in society, including environmental injustice, bias in academia, and the public health threat of racism."

And so, perhaps now more than ever, employers, workers, and society are in need of solutions to overcome disparities in how opportunities and resources are allocated. While inequality in the labor market is certainly not a new problem, recent circumstances have led to an outpouring of corporate statements and initiatives aimed at mitigating it. As we argue in this brief, however, traditional people science is suboptimal for goals like diversity and inclusion, meaning it cannot provide the necessary foundation for these employers to support meaningful progress. The new people science, on the other hand, is well suited to rise to the occasion of the present moment.

CURRENT DISCOURSE ON REDEPLOYMENT

It is beyond the scope of this brief to grapple with all the questions facing today's employers and workers. However, a variant of the job models discussed in Section IV can help answer an important question: How can workers affected by COVID-19 be redeployed in an efficient and equitable manner? Before

presenting this use case, it is worth describing current discourse surrounding the issue in two parts: first, how workers can be reskilled; and second, how employers should approach evaluating workers for jobs.

The notion that large parts of the workforce will need to be reskilled, retrained, or upskilled in the aftermath of COVID-19 is fairly intuitive in light of the rapid growth seen in industries like healthcare and logistics, but it has not necessarily been a focus of policymakers in recent months. As Enders et al. note, "Many governments have focused on providing special unemployment benefits to laid-off workers. However, few programs have tried to train and entice workers to switch over to understaffed sectors of the economy." This lack of attention to retraining may perhaps be reflective of the fact that public sector programs have been relatively ineffective in the past, as described in Section IV of this brief. At the U.S. federal level, the primary response to the question of redeploying displaced workers has been White House support for an ad campaign called "Find Something New," which directs the unemployed to a website with links to job search and training resources. Launched by the Ad Council in July, the campaign has received funding from organizations such as the Department of Commerce, Apple, and IBM, though it was sharply criticized on social media for being "tone deaf" and "inadequate." At the state and local levels, a variety of technology companies and academic institutions have partnered with agencies serving the unemployed to dramatically increase access to online learning solutions.

On the note of how employers who are hiring should actually go about identifying and evaluating candidates in the post–COVID-19 era, concerted solutions are fairly rare. Some thought leaders have articulated the need to shift away from traditional credentials; in one statement regarding the Find Something New campaign, IBM Executive Chairperson Ginni Rometty reiterated her belief that "new collar" careers (jobs in a high-tech economy that do not require a four-year degree) are a crucial pathway to social mobility during times of economic transition. The Trump administration conveyed a similar sentiment in a June 26, 2020 Executive Order on government hiring practices, directing agencies to move away from degree requirements and toward "skills- and competency-based hiring" that "will hold the civil service to a higher standard." With respect to identifying more diverse candidates, many employers have also made broad claims about intentions to increase workforce diversity, and some concrete steps are being taken. Blackstone, for example, announced intentions to conduct on-campus recruiting at historically Black and women's colleges. PepsiCo similarly set a target to increase the number of Black people in managerial positions at the company by 30% by 2025, in addition to mandating company-wide anti-bias training.

To summarize the current state of workforce transitions today, displaced workers may have ample access to e-learning materials that could position them to acquire new types of roles, but minimal personalized guidance on how to navigate the myriad of options. Employers are invested in the notion of showing real gains on diversity and inclusion efforts, and while discourse suggests that they are open to abandoning conventional hiring strategies, technological solutions for doing so are not yet part of the picture. If

properly deployed, the new people science has the potential to address both of these shortcomings simultaneously.

AN EXPLORATION: MATCHING DISPLACED WORKERS TO HIGH-DEMAND ROLES BASED ON SOFT SKILLS

Background/Data

The following evidence again comes from pymetrics' models, this time built for *specific employers* who have experienced job losses. The goal is to demonstrate the relationship between a model that once evaluated candidates for an at-risk or declining job, such as an airline pilot, and models that evaluate fit to growing jobs. This could either be in a context like internal mobility (e.g., an airline company wants to retrain affected workers for a different role within the company) or off-boarding guidance for laid-off employees.

Two analyses demonstrate how the redirecting of workers across roles, using their underlying cognitive, social, and emotional traits, can work in practice. Specifically, three employers *(airline, retail,* and *hospitality industries)* who previously used pymetrics' models to evaluate candidates for three different roles *(pilot, retail salesperson,* and *front desk staff positions)* wanted to understand the redeployment prospects for their incumbents who now faced job losses due to COVID-19. The aviation industry employer was interested in pilots' alignment with four engineering roles, and the hospitality and retail employers were both interested in alignment with one digital marketing role. Typically, such an exercise would likely only incorporate a comparison of hard skills between the roles. However, this situation provided an opportunity to evaluate the additional insights provided by a fit-based analysis using soft skills.

Soft-skills comparison methodology

To conduct this evaluation, we took the three client-specific job models that were once used to evaluate candidates for the now-declining roles and identified the key cognitive, emotional, and social soft-skills associated with success in each position. We then mapped this information to the industry-level models presented in Section IV, which are used to evaluate fit to five high-growth roles. With this comparison, we determined the proportion of employees in now-declining jobs whose soft skills matched the soft skills associated with the high-growth jobs.

Hard-skills comparison methodology

In addition to the soft-skills comparison, we conducted a similar analysis on hard-skills gaps. This was done using data from a people analytics vendor called Burning Glass Technologies. The Boston-based company tracks millions of job listings across thousands of job boards and corporate websites to analyze the skills described in job descriptions for various occupations. Burning Glass is one of many indices that align jobs

with skills, but they are particularly involved in research on how the skills gap is disrupting modern job markets. Burning Glass's repository was used to obtain data on the top 50 skills that are most prevalent in each of the shrinking and growing roles (see Table 2), indexed by O*NET code.

We report the results for these investigations as two studies: the pilot vs. engineering comparisons (Study 1) and the front desk staff vs. digital marketing and retail sales vs. digital marketing comparisons (Study 2).

Results – Study 1: Airline client and engineering roles

Table 5 summarizes aggregate soft-skills and hard-skills overlap between airline pilots with the four engineering roles presented in Section IV: data science, front-end engineering, software development, and systems engineering. These results can be interpreted as the proportion of pilots that show a good fit to these engineering roles. While the hard-skills analysis alone would indicate that a fairly small percentage of pilots are well suited for any of the high-growth roles (average = 11%), the soft-skills analysis is much more optimistic about their prospects (average fit = 39%).

Table 5: Soft- and Hard-Skills Overlap between Pilot and Engineering Roles

PILOT: FIT TO 4 GROWING ROLES

ROLE	SOFT SKILLS FIT	HARD SKILLS FIT
Data Science	36%	8%
Front-End Engineering	48%	12%
Software Development	25%	13%
Systems Engineering	48%	10%

