Intelligence 87 (2021) 101561

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

Reducing adverse impact in high-stakes testing

Alexander P. Burgoyne *, Cody A. Mashburn, Randall W. Engle

Georgia Institute of Technology, USA

ARTICLE INFO

Keywords:
Adverse impact
High-stakes tests
Equity
Attention control
ASVAB
Personnel selection

ABSTRACT

A critical goal for psychological science in the 21st century is to foster diversity, equity, and inclusion in occupational contexts. One arena which will continue to benefit from a focus on equity is *high-stakes testing*, such as the assessments used for personnel selection and classification decisions. We define an equitable test as one that minimizes group differences based on protected classes such as race, sex, and ethnicity, while predicting criterion performance equivalently across groups. In this article, we provide an overview of the concepts of test equity, adverse impact, and predictive bias. We discuss how group differences in performance on high-stakes tests such as the Armed Services Vocational Aptitude Battery (ASVAB) could be driven by differences in crystallized intelligence (i.e., acquired knowledge), which is emphasized by the ASVAB subtests and related to socioeconomic status. We suggest that shifting the focus of some high-stakes assessments away from crystallized intelligence or supplementing them with other cognitive constructs could mitigate group differences in performance without sacrificing criterion validity. In particular, we provide evidence that tests of *attention control*—the domain-general ability to maintain focus on task-relevant information and resist distraction—could provide a more equitable path forward.

1. Introduction

A critical goal for psychological science in the 21st century is to foster diversity, equity, and inclusion in occupational contexts. As the Association for Psychological Science pledged in a recent statement:

> Psychological science has the ability to transform society for the better and can and must play a central role in advancing human welfare and the public interest. To that end, we support the pursuit of a wide variety of scientific work that furthers our understanding of the causes and harmful effects of racism, stereotypes, and inequities; the psychological and societal benefits of diversity, equity, and inclusion; and the most effective ways to foster these outcomes and advance a more just and equitable world (Association for Psychological Science, 2020).

One arena which will continue to benefit from a focus on equity is *high-stakes testing*—testing situations in which the examinee's performance has significant consequences, such as in personnel selection and classification decisions. Although considerable progress has been made in this domain (for a review, see Sackett, Schmitt, Ellingson, & Kabin, 2001), there is still room for improvement.

Intelligence tests have come a long way since 1904, when Alfred Binet and Theodore Simon developed the first "modern" intelligence scale to select academically at-risk children for special education classes (Binet & Simon, 1904). Today, increasingly sophisticated tools are used to measure individual differences in general intelligence and specific cognitive abilities, such as computerized adaptive tests that change in difficulty depending on the examinee's performance (van der Linden, 2000). Nevertheless, 21st century intelligence research is not without its own issues. Here, we focus on just one of them: *test equity*. We define an equitable test as one that minimizes group differences based on protected classes such as race, sex, and ethnicity, while predicting criterion performance equivalently across groups. In this article, we provide an overview of the concepts of test equity, adverse impact, and predictive bias, followed by a discussion of how new tests of cognitive ability, and in particular, tests of attention control, could be used to reduce adverse impact in high-stakes assessments.

2. Adverse impact

High-stakes tests have a tremendous impact on professional opportunities, and in turn, economic outcomes. For example, personnel selection instruments such as the Wonderlic Personnel Test (Wonderlic, 2007) have determined access to occupational opportunities for millions

* Corresponding author.
E-mail address: burgoyn4@gmail.com (A.P. Burgoyne).