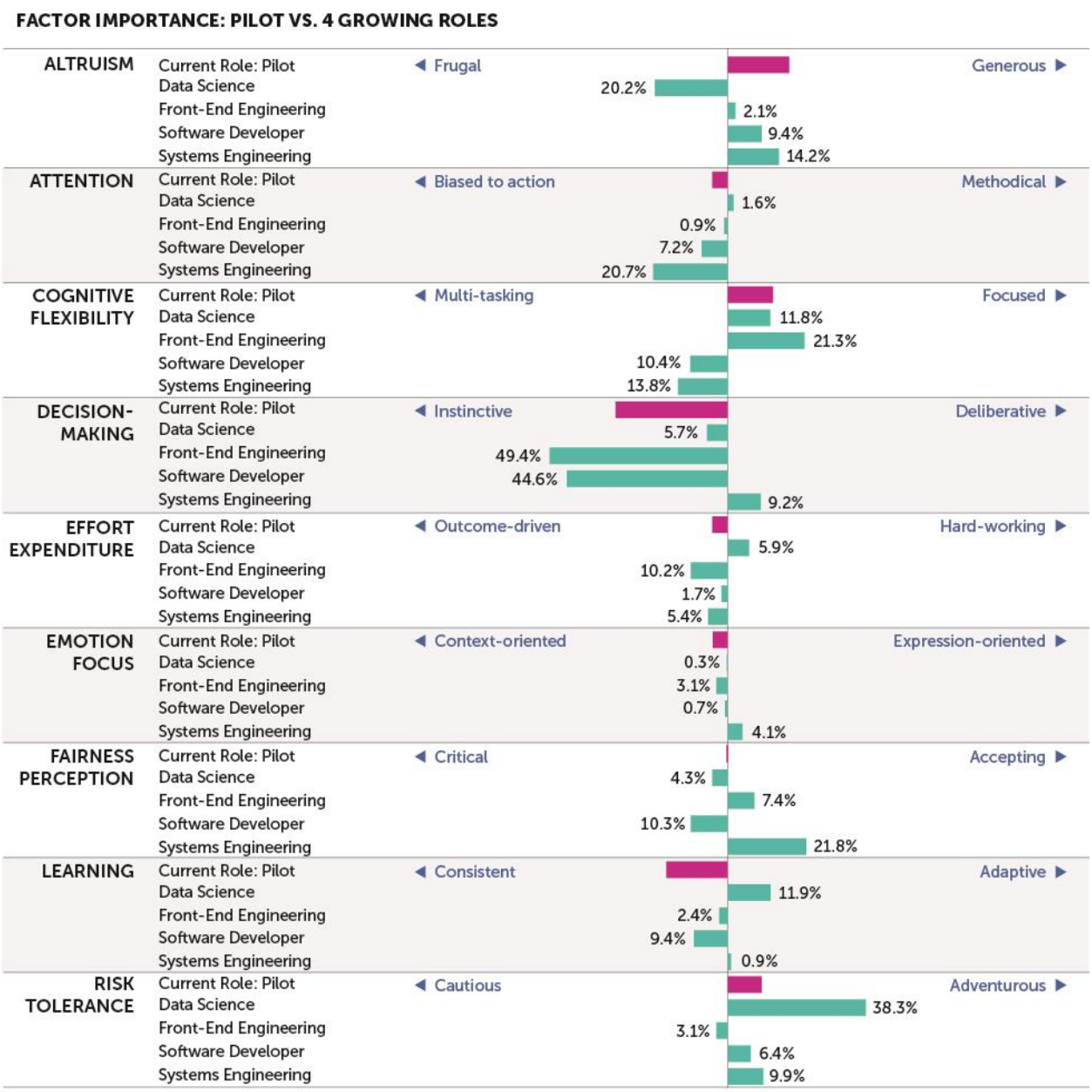
To better understand the nature of the respective hard- and soft-skills gaps, we compared each in Table 6 and Table 7. As shown in Table 6, pilots will inevitably require instruction to become proficient in Java, SQL, and Operating Systems if they are interested in pursuing one of these high-growth opportunities. However, for those whose underlying aptitudes suggest strong potential to thrive in a new industry, such a training investment may very well be worthwhile. As Table 7 suggests, the trait of instinctive decision-making is important for pilots, front-end engineers, and software developers alike.

Table 6: Hard-Skills Overlap between Pilot and Engineering Roles

COMPARING HARD SKILLS: PILOT VS. 4 GROWING ROLES

	SKILLS POSSESSED IN CURRENT ROLE	SKILLS REQUIRED FOR GROWING ROLES			
Hard Skills	Pilot	Data Science	Front-End Engineering	Systems Engineering	Software Developer
Business Process and Analysis	✓				
Business Strategy	✓				
Data Techniques	✓				
Java	GAP (4/4)				
Market Analysis	GAP (2/4)				
Operating Systems	GAP (4/4)				
People Management	✓				
Project Management	✓				
Quality Assurance and Control	✓				
Research Methodology	GAP (2/4)				
SQL Databases and Programming	GAP (4/4)				
Statistical Software	GAP (1/4)				
Technical Support	✓				
Web Analytics	GAP (1/4)				
Web Design	GAP (1/4)				

Table 7: Soft-Skills Comparison between Pilot Role and Engineering Roles



Results – Study 2: Digital marketing, front desk staff, and retail salespeople

Table 8 summarizes aggregate soft-skills and hard-skills overlap between front desk staff and digital marketing personnel, as well as retail salespeople and digital marketing personnel. While the disparity between soft-skills fit and hard-skills fit is smaller than in Study 1, it is notable that soft-skills fit is still higher for both of the shrinking roles. For example, nearly 1 in 3 retail salespeople demonstrate a soft-skills alignment with the digital marketing position, though only 1 in 6 are a hard-skills fit.

Table 8: Soft- and Hard-Skills Overlap

FRONT DESK STAFF AND RETAIL SALES ROLES: FIT TO DIGITAL MARKETING ROLE

ROLE	SOFT SKILLS FIT	HARD SKILLS FIT
Front Desk Staff	22%	18%
Retail Sales	30%	16%

As with Study 1, we also took a more granular look at the hard-skills (see Table 9) and soft-skills (see Table 10) overlap and gaps between hotel front desk staff and retail sales roles versus those in digital marketing. Compared to front desk staff, retail salespeople actually have an even larger number of hard-skills gaps to fill in order to work in the high-growth digital marketing role (e.g., Brand Management and Marketing Management in Table 9), again running counter to the findings of the soft-skills analysis. However, all three of these roles share in relevant soft-skills traits, such as an inclination toward generosity and risk tolerance (see Table 10).

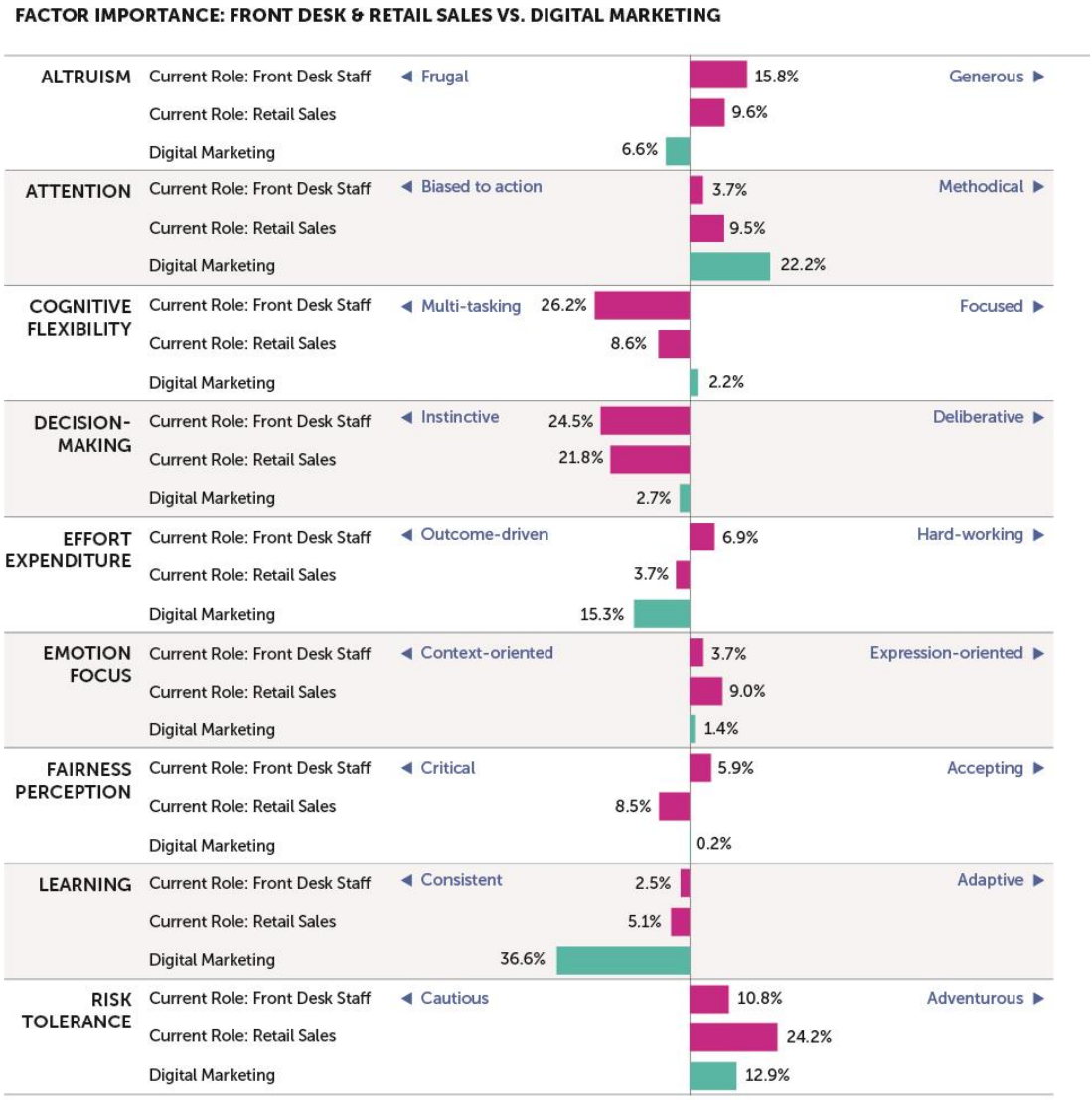
Table 9: Hard-Skills Overlap

COMPARING HARD SKILLS: FRONT DESK & RETAIL SALES VS. DIGITAL MARKETING

	SKILLS POSSESSED IN CURRENT ROLE		SKILLS REQUIRED FOR GROWING ROLES
Hard Skill	Front Desk Staff	Retail Sales	Digital Marketing
Account Management	GAP	GAP	
Administrative Support	✓	✓	
Advanced Customer Service	✓	✓	
Basic Customer Service	✓	✓	
Billing and Invoicing	✓	✓	
Brand Management	✓	GAP	
Business Strategy	GAP	GAP	
Content Development	GAP	GAP	
E-Commerce	GAP	GAP	
General Sales Practices	✓	✓	
Marketing Management	✓	GAP	
Media Strategy and Planning	GAP	GAP	
Merchandising	✓	✓	
People Management	✓	✓	
Product Management	GAP	GAP	
Web Development	GAP	GAP	

Table 10: Soft-Skills Comparison

Front Desk Staff and Retail Sales Roles vs. Digital Marketing Role



DISCUSSION

The above demonstrations reveal a few key reasons why the incorporation of soft skills in the context of redeployment is critical. The first and foremost reason is that hard-skills gap analysis alone cannot tell the whole story when workers are moving between non-skills-proximal industries or roles. *Skills proximity* is a concept that was well outlined in a report by Burning Glass Technologies and the World Economic Forum. This concept refers to the extent to which skills co-occur and have been found to "cluster" together in quantified analyses of labor markets; while important, the reality is that it is only one piece of the puzzle. For workers whose prior positions have all but evaporated—airport personnel, restaurant hosts/hostesses,

babysitters, massage therapists—a sole focus on skills proximity will likely produce no actionable insights about how they might fit into the future of work. In these cases, soft-skills analysis may very well be the only option to inform redeployment strategies. Notably, the behavioral assessment of soft skills becomes particularly important in these contexts because self-reports will inevitably miss associations between roles that are not obvious.

The second crucial benefit of soft skills for hiring in periods of economic transformation relates to social mobility and equality of opportunity. Because behavioral assessment data does not rely on any metrics that often restrict job candidates from better-paying jobs—such as degree requirements or employee referrals—the new people science can provide a rare chance for marginalized applicants to demonstrate their potential to succeed in a role. Study 2 serves as an example of this: While front desk staff and retail salespeople are paid less than digital marketing personnel on average, a considerable proportion of these at-risk lower-wage workers have a natural propensity to thrive in a digital marketing position. This alternative strategy of identifying candidates may be particularly useful for employers who are sincere in their intentions to increase the diversity of their workforce in the aftermath of COVID-19, particularly since a retail sales résumé would normally be overlooked by a traditional hiring process.

A final reason soft skills should be incorporated into the discourse around redeployment addresses the need to provide displaced workers with guidance on navigating the future of work. As Section IV of this brief summarizes, one-size-fits-all approaches to retraining fall short because they fail to account for differences between prospective workers. But on the other end of the spectrum, large investments in retraining and reskilling programs conducted via e-learning have created far too many options for trainees to meaningfully choose from. Behavioral assessment data provides the opportunity for displaced workers to optimize their retraining process by first identifying role types for which they are well suited. This can help mitigate undesirable and costly situations, such as a worker investing six months in an online course in cloud computing, only to realize that the resulting job is a bad fit for their personality and cognitive style.

Conclusion

The fact that the systems used to evaluate human potential have massive effects on society is both painfully obvious and remarkably forgotten in popular discourse. History has demonstrated that the consequences of ineffective employment selection tools include painful economic transitions and entrenched systemic inequalities, yet traditional hiring practices have remained largely undisrupted for decades. The goal of this brief has been to call attention to the need for fundamental change in the science of employment, fueled by pragmatic insights from cognitive science and related disciplines. A new people science is very

possible; and if deployed in a considered manner, it can provide the foundation for a dynamic and inclusive modern economy.

While the prospect of retraining the workforce is often viewed in terms of barriers—for example, what credentials or abilities are these people lacking?—the new people science can change this discourse to one of opportunities. The evaluation of job candidates in terms of their aptitudes is essentially a means of *optimizing* the redeployment of displaced and under-skilled individuals in an unprecedented manner. Behavioral assessments, backed by decades of well-established research, allow for the accurate and real-time measurement of soft skills that provide the basis for job matching. If humans can be evaluated in terms of their unique potential, and that potential can be aligned with the needs of employers, the net result can only be a more prosperous society.

Importantly, the new people science can also serve as a strategy for addressing intractable societal problems. Progress on racial inequality in the workforce has arguably never been more pressing than it is right now, and systems of allocating economic opportunities are an obvious piece of this complex puzzle. In the post–COVID-19 world, in order to effectively redirect talent and equip workers with useful skills, we must be able to separate true propensity for success from biased and irrelevant assessments. This is the promise of the new people science—revising what we measure, how we measure it, and how we think about those measurements—to prioritize both fairness and validity.

In many ways, the circumstance of a global pandemic simply accelerated changes in the workforce that were already being anticipated over the past decade. At the same time, if 2020 has taught us anything, it is that we cannot truly know what the next iteration of "the future of work" will entail. In order for any system of employment selection to remain relevant, it cannot be rooted in a particular context or technology. Instead, an impactful people science is one that allows for the redirection of entire jobs, industries, and workforces in an efficient and equitable manner. Accurate measurement of underlying soft skills is our best option for leveraging the full potential of human capability.

References

Adams-Prassl, A., Boneva, T., Golin, M., and Rauh, C. (2020). Inequality in the Impact of the Coronavirus Shock: Evidence from Real Time Surveys. IZA Discussion Paper No. 13183, Available at SSRN: https://ssrn.com/abstract=3590881.

Adler, P., Falk, C., Friedler, S.A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. (2018). Auditing Black-Box Models for Indirect Influence. *Knowledge and Information Systems*, 54(1), 95–122.

Agarwal, A. (2018). Data Reveals Why The "Soft" In "Soft Skills" Is A Major Misnomer. *Forbes*. Retrieved from forbes.com/sites/anantagarwal/2018/10/02/data-reveals-why-the-soft-in-soft-skills-is-a-major-misnomer/#3807b15f6f7b.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H. (2018). A Reductions Approach to Fair Classification. Proceedings of the 35th International Conference on Machine Learning, PMLR, 80, 60–69.

Alabdulkareem, A., Frank, M.R., Sun, L., AlShebli, B., Hidalgo, C., and Rahwan, I. (2018). Unpacking the Polarization of Workplace Skills. *Science Advances*, 4(7), eaao6030.

Algera, J. A., Jansen, P. G., Roe, R. A., & Vijn, P. (1984). Validity generalization: Some critical remarks on the Schmidt-Hunter procedure. *Journal of Occupational Psychology, 57*(3), 197-210.