https://doi.org/10.1016/j.intell.2021.101561
Received 12 February 2021; Received in revised form 26 May 2021; Accepted 26 May 2021
Available online 3 June 2021

of job applicants since their inception (Hicks, Harrison, & Engle, 2015). In the military sector, the Armed Services Vocational Aptitude Battery (ASVAB) is similarly used to select and classify applicants, high school students, and undergraduates for military careers (ASVAB Enlistment Testing Program, 2020a). Performing well on these tests can open doors to prestigious careers and top military professions, such as intelligence officer or communications specialist. By contrast, performing poorly can limit one's employment prospects and lead to potentially more dangerous job assignments such as Army infantryman. Intelligence tests are used for selection and classification because decades of research have consistently shown they are one of, if not the single best predictor of job performance (Schmidt & Hunter, 1998), training performance (Earles & Ree, 1992), and academic success (Kuncel, Hezlett, & Ones, 2004), with meta-analytic average correlations ranging from $r = .47$ to $r = .68$ after accounting for range restriction and other psychometric limitations (Ones & Dilchert, 2004; Sackett, Borneman, & Connelly, 2008).

Unfortunately, mean scores on high-stakes tests often differ by race, sex, and ethnicity (see Neisser et al., 1996). For example, Bobko and Roth (2013) conducted a meta-analysis and found an average difference of approximately three-quarters of one standard deviation between Black and White job applicants on cognitive ability tests, and that the group difference was larger for jobs that were less complex (i.e., required less information processing; $d = 0.86$ for low complexity jobs such as factory line worker; $d = 0.72$ for moderately complex jobs such as first-level supervisor). Depending on how these high-stakes tests are used by organizations to make selection decisions, they can result in *adverse impact*—the disproportionate selection of members of one group over another.

Adverse impact occurs when a selection procedure results in inequitable hiring, promotion, or membership opportunities for members of a race, color, religion, sex, national origin group, or other protected class (Zedeck, 2011). In the United States, there are two primary approaches for determining whether a selection procedure causes adverse impact: the "four-fifths rule" (i.e., the "80% rule"; Uniform Guidelines on Employee Selection Procedures, 1978) and statistical significance tests (i.e., the *z*-test; Office of Federal Contract Compliance Programs, 1993). According to the four-fifths rule, a selection procedure has adverse impact if it results in a selection rate for any protected group that is less than 80% of the rate of the group with the highest selection rate. For example, if a test leads to the selection of 32 of 40 White applicants (an 80% selection rate) but only 10 of 25 Black applicants (a 40% selection rate), that test would have adverse impact because the selection rate for Black applicants is 50% of the selection rate for White applicants (40/80 = 50%; see Table 1).

As a supplement to the four-fifths rule, statistical significance tests can be conducted on the difference between selection rates (Morris & Lobsenz, 2000; Roth, Bobko, & Switzer III, 2006). Typically, experts will perform a *z*-test comparing the difference between independent proportions. In the example provided in Table 1, the difference in selection rates is statistically significant ($z = 3.28$, $p = .001$), indicating adverse impact.

Although the four-fifths rule and *z*-test converge on similar conclusions in this case, they frequently produce different results (Collins & Morris, 2008). Part of the issue is that the four-fifths rule evaluates the practical size of the effect resulting from the selection procedure (i.e., the impact ratio), whereas the *z*-test assesses the null hypothesis that the selection rates are the same. In other words, the four-fifths rule examines the *ratio* of selection rates, whereas the *z*-test often examines the *difference* between the selection rates (but see Morris & Lobsenz, 2000). This can lead to different conclusions simply because the tests are based on different effects. Both tests are also susceptible to sampling error; when sample sizes are small, there may be a large difference in selection rates in violation of the four-fifths rule that is not statistically significant due to low statistical power. Conversely, given a large enough sample, even a small difference in selection rates will be statistically significant.

In practice, courts will base their determination of adverse impact on not only the four-fifths rule and or the *z*-test, but other factors as well, such as whether the hiring organization's recruitment practices seem to discourage minority applicants (Collins & Morris, 2008). Regardless of how adverse impact is determined by the courts, when a test has been deemed to have adverse impact, it is incumbent on the hiring organization to demonstrate that it is valid, relevant to the job, and that alternative selection procedures have been examined (Uniform Guidelines on Employee Selection Procedures, 1978).

Table 1
An example of a selection process resulting in adverse impact according to the four-fifths rule.

Race	Applicants	Hires	Selection Rate	Adverse Impact Ratio
White	40	32	80%	–
Black	25	10	40%	50%

Note: The adverse impact ratio is calculated by dividing the selection rate of one group by the selection rate of the group with the highest selection rate. Adapted from Zedeck (2011).

3. Predictive bias

Complicating matters, however, is the issue of *predictive bias*, or differential prediction. Predictive bias occurs when the criterion validity of an item or test score differs across groups, or when, given the same score on a selection test, the predicted level of criterion performance differs by group (Cleary, 1968; Neisser et al., 1996; Schmidt & Hunter, 1998). Experts assess predictive bias by conducting a moderated regression analysis in which the outcome variable (e.g., job performance) is regressed on selection test scores, group membership, and their interaction term (Society for Industrial and Organizational Psychology, 2018); a difference in the slope or intercept across groups indicates predictive bias. Thus, whereas adverse impact can be understood in statistical terms as resulting from mean differences between groups on a selection test, predictive bias is indicated by differences in the slope of the regression line relating test performance to a criterion, or differences in its intercept across groups (see Fig. 1A and B). In other words, if the same test score leads to different predictions for different groups, the test may be considered biased. *Underprediction* is particularly problematic because it signals bias against a group (Society for Industrial and Organizational Psychology, 2018). Underprediction occurs when the overall regression line (i.e., based on the combined sample; see the black lines in Fig. 1) predicts that an individual will have lower job performance than would actually be the case (see Group 1 in Fig. 1A and B).[1] In the United States, the use of a biased selection procedure constitutes unlawful discrimination as outlined by the Equal Employment Opportunity Act of 1972 (and further clarified in the 1979 FAQ).

Thus, a test can have adverse impact without having predictive bias, provided the regression line reflecting the relationship between selection test performance and criterion performance is the same for each group (see Fig. 1C). The ASVAB provides a useful example of this distinction. As a result of mean differences in performance on the ASVAB, the qualification rate for Black applicants is less than 80% of the qualification rate for White applicants, both for entry to the military as well as eligibility for enlistment incentives (ASVAB Enlistment Testing Program, 2020b). Therefore, according to the four-fifths rule, the use of the ASVAB results in adverse impact. Nevertheless, a large-scale study by Wise et al. (1992) suggested that the ASVAB was "fair" and "sensitive" across groups (p. ii). Specifically, Wise et al. (1992) found that

[1] *Overprediction* is also indicative of predictive bias, although it does not signal bias against the overpredicted group. Overprediction occurs when a group has greater predicted performance by a common regression line than would actually be the case (see Group 2 in Fig. 1A and B).