Alon, T., Doepke, M., Olmstead-Rumsey, J., and Tertilt, M. (April 2020). The Impact of COVID-19 on Gender Equality (No. w26947). National Bureau of Economic Research.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). The Standards for Educational and Psychological Testing. American Educational Research Association.

Atkeson, A. (2020). What Will Be the Economic Impact of COVID-19 in the U.S.? Rough Estimates of Disease Scenarios (No. w26867). National Bureau of Economic Research.

Austin, A. (2008). What a Recession Means for Black America. Economic Policy Institute.

Balcar, J. (2014). Soft Skills and Their Wage Returns: Overview of Empirical Literature. *Review of Economic Perspectives*, 14(1), 3–15.

Barrero, J.M., Bloom, N., and Davis, S.J. (2020). COVID-19 Is Also a Reallocation Shock (No. w27137). National Bureau of Economic Research.

Bear, M., Connors, B., and Paradiso, M.A. (2020). *Neuroscience: Exploring the Brain, Enhanced Fourth Edition*. Jones & Bartlett Learning, LLC.

Behaghel, L., Crépon, B., and Le Barbanchon, T. (2015). Unintended Effects of Anonymous Résumés. *American Economic Journal: Applied Economics*, 7(3), 1–27.

Behroozi, M., Shirolkar, S., Barik, T., and Parnin, C. Does Stress Impact Technical Interview Performance? Retrieved from http://chrisparnin.me/pdf/stress_FSE_20.pdf.

Bemis, S.E. (1968). Occupational Validity of the General Aptitude Test Battery. *Journal of Applied Psychology*, 52(3), 240–244.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1), 122–142.

Bersin, J. (2019). Let's Stop Talking About Soft Skills: They're Power Skills. Retrieved from https://joshbersin.com/2019/10/lets-stop-talking-about-soft-skills-theyre-power-skills/.

Bhattacharya, J., and Petsche, H. (2005). Drawing on mind's canvas: Differences in cortical integration patterns between artists and non-artists. *Human Brain Mapping*, 26(1), 1–14.

The Black Lives Matter Movement (Special Report). (2020). *Scientific American*. Retrieved https://www.scientificamerican.com/report/the-black-lives-matter-movement/.

Bluestone, P., Chike, E., and Wallace, S. (2020). The Future of Industry and Employment: COVID-19 Effects Exacerbate the March of Artificial Intelligence. Center for State and Local Finance Brief. (Pub. No. 60). Andrew Young School of Policy Studies. Retrieved from https://cslf.gsu.edu/download/covid-19-ai/?wpdmdl=6496041&refresh=5ea830afd2a471588080815.

Bosco, F., Allen, D.G., and Singh, K. (2015). Executive Attention: An Alternative Perspective on General Mental Ability, Performance, and Subgroup Differences. *Personnel Psychology*, 68(4), 859–898.

Brown, T.A. (2015). *Confirmatory Factor Analysis for Applied Research (Second Edition)*. Guilford Press.

Browne, M.W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44(1), 108–132.

Buchanan, L., Bui, Q., and Patel, J.K. (2020). Black Lives Matter May Be the Largest Movement in U.S. History. *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html.

Buckland, M., and Gey, F. (1994). The Relationship Between Recall and Precision. *Journal of the American Society for Information Science*, 45(1), 12–19.

Buhrmester, M., Kwang, T., and Gosling, S.D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1), 3–5.

Cable, D.M., and Judge, T.A. (August 1995). The Role of Person–Organization Fit in Organizational Selection Decisions. Paper presented at the Academy of Management Annual Meetings, Vancouver, Canada.

Cabrera, M.A.M., and Nguyen, N.T. (2001). Situational Tests: A Review of Practice and Constructs Assessed. *International Journal of Selection and Assessment*, 9(1–2), 103–113.

Caldwell, D.F., and O'Reilly, C.A. III. (1990). Measuring Person–Job Fit Using a Profile-Comparison Process. *Journal of Applied Psychology*, 75(6), 648–657.

Camerer, C. (2011). The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List. Retrieved from https://ssrn.com/abstract=1977749.

Carney, M. (2020). Mark Carney on how the economy must yield to human values. *The Economist*. Retrieved from economist.com/by-invitation/2020/04/16/mark-carney-on-how-the-economy-must-yield-to-human-values.

Carpenter, J., and Myers, C.K. (2010). Why volunteer? Evidence on the role of altruism, image, and incentives. *Journal of Public Economics*, 94(11–12), 911–920.

Casaletto, K.B., and Heaton, R.K. (2017). Neuropsychological Assessment: Past and Future. *Journal of the International Neuropsychological Society*, 23(9–10), 778–790.

Chapman, B.P., Duberstein, P.R., Sörensen, S., and Lyness, J.M. (2007). Gender Differences in Five Factor Model Personality Traits in an Elderly Cohort: Extension of Robust and Surprising Findings to an Older Generation. *Personality and Individual Differences*, 43(6), 1594–1603.

Chen, L., Mislove, A., and Wilson, C. (2015). Peeking Beneath the Hood of Uber. In IMC '15: Proceedings of the 2015 Internet Measurement Conference (pp. 495–508).

Chen, C., Chen, Y., Hsu, P-H., and Podolski, E.J. (2016). Be nice to your innovators: Employee treatment and corporate innovation performance. *Journal of Corporate Finance*, 39, 78–98.

Chetty, R., Friedman, J., Saez, E., Turner, N., and Yagan, D. (2017). Mobility Report Cards: The Role of Colleges in Intergenerational Mobility (No. w23618). National Bureau of Economic Research.

Christiansen, N., Sliter, M., and Frost, C.T. (2014). What employees dislike about their jobs: Relationship between personality-based fit and work satisfaction. *Personality and Individual Differences*, 71, 25–29.

Cimatti, B. (2016). Definition, Development, Assessment of Soft Skills and Their Role for the Quality of Organizations and Enterprises. *International Journal for Quality Research*, 10(1) 97–130.

Cirillo, P., and Taleb, N.N. (2020). Tail risk of contagious diseases. *Nature Physics*, 16, 606–613.

Cohen, P. (2020). Many Jobs May Vanish Forever as Layoffs Mount. *The New York Times*. Retrieved from nytimes.com/2020/05/21/business/economy/coronavirus-unemployment-claims.html.

Coleman, J. (2020). White House campaign advocates new "pathways" to jobs amid pandemic. *The Hill*. Retrieved from thehill.com/homenews/administration/507203-white-house-campaign-advocates-new-pathways-to-jobs-amid-pandemic.

Cook, D.A., and Beckman, T.J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, 119(2), 166-e7.

Cronbach, L.J., and Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*. 52(4): 281–302.

Daniel, R.P. (1932). Basic considerations for valid interpretations of experimental studies pertaining to racial differences. *Journal of Educational Psychology*, 23(1), 15–27.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Deb, T. (2009). *Managing Human Resource and Industrial Relations*. Excel Books.

Debelak, R., Egle, J., Köstering, L., and Kaller, C.P. (2016). Assessment of planning ability: Psychometric analyses on the unidimensionality and construct validity of the Tower of London Task (TOL-F). *Neuropsychology*, 30(3), 346–360.

Deming, D.J. (2017). The value of soft skills in the labor market. *NBER Reporter*, (4), 7–11.

Deniz, N., Noyan, A., and Ertosun, Ö.G. (2015). Linking person-job fit to job stress: The mediating effect of perceived person-organization fit. *Procedia—Social and Behavioral Sciences*, 207, 369–376.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, 64, 135–168.

Dykes, M. (2020). US Court Ruling: You Can Be "Too Smart" to Be a Cop. Global Research. Retrieved from globalresearch.ca/us-court-ruled-you-can-be-too-smart-to-be-a-cop/5420630.

Edwards, A.L. (1982). *The Social Desirability Variable in Personality Assessment and Research*. Greenwood Press.

Enders, A., Haggstrom, L., and Lalive, R. (2020). How Reskilling Can Soften the Economic Blow of Covid-19. *Harvard Business Review*. Retrieved from hbr.org/2020/06/how-reskilling-can-soften-the-economic-blow-of-covid-19.

Enforcement and Litigation Statistics. (n.d.). U.S. Equal Employment Opportunity Commission. Retrieved July 19, 2020, from https://www.eeoc.gov/statistics/enforcement-and-litigation-statistics.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14(4), 583–610.

Eriksen, B.A., and Eriksen, C.W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics*, 16, 143–149.

European Union Independent High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI.

Eysenck, M.W., Hunt, E.B., Ellis, A., and Johnson-Laird, PN. (1994). *The Blackwell Dictionary of Cognitive Psychology*. Wiley.

Fadulu, L. (2018). Why Is the U.S. So Bad at Worker Retraining? *The Atlantic*. Retrieved from theatlantic.com/education/archive/2018/01/why-is-the-us-so-bad-at-protecting-workers-from-automation/549185/.

Famighetti, C., and Hamilton, D. (2019). The Great Recession, education, race, and homeownership. Economic Policy Institute. Retrieved from epi.org/blog/the-great-recession-education-race-and-homeownership/.

Fatourechi, M., Ward, R.K., Mason, S.G., Huggins, J., Schlögl, A., and Birch, G.E. (December 2008). Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets. In 2008 Seventh International Conference on Machine Learning and Applications (pp. 777–782). IEEE.

Fiarman, S.E. (2016). Unconscious Bias: When Good Intentions Aren't Enough. *Educational Leadership*, 74(3), 10–15.

Fischer, C.S., Voss, K., Swidler, A., Lucas, S.R., Jankowski, M.S., and Hout, M. (1996). *Inequality by Design: Cracking the Bell Curve Myth*. Princeton University Press.

Forrester. (2019). The Future of Work Is an Adaptive Workforce. *Forbes*. Retrieved from forbes.com/sites/forrester/2019/08/01/the-future-of-work-is-an-adaptive-workforce/#31bc53aa3fa2.

Forsythe, R., Horowitz, J. L., Savin, N., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior, 6*(3), 347-369.

Frankish, K., and Ramsey, W.M. (2012). *The Cambridge Handbook of Cognitive Science*. Cambridge University Press.

Franzen, A., and Pointner, S. (2013). The external validity of giving in the dictator game: A field experiment using the misdirected letter technique. *Experimental Economics*, 16(2), 155–169.

2017 Future of Work Outlook Survey. (2017). Future Enterprise. Retrieved from futureenterprise.com/blog/2017/2/2/future-of-work-survey.

Gardiner, E., and Jackson, C.J. (2012). Workplace mavericks: How personality and risk-taking propensity predicts maverickism. *British Journal of Psychology*, 103(4), 497–519.

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books.

Germine, L., Nakayama, K., Duchaine, B.C., Chabris, C.F., Chatterjee, G., and Wilmer, J.B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin Review*, 19(5), 847–857.

Geronimus, A.T., and Bound, J. (1998). Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples. *American Journal of Epidemiology*, 148(5), 475–486.

Goldstein, H.W., Zedeck, S., and Goldstein, I.L. (2002). g: Is This Your Final Answer? *Human Performance*, 15(1–2), 123–142.

Goodman, P.S., Cohen, P., and Chaundler, R. (2020). 'It oddly hasn't been a stressful time': European nations handled COVID-19 unemployment differently than the United States. Take a look. *Chicago Tribune*. Retrieved from chicagotribune.com/coronavirus/ct-nw-nyt-europe-unemployment-rate-covid-19-20200704-ihuzdo2kbngdxjnoxqxymejhj4-story.html.

Gottfried, M. (2020). Blackstone Revamps Approach on Recruiting Process to Aid Diversity. *The Wall Street Journal* (6/25/20).

Gould, S.J. (1981). *The Mismeasure of Man*. W. W. Norton & Company.

Guterres, A. (2020). The recovery from the COVID-19 crisis must lead to a different economy. United Nations. Retrieved from un.org/en/un-coronavirus-communications-team/launch-report-socio-economic-impacts-covid-19.

Handbook of Human Factors and Ergonomics. (2012). Salvendy, G. (ed.) John Wiley & Sons.

The Handbook of Work Analysis: Methods, Systems, Applications and Science of Work Measurement in Organizations. (2012). Wilson, M.A., Bennett, W. Jr., Gibson, S.G. and Alliger, GM (eds.). Routledge.

Hardt, M., Price, E., Srebo, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 3315–3323.

Harter, J.K., Schmidt, F.L., and Hayes, T.L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. *Journal of Applied Psychology*, 87(2), 268–279.

Hartman, M. (2020). COVID-19 wreaks economic havoc, spurs health care hiring. Marketplace. Retrieved from marketplace.org/2020/03/13/demand-health-care-workers-up-covid-19/.

Heckman, J.J., and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4), 451–464.

Heller, D. (2019). Work experience poor predictor of future job performance. Florida State University. Retrieved from phys.org/news/2019-05-poor-predictor-future-job.html.

Herring, C. (2009). Does diversity pay?: Race, gender, and the business case for diversity. *American Sociological Review*, 74(2), 208–224.

Herrnstein, R., and Murray, C. (1994). *The Bell Curve*. Free Press.

Holger, D. (2019). The Business Case for More Diversity. *Wall Street Journal*. Retrieved from wsj.com/articles/the-business-case-for-more-diversity-11572091200.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., and Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems (pp. 1–16).

Hossin, M., and Sulaiman, M.N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 1.

House, R.J., Filley, A.C., and Gujarati, D.N. (1971). Leadership style, hierarchical influence, and the satisfaction of subordinate role expectations: A test of Likert's influence proposition. *Journal of Applied Psychology*, 55(5), 422–432.

Human Rights Watch. (2020). US: Covid-19 Disparities Reflect Structural Racism, Abuses. Retrieved from https://www.hrw.org/news/2020/06/10/us-covid-19-disparities-reflect-structural-racism-abuses.

Hunt, V., Prince, S., Dixon-Fyle, S., and Yee, L. (2018). Delivering through diversity. McKinsey & Company Report. Retrieved April, 3, 2018.

Hunter, J.E., and Schmidt, F.L. (1982). Ability tests: Economic benefits versus the issue of fairness. *Industrial Relations: A Journal of Economy and Society*, 21(3), 293–308.

Hunter, J.E., Schmidt, F.L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. SAGE Publications.

Hutchinson, B., and Mitchell, M. (2019). 50 Years of Test (Un)fairness: Lessons for Machine Learning. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).

Ivancevich, J.M., and Donnelly Jr., J.H. (1975). Relation of Organizational Structure to Job Satisfaction, Anxiety-Stress, and Performance. *Administrative Science Quarterly*, 20(2), 272–280.

Jacoby, S.M. (2008). Employee attitude surveys in historical perspective. *Industrial Relations: A Journal of Economy and Society*, 27(1), 74–93.

Jencks, C., and Phillips, M. (1998). *The Black-White Test Score Gap: Why It Persists and What Can Be Done*. Brookings Institution Press.

Jensen, A.R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 39(1), 1–123.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.

Judge, T.A., and Bretz Jr., R.D. (1994). Person–organization fit and the theory of work adjustment: Implications for satisfaction, tenure, and career success. *Journal of Vocational Behavior*, 44(1), 32–54.

Judge, T.A., Thoresen, C.J., Bono, J.E., and Patton, G.K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, 127(3), 376–407.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kaller, C.P., Unterrainer, J.M., and Stahl, C. (2012). Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. *Psychological Assessment*, 24(1), 46.

Karsten, J., Penninx, B.W., Riese, H., Ormel, J., Nolen, W.A., and Hartman, C.A. (2012). The state effect of depressive and anxiety disorders on big five personality traits. *Journal of Psychiatric Research*, 46(5), 644–650.

Kearns, M., Neel, S., Roth, A., and Wu, Z.S. (January 2019). An empirical study of rich subgroup fairness for machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 100–109).