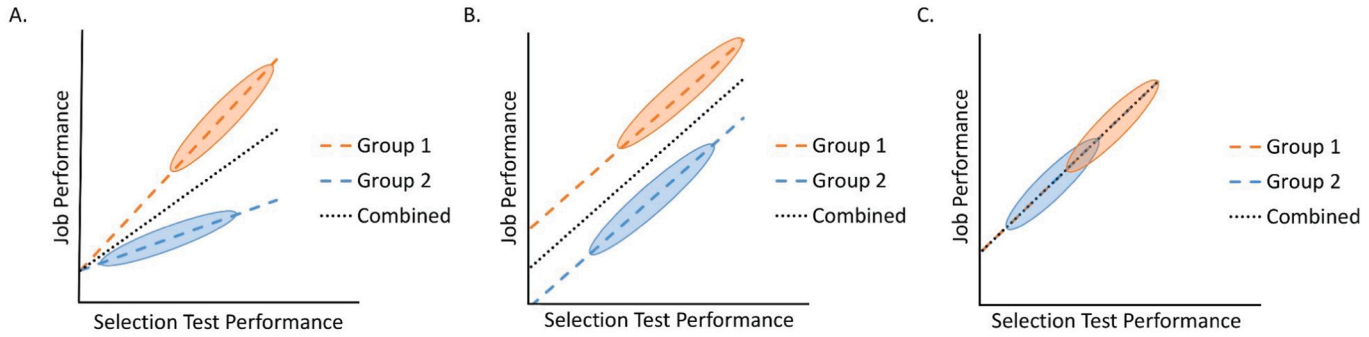

Fig. 1. Hypothetical relationships between selection test performance and job performance for two groups. Ovals represent distributions of scores for each group; lines represent regression equations relating selection test performance to job performance. In all three panels, the groups differ on selection test scores, indicating that the use of the selection test could result in adverse impact for Group 2. In panel (A), the test has predictive bias because the slopes of the regression lines differ while the intercepts are the same, and would be considered biased against Group 1 because of underprediction. In panel (B), the test has predictive bias because the intercepts for the groups differ while the slopes are the same, and, like in panel (A), the test underpredicts job performance for Group 1. In panel (C), the test is considered unbiased because the regression line is identical for each group; members of Group 1 and Group 2 with the same selection test score will have the same predicted level of job performance. Nevertheless, the use of the selection test could result in adverse impact because mean selection test performance differs by group.

individuals with the same score had the same average outcome regardless of group membership (i.e., fairness, or a lack of predictive bias), and differences in test scores were predictive of differences in outcomes (i.e., sensitivity). As Wise et al. (1992) stated: "The results indicate that the current [ASVAB] technical composites are sensitive and fair for females and blacks. Nonetheless, use of the technical composites does create a significantly greater barrier for these groups in comparison to males and whites" (p. ii). To sum up, a test can be considered fair, sensitive, and unbiased according to the definitions above while resulting in adverse impact for historically and systemically disadvantaged groups, an issue we think warrants attention and remediation.

That is, setting aside the legal ramifications of adverse impact, organizations may seek to use more equitable high-stakes tests for moral and instrumental reasons, too (Moses, 2010). From a moral standpoint, selection procedures that reduce adverse impact appeal to a desire for greater unity, integration, justice, and representation. For example, a company might value having a workforce that reflects the population at large (Sackett et al., 2001). On the other hand, from an instrumental standpoint, reducing adverse impact can have tangible benefits for organizations. For instance, diversity is positively associated with decision-making processes, possibly because it increases creativity and innovation by providing individuals with the opportunity to interact with a breadth of perspectives they might not otherwise encounter (De Dreu & West, 2001; McLeod, Lobel, & Cox Jr, 1996). Taken together, the development and adoption of more equitable high-stakes tests is motivated by legal, moral, and instrumental considerations. In the next section, we describe how our laboratory and others have attempted to use a theory-driven approach to understand and reduce group differences in high-stakes test performance.

4. Reducing adverse impact

There are a number of hypothesized causes of group differences in high-stakes test performance, ranging from systematic inequalities that lead to economic and educational differences across groups to motivational factors such as stereotype threat (Spencer, Logel, & Davies, 2016). Socioeconomic status (SES)—a composite metric comprising family income, parental education, and professional status—seems to play a role. For example, Sackett, Kuncel, Arneson, Cooper, and Waters (2009) found that SES correlated strongly with standardized test scores (r = .42), and US census data indicates that SES differs across racial and ethnic groups (McKinnon, 2002; Ryan & Siebens, 2012). The evidence suggests that higher SES is beneficial to the development of abilities measured by high-stakes tests, perhaps because families with higher SES can afford better schooling, supplemental instruction, nutrition, enrichment, and so on (Bradley & Corwyn, 2002).