Kim, Y., Huang, J., and Emery, S. (2016). Garbage in, garbage out: Data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of Medical Internet Research*, 18(2), e41.

Ko, L.W., Komarov, O., Hairston, W.D., Jung, T.P., and Lin, C.T. (2017). Sustained Attention in Real Classroom Settings: An EEG Study. *Frontiers in Human Neuroscience*, 11, 388.

Kodrzycki, Y.K. (1997). Training programs for displaced workers: What do they accomplish? *New England Economic Review*, 39–59.

Kohavi, R. (August 1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* 2, 1137–1145.

Kolb, B., and Whishaw, I.Q. (2009). *Fundamentals of Human Neuropsychology*. Worth Publishers.

Konradt, U., Garbers, Y., Böge, M., Erdogan, B., and Bauer, T.N. (2016). Antecedents and Consequences of Procedural Fairness Perceptions in Personnel Selection: A Three-Year Longitudinal Study. *Business Faculty Publications and Presentations. 54*.

Koppenol-Gonzalez, G.V., Bouwmeester, S., and Boonstra, A.M. (2010). Understanding planning ability measured by the Tower of London: An evaluation of its internal structure by latent variable modeling. *Psychological Assessment*, 22(4), 923.

Koppes, L.L. (Ed.). (2014). *Historical Perspectives in Industrial and Organizational Psychology*. Psychology Press.

Koys, D.J. (2001). The effects of employee satisfaction, organizational citizenship behavior, and turnover on organizational effectiveness: A unit-level, longitudinal study. *Personnel Psychology*, *54*(1), 101–114.

Kozlowski, S.W. (2012). *The Oxford Handbook of Organizational Psychology*. Oxford University Press.

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 94, 1372–1381.

Kristof, A.L. (1996). Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology*, *49*(1), 1–49.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Landers, R.N. (2015). An introduction to game-based assessment: Frameworks for the measurement of knowledge, skills, abilities and other human characteristics using behaviors observed within videogames. *International Journal of Gaming and Computer-Mediated Simulations*, *7*(4), iv-viii.

Landy, F.J., and Conte, J.M. (2017). *Work in the 21st Century: An Introduction to Industrial and Organizational Psychology*. John Wiley & Sons.

LeBreton, J.M., Schoen, J.L., and James, L.R. (2017). Situational Specificity, Validity Generalization, and the Future of Psychometric Meta-analysis. In J.L. Farr and N.T. Tippins (Eds.), *Handbook of Employee Selection,* Second Edition (pp. 93–114). Routledge, Taylor & Francis Group.

Lee, J. (2009). Partnerships with industry for efficient and effective implementation of TVET. *International Journal of Vocational Education and Training*.

Leigh, D.E. (1990). Does Training Work for Displaced Workers? A Survey of Existing Evidence. W.E. Upjohn Institute for Employment Research.

Lejuez, C.W., Read, J.P., Kahler, C.W., Richards, J.B., Ramsey, S.E., Stuart, G.L., Ramsey, S.E., Strong, D.R., and Brown, R.A. (2002). Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology Applied*, 8(2), *75*–84.

Lezak, M.D., Howieson, D.B., Loring, D.W., Hannay, H.J., and Fischer, J.S. (2004). *Neuropsychological Assessment*. Oxford University Press.

Lin, W.L., and Yao, G. (2014). Concurrent Validity. In: Michalos, A.C. (ed.), *Encyclopedia of Quality of Life and Well-Being Research*. Springer, Dordrecht.

Link, H.C. (1919). *Employment Psychology: The Application of Scientific Methods to the Selection, Training, and Grading of Employees*. Macmillan.

Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

MacCoby, E.E., and Jacklin, C.N. (1991). *The Psychology of Sex Differences*. Stanford University Press.

MacGillis, A. (2020). How Germany Saved Its Workforce from Unemployment While Spending Less Per Person Than the U.S. ProPublica. Retrieved from propublica.org/article/how-germany-saved-its-workforce-from-unemployment-while-spending-less-per-person-than-the-u-s.

Magrass, Y., and Upchurch, R.L. (1988). Computer literacy: People adapted for technology. *ACM SIGCAS Computers and Society*, 18(2), 8–15.

MarketScreener. Retrieved from marketscreener.com/THE-BLACKSTONE-GROUP-INC-60951400/news/Blackstone-Revamps-Approach-On-Recruiting-Process-to-Aid-Diversity-WSJ-30822029/.

Mattern, K.D., Patterson, B.F., Shaw, E.J., Kobrin, J.L., and Barbuti, S.M. (2008). Differential Validity and Prediction of the SAT®. Research Report No. 2008-4. College Board.

Mayfield, J.W. (1997). Black-white differences in memory test performance among children and adolescents. *Archives of Clinical Neuropsychology*, 12(2), 111–122.

McChesney, J., Roberts, Z., Dolphin, J., and Thissen-Roe, A. (2020). Relationships Between Personality & Behavior in Employment Games. Poster presented at the Society for Industrial-Organizational Psychology 35th Annual Conference, Austin, TX.

McKnight, D.H., and Chervany, N.L. (2000). What is trust? A conceptual analysis and an interdisciplinary model. AMCIS 2000 Proceedings, 382.

McLaren, S. (2019). Candidates' Soft Skills Are Notoriously Hard to Assess, But Following These 6 Steps Will Help. LinkedIn Talent Solutions. Retrieved from business.linkedin.com/talent-solutions/blog/recruiting-strategy/2019/soft-skills-are-hard-to-assess-but-these-6-steps-can-help.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.

Miller, S., and Hughes, D. (2017). The Quant Crunch: How the demand for data science skills is disrupting the job market. Burning Glass Technologies. http://www.burning-glass.com/wp-content/uploads/The_Quant_Crunch.pdf.

Mitchell, A.G. (1998). Strategic training partnerships between the State and enterprises. International Labour Organization. Geneva.

Mitroff, S.R., Biggs, A.T., Adamo, S.H., Dowd, E.W., Winkle, J., and Clark, K. (2015). What can 1 billion trials tell us about visual search? *Journal of Experimental Psychology: Human Perception and Performance*, 41(1), 1–5.

Morath, E., and Omeokwe, A. (2020). Coronavirus Obliterated Best African-American Job Market on Record. *The Wall Street Journal*. Retrieved from wsj.com/articles/coronavirus-obliterated-best-african-american-job-market-on-record-11591714755.

Muhlhausen, D. (2017). Federal Job Training Fails Again. The Heritage Foundation.

Mullainathan, S. (2019). Biased Algorithms Are Easier to Fix Than Biased People. *The New York Times*. Retrieved from https://nyti.ms/38brSto.

Mullainathan, S., and Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, 107(5), 476–480.

Müller, V.C. (2005). Ethics of Artificial Intelligence and Robotics. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). Metaphysics Research Lab, Stanford University.

Muro, M., Maxim, R., and Whiton, J. (2020). The robots are ready as the COVID-19 recession spreads. Brookings. Retrieved from brookings.edu/blog/the-avenue/2020/03/24/the-robots-are-ready-as-the-covid-19-recession-spreads/.

Murphy, K.R. (2013). *Validity Generalization: A Critical Review*. Taylor & Francis.

Murphy, K.R. (2008). Explaining the Weak Relationship Between Job Performance and Ratings of Job Performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1(2), 148–160.

Murphy, K.R., and Shiarella, A.H. (1997). Implications of the Multidimensional Nature of Job Performance for the Validity of Selection Tests: Multivariate Frameworks for Studying Test Validity. *Personnel Psychology*, 50(4), 823–854.

Naef, M., and Schupp, J. (2009). Measuring Trust: Experiments and Surveys in Contrast and Combination. IZA, DP No. 487.

Nagel, T. (2016). *The Possibility of Altruism*. Princeton University Press.

National Center for O*NET Development. O*NET OnLine. Retrieved from https://www.onetonline.org/.

Nietzel, M.T. (2020). The Latest from Coursera: Free Courses for Newly Unemployed Workers Across the Globe. *Forbes*. Retrieved from forbes.com/sites/michaeltnietzel/2020/04/24/the-latest-from-coursera-free-courses-for-newly-unemployed-workers-across-the-globe/#26dcc19f6546.

Nye, C., Su, R., Rounds, J., and Drasgow, F. (2012). Vocational Interests and Performance: A Quantitative Summary of Over 60 Years of Research. *Perspectives on Psychological Science*, 7(4), 384–403.

O'Donnell, R. (2018). Eye tracking study shows recruiters look at résumés for 7 seconds. HR Dive. Retrieved from hrdive.com/news/eye-tracking-study-shows-recruiters-look-at-resumes-for-7-seconds/541582/.

Organisation for Economic Co-operation and Development Committee on Digital Economy Policy. (2019). Recommendation of the Council on Artificial Intelligence. Retrieved from: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Osoba, O.A., Boudreaux, B., Saunders, J.M., Irwin, J.L., Mueller, P.A., and Cherney, S. (2019). Algorithmic equity: A framework for social applications. Santa Monica, CA: RAND.

Outtz, J.L. (2002). The role of cognitive ability tests in employment selection. *Human Performance*, 15(1-2), 161–172.

Paulhus, D.L., and Vazire, S. (2007). The self-report method. *Handbook of Research Methods in Personality Psychology*, 1, 224–239.

Personal Data Protection Commission Singapore. (2019). A Proposed Model AI Governance Framework for Public Consultation.

Petitto, L.A., and Dunbar, K. (October 2004). New findings from Educational Neuroscience on Bilingual Brains, Scientific Brains, and the Educated Mind. In Conference on Building Usable Knowledge in Mind, Brain, & Education. Harvard Graduate School of Education (pp. 1–20).

Petty, M.M., McGee, G.W., and Cavender, J.W. (1984). A meta-analysis of the relationships between individual job satisfaction and individual performance. *The Academy of Management Review*, 9(4), 712–721.

Pizzagalli, D.A., Iosifescu, D., Hallett, L.A., Ratner, K.G., and Fava, M. (2008). Reduced hedonic capacity in major depressive disorder: Evidence from a probabilistic reward task. *Journal of Psychiatric Research*, 43(1), 76–87.

Ployhart, R.E., and Holtz, B.C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, 61(1), 153–172.

Polli, F., Trindel, K., Baker, L, and Pettiford, A. (2019). Technical Brief for pymetrics.

Powers, D.M. (2011). Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

pymetrics, inc. (in prep). Behavioral Factors Linked to ONET Knowledge, Skills and Abilities.

pymetrics, inc. (2020). aduit-AI: How we use it and what it does. Retrieved from https://github.com/pymetrics/audit-ai/blob/master/examples/implementation_suggestions.md.

Quillian, L., Pager, D. Hexel, O., and Midtbøen, A.H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, 114(41).

Rainwater Jr., J.H., Michael, W.B., and Stewart, R. (1963). Predictive Validity of Mental Ability Tests for Selecting Clerical Employees. *Psychological Reports*, 12(2), 435–438.

Ranosa, R. (2020). These jobs are growing despite COVID-19. *Human Resources Director*. Retrieved from hcamag.com/us/news/general/these-jobs-are-growing-despite-covid-19/219859.

Revelle, W., Condon, D.M., and Wilt, J. (2011). Methodological advances in differential psychology. In T. Chamorro-Premuzic, S. von Stumm, and A. Furnham (Eds.), *The Wiley-Blackwell Handbooks of Personality and Individual Differences*. (pp. 39–73). Wiley-Blackwell.

Richardson J.T.E. (2007). Measures of short-term memory: A historical review. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, 43(5), 635–650.

Richardson, K., and Norgate, S. H. (2015). Does IQ Really Predict Job Performance? *Applied Developmental Science*, 19(3), 153–169.

Sackett, P.R., and Ellingson, J.E. (1997). The Effects of Forming Multi-Predictor Composites on Group Differences and Adverse Impact. *Personnel Psychology*, 50(3), 707–721.

Sackett, P.R., and Walmsley, P.T. (2014). Which Personality Attributes Are Most Important in the Workplace? *Perspectives on Psychological Science*, 9(5), 538–551.

Sadler, P.J. (1970). Leadership style, confidence in management, and job satisfaction. *The Journal of Applied Behavioral Science*, 6(1), 3–19.

Santesso, D.L., Dillon, D.G., Birk, J.L., Holmes, A.J., Goetz, E., Bogdan, R., and Pizzagalli, D.A. (2008). Individual differences in reinforcement learning: Behavioral, electrophysiological, and neuroimaging correlates. *Neuroimage*, 42(2), 807–816.

Sarter, M., Givens, B., and Bruno, J.P. (2001). The cognitive neuroscience of sustained attention: Where top-down meets bottom-up. *Brain Research Reviews*, 35(2), 146–160.

The SAT: Practice Test #1. (2016). The College Board. Retrieved from https://collegereadiness.collegeboard.org/pdf/sat-practice-test-1.pdf.

Schmidt, F.L., and Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262–274.

Schneider, W.J., and McGrew, K.S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In D.P. Flanagan and E.M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 73–163). Guilford Press.

Schneider, W.J., and Newman, D.A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review*, 25(1), 12–27.

Schulam, P., and Saria, S. (April 2019). Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. In The 22nd International Conference on Artificial Intelligence and Statistics (pp. 1022–1031).

Scott, W.D. (1911). *Increasing Human Efficiency in Business: A Contribution to the Psychology of Business*. Macmillan Press.

Sekiguchi, T. (2004). Person-organization fit and person-job fit in employee selection: A review of the literature. *Osaka Keidai Ronshu*, 54(6), 179–196.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422), 486–494.

Shearer, C.B., and Karanian, J.M. (2017). The Neuroscience of Intelligence: Empirical Support for the Theory of Multiple Intelligences? *Trends in Neuroscience and Education*, 6, 211–223.

Siddique, C.M. (2004). Job analysis: A strategic human resource management practice. *The International Journal of Human Resource Management*, 15(1), 219–244.

Singer, M.S., and Singer, A.E. (1986). Relation between transformational vs. transactional leadership preference and subordinates' personality: An exploratory study. *Perceptual and Motor Skills*, 62(3), 775–780.

Singh, N., and Krishnan, V.R. (2007). Transformational leadership in India: Developing and validating a new scale using grounded theory approach. *International Journal of Cross-Cultural Management*, 7(2), 219–236.

Slaughter, J.E., Christian, M.S., Podsakoff, N.P., Sinar, E.F., and Lievens, F. (2013). On the limitations of using situational judgment tests to measure interpersonal skills: The moderating influence of employee anger. *Personnel Psychology*, 67(4), 847–885.

Smith, B. (2020). Microsoft launches initiative to help 25 million people worldwide acquire the digital skills needed in a COVID-19 economy. Microsoft. Press Release. Retrieved from blogs.microsoft.com/blog/2020/06/30/microsoft-launches-initiative-to-help-25-million-people-worldwide-acquire-the-digital-skills-needed-in-a-covid-19-economy/.

Smith, C.S. (2020). Dealing with Bias in Artificial Intelligence. *The New York Times*. Retrieved from https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html.

Sommers, D., and Austin, J. (2002). Using O*NET in Dislocated Worker Retraining: The Toledo Dislocated Worker Consortium Project. Center on Education and Training for Employment College of Education, The Ohio State University.

Stevens, P. (2020). Companies are making bold promises about greater diversity, but there's a long way to go. CNBC. Retrieved from cnbc.com/2020/06/11/companies-are-making-bold-promises-about-greater-diversity-theres-a-long-way-to-go.html.

Superville, D. (2020). White House-backed campaign pushes alternate career paths. *Associated Press*. Retrieved from apnews.com/32959d751de0f9cc327a92ff60a49b20.

Taylor, F.W. (1919). *The Principles of Scientific Management*. Harper & Brothers.