However, a criticism of many high-stakes tests—notably, the Wonderlic and the ASVAB—is that they elevate the role of *crystallized intelligence*, or acquired knowledge, on overall performance. For example, the Wonderlic is comprised of verbal and numerical test questions (e.g., analogies, word comparisons, and math problems), as well as spatial reasoning items. Perhaps unsurprisingly, Matthews and Lassiter (2007) found that performance on the Wonderlic was significantly related to individual differences in crystallized intelligence but not fluid intelligence (i.e., novel problem solving ability). Similarly, Roberts et al. (2000) argued that the subtests of the ASVAB disproportionately measure acculturated learning (see Table 2), which may be especially sensitive to group differences in SES and educational history (Bosco, Allen, & Singh, 2015; Outtz & Newman, 2011; Sternberg & Wagner, 1993). Consistent with this possibility, Outtz and Newman (2011) found that the subtests of the ASVAB with the largest White-Black differences were those that measured technical knowledge (e.g., Auto and Shop Information, Mechanical Comprehension, and Electronics Information) and verbal/crystallized intelligence (e.g., General Science, Word Knowledge, and Paragraph Comprehension). Further evidence is provided by Hough, Oswald, and Ployhart (2001), who found that tests of crystallized intelligence (e.g., verbal ability, quantitative ability, and science

Table 2
Subtests of the Armed Services Vocational Aptitude Battery (ASVAB).

Subtest	Description	Content Area
General Science	Knowledge of physical and biological sciences	Science/ Technical
Arithmetic Reasoning	Ability to solve arithmetic word problems	Math
Word Knowledge	Ability to select the correct meaning of words presented in context and to identify the best synonym for a given word	Verbal
Paragraph Comprehension	Ability to obtain information from written passages	Verbal
Math Knowledge	Knowledge of high school mathematics principles	Math
Electronics Information	Knowledge of electricity and electronics	Science/ Technical
Auto Information	Knowledge of automobile technology	Science/ Technical
Shop Information	Knowledge of tools and shop terminology and practices	Science/ Technical
Mechanical Comprehension	Knowledge of mechanical and physical principles	Science/ Technical
Assembling Objects	Ability to determine how an object will look when its parts are put together	Spatial

Note. This table was adapted from ASVAB Enlistment Testing Program (2020c).

achievement) tended to have larger group differences than tests of memory, spatial ability, and processing speed (see also Verive & McDaniel, 1996). Thus, to the extent that group differences in performance on high-stakes tests such as the ASVAB are a result of SES-associated differences in acquired knowledge, supplementing high-stakes assessments with non-crystallized intelligence measures could mitigate group differences in performance (Bosco et al., 2015).

Of course, there is a good reason why the military and other employers sometimes include tests of domain-specific knowledge in personnel selection assessments: job relevance. High-stakes tests may assess examinees' domain-specific knowledge because it is required for competency in the domain. As just one example, extensive knowledge about auto and shop information may be particularly valuable for mechanics in the military. Although job-specific knowledge and skills can be learned, there are situations in which employers would prefer to select individuals who already have a modicum of proficiency in an area. In these cases, removing tests of acquired knowledge would not be warranted, but supplementing them with other job-relevant tests that demonstrate less adverse impact could be beneficial.

What, then, could supplement the ASVAB's focus on crystallized intelligence? Tests of other cognitive abilities that are less dependent on acquired knowledge, such as attention control, working memory capacity, and fluid intelligence could help reduce adverse impact (Hough et al., 2001; Outtz & Newman, 2011). *Attention control* refers to the domain-general ability to maintain focus on goal-relevant information and disengage from no-longer-relevant information while resisting interference and distraction by irrelevant thoughts and events (Burgoyne & Engle, 2020; Engle & Kane, 2004). It is closely related to *working memory capacity*, the amount of information an individual can maintain in a readily accessible state. *Fluid intelligence*, on the other hand, refers to an individual's capacity for novel learning, reasoning, and problem solving.

Individual differences in these cognitive abilities predict academic and work-relevant outcomes (Best, Miller, & Naglieri, 2011; Martin, Mashburn, & Engle, 2020; Rohde & Thompson, 2007), and in fact, some tests of these abilities are already used in the occupational sector for personnel selection. AON Assessment Solutions, for example, advertises a gamified version of a working memory capacity task on their website, noting that "[b]y focusing on attention control instead of learned knowledge, G.A.M.E. [Global Adaptive-Memory Evaluation] diminishes the high levels of adverse impact associated with traditional cognitive ability tests" (Martin & LaPort, 2017, par. 5).