Tead, O., and Metcalf, H.C. (1920). *Personnel Administration: Its Principles and Practice*, Issue 18. McGraw-Hill Book Company, Incorporated.

Tett, R.P., Jackson, D.N., and Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44(4), 703–742.

Thorndike, E.L. (1918). Individual differences. *Psychological Bulletin*, 15(5), 148–159.

Treadway, M.T., Buckholtz, J.W., Schwartzman, A.N., Lambert, W.E., and Zald, D.H. (2009). Worth the 'EEfRT'? The effort expenditure for rewards task as an objective measure of motivation and anhedonia. *PLOS One*, 4(8), e6598.

Tukey, J.W. (1953). The problem of multiple comparisons. Mimeographed notes, Princeton University.

Tulchinsky, A. (2019). Why Explainable AI (XAI) is the Future of Marketing and e-Commerce. The Future of Customer Engagement and Experience. Retrieved from https://www.the-future-of-commerce.com/2019/03/11/what-is-explainable-ai-xai/.

Umebayashi, K., and Okita, T. (2010). An ERP investigation of task switching using a flanker paradigm. *Brain Research*, 1346, 165–173.

United Nations, Economic Development. (2020). COVID-19: Impact could cause equivalent of 195 million job losses, says ILO chief [Press release]. Retrieved from https://news.un.org/en/story/2020/04/1061322.

U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (1978). Uniform guidelines on employee selection procedures. Federal Register, (43)166, 38290-38315.

The United States Civil Rights Act of 1968—Title VIII Fair Housing Act (1968).

United States. Executive Office of the President. (2020). Executive Order on Modernizing and Reforming the Assessment and Hiring of Federal Job Candidates. Retrieved from https://www.whitehouse.gov/presidential-actions/executive-order-modernizing-reforming-assessment-hiring-federal-job-candidates/.

The United States Fair Credit Reporting Act (FCRA). (1970), 15 U.S.C. § 1681.

Unterrainer, J.M., Rahm, B., Kaller, C.P., Leonhart, R., Quiske, K., Hoppe-Seyler, K., Meier, C., Müller, C., and Halsband, U. (2004). Planning abilities and the Tower of London: Is this task measuring a discrete cognitive function? *Journal of Clinical and Experimental Neuropsychology*, 26(6), 846–856.

Van Iddekinge, C.H., Raymark, P.H., Eidson, Jr., C.E., and Attenweiler, W.J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance*, 17(1), 71–93.

Vanian, J. (2019). Google's Hate Speech Detection A.I. Has a Racial Bias Problem. *Forbes*. Retrieved from https://fortune.com/2019/08/16/google-jigsaw-perspective-racial-bias/.

Vazire, S., and Carlson, E.N. (2010). Self-knowledge of personality: Do people know themselves? *Social and Personality Psychology Compass*, 4(8), 605–620.

Venables, N.C., Foell, J., Yancey, J.R., Kane, M.J., Engle, R.W., and Patrick, C.J. (2018). Quantifying inhibitory control as externalizing proneness: A cross-domain model. *Clinical Psychological Science*, 6(4), 561–580.

Venkatadri, G., Lucherini, E., Sapiezynski, P., and Mislove, A. (2019). Investigating sources of PII used in Facebook's targeted advertising. In Proceedings on Privacy Enhancing Technologies, 1, 227–244.

Vrieze, E., Ceccarini, J., Pizzagalli, D.A., Bormans, G., Vandenbulcke, M., Demyttenaere, K., Van Laere, K, and Claes, S. (2013). Measuring extrastriatal dopamine release during a reward learning task. *Human Brain Mapping*, 34(3), 575–586.

Wallen, N.E. (1962). Chapter II: Development and Application of Tests of General Mental Ability. *Review of Educational Research*, 32(1), 15–24.

Wangenheim, F.V., Evanschitzky, H., and Wunderlich, M. (2007). Does the employee–customer satisfaction link hold for all employee groups? *Journal of Business Research*, 60(7), 690–697.

Wanous, J.P., Sullivan, S.E., and Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology*, 74(2), 259–264.

Washington University in St. Louis. (2019). Change the bias, change the behavior? Maybe not. *ScienceDaily*. Retrieved from www.sciencedaily.com/releases/2019/08/190802144415.htm.

Weed, S.E., Mitchell, T.R., and Moffitt, W. (1981). Leadership style, subordinate personality and task type as predictors of performance and satisfaction with supervision. In *Psychology and Industrial Productivity* (pp. 123–140). Palgrave Macmillan.

Werbel, J.D., and Gilliland, S.W. (1999). Person–environment fit in the selection process. In G.R. Ferris (Ed.), *Research in Human Resources Management*, Vol. 17 (pp. 209–243). Elsevier Science/JAI Press.

White House campaign advice to jobless: "Find something new." (2020). CBS News. Retrieved from cbsnews.com/news/unemployed-find-something-new-white-house-campaign/.

Whitmore, P.G., Fry, J.P., and Human Resources Research Organization. (1974). Soft Skills: Definition, Behavioral Model Analysis, Training Procedures. National Technical Information Service.

Wilkie, D. (2019). Employers say students aren't learning soft skills in college. *Society for Human Resource Management,* (1)2, 3.

Wilson, T.D. (2004). *Strangers to Ourselves*. Harvard University Press.

World Economic Forum. (2016). The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution. *Global Challenge Insight Report*. World Economic Forum.

World Economic Forum Boston Consulting Group (BCG). (2018). Towards a reskilling revolution: A future of jobs for all. World Economic Forum, Geneva, Switzerland.

Wu, Y., Zeng, Y., Zhang, L., Wang, S., Wang, D., Tan, X., Zhu, X., and Zhang, J. (2013). The role of visual perception in action anticipation in basketball athletes. *Neuroscience*, 237, 29–41.

Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems, 1–12.

Young, J.W. (2001). Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis. Research Report No. 2001-6. College Entrance Examination Board.

Yuste, R., Goering, S., Bi, G., Carmena, J.M., Carter, A., Fins, J.J., ... and Kellmeyer, P. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, 551(7679), 159.

Zeisel, J.S. (1963). Manpower and Training: Trends, Outlook, Programs. United States: U.S. Department of Labor, Office of Manpower, Automation and Training.

Zhang, L., Wu, Y., and Wu, X. (2017). A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 3929–3935.

Endnotes

1 Verbal-linguistic, logical-mathematical, spatial-visual, bodily-kinesthetic, musical, intrapersonal, interpersonal, naturalist, and existential.

2 Notably, an extensive debate exists about the extent to which soft skills are innate or learned, but it is beyond the scope of this brief to review this literature. For the purposes of employment selection, it is somewhat irrelevant whether a job applicant demonstrates certain soft skills due to genetic, developmental, or educational factors. Rather, the goal of incorporating soft skills into employment science is to evaluate people as they are and use the information to optimize hiring decisions. Doing so does not suggest that soft skills are completely static and cannot evolve over time.

3 Neuropsychological concepts like the Boston Process Approach, which de-emphasize final, unitary scores but instead focus on how an individual performs a task, also inform the new people science. See Nancy, H., Kaplan, E., and Milberg, W. (2009). *The Boston Process Approach to Neuropsychological Assessment: A Practitioner's Guide*. Oxford University Press.

4 The same authors provide an example: "That introversion is associated with better performance on exams could be because introverts are smarter than their more extroverted colleagues. But with a stress manipulation that reverses the rank orders of introversion and performance, we can rule out an ability explanation."

5 Such findings where tests predict outcomes better for some than for others should cause alarm: When the validity of an assessment differs across groups, it indicates mismeasurement; it indicates that a test better captures the abilities of some and misses the value of others.

6 Notably, before any aggregation occurs, a thorough job analysis is conducted to ensure that the roles are analogous across environments.

7 While tasks like the Dictator Game have traditionally been performed with real stakes—allocating real money—the effect of stakes on in-game performance remains unclear: Some researchers have found negligible effects of increased stakes, others have observed small effects, and still others have suggested that effects may depend on the populations and may vary across regions.