While tests of attention control, working memory, and fluid intelligence may all be less sensitive to cultural differences than tests of crystallized intelligence, our position is that attention control is a central cognitive construct underpinning individual differences in these abilities (Burgoyne & Engle, 2020). Attention control is typically measured by tasks that require inhibiting a prepotent response, such as the antisaccade task (Hallett, 1978), in which participants must look *away* from a flashing asterisk on one side of the screen to detect a briefly presented letter on the opposite side of the screen.[2] Another example is the Stroop task (Stroop, 1935), in which participants are shown color words in different colors (e.g., "RED" printed in green), and must indicate the color of the word, *not* the color the word refers to. In both cases, test takers must maintain focus on task goals amidst interference and distraction to perform well.

Our theoretical framework holds that the domain-general ability to control attention underpins performance on a variety of cognitive tasks, from learning and reasoning to memory and multitasking, helping to explain why measures of cognitive ability tend to correlate positively with one another (Engle, 2018; for a review, see (Burgoyne, Tsukahara, Draheim, & Engle, 2020). For example, attention control appears to be the primary 'active ingredient' tapped by working memory capacity tasks (Engle, Tuholski, Laughlin, & Conway, 1999), which challenge participants to juggle the cognitive demands of information storage and active processing (e.g., remembering letters while solving math equations, as in the Operation Span task). Large-scale latent variable analyses have shown that attention control drives the relationship between working memory capacity and higher-order constructs such as fluid intelligence (Draheim, Tsukahara, Martin, Mashburn, & Engle, 2020) and aspects of rational thinking (Burgoyne, Mashburn, Tsukahara, Hambrick, & Engle, 2021). Attention control also supports more 'primitive' functions such as sensory discrimination—the ability to differentiate between visual, auditory, or other sensory stimuli (Tsukahara, Harrison, Draheim, Martin, & Engle, 2020).

Recent evidence suggests that tests of attention control and working memory capacity could reduce adverse impact relative to traditional high-stakes tests without loss in criterion validity. For example, across a series of studies of 273 bank employees and 197 undergraduates, Bosco et al. (2015) compared the validity of tests of attention control (the Flanker task) and working memory (Operation Span and Reading Span) to a conventional test of mental ability (the Wonderlic Personnel Test) for predicting supervisor ratings and performance in a management simulation. In each study, the criterion validity of the attention control-related measures rivaled that of the Wonderlic Personnel Test while resulting in smaller group differences between Black and White participants. Indeed, a meta-analysis of the studies revealed that the combined attention control and working memory measures reduced group differences by approximately half of one standard deviation compared to the Wonderlic Personnel Test (compare $d = 0.68$ to $d = 1.09$). Further, it did so without sacrificing criterion validity; for management simulation performance, the meta-analytic correlations were $r = .35$ for the combined attention control and working memory capacity measures and $r = .33$ for the Wonderlic Personnel Test. While these results are promising, we note that Bosco et al. (2015) used one measure of attention control and two indirect measures (i.e., working memory tasks) which are potentially susceptible to contamination by domain-specific acquired knowledge (math skill in the Operation Span task and reading skill in the Reading Span task; see Hambrick, Kane, & Engle, 2005). Thus, it is possible that a battery of attention control tasks—which require little acculturated knowledge—could fare better.

Results from our own lab indicate that as a construct, attention control can predict multitasking ability above and beyond the ASVAB while possibly reducing adverse impact (Martin et al., 2020). Multitasking is a ubiquitous cognitive demand of modern military and civilian jobs, particularly given their increasing reliance on information technology (Burgoyne, Hambrick, & Altmann, 2020; Hambrick et al., 2011; Martin et al., 2020). As a case in point, the O*Net occupational database currently lists more than 800 professions that require "the ability to shift back and forth between two or more activities or sources of information" (National Center for O*NET Development, 2020). In a sample of 171 young adults aged 18–35, Martin et al. (2020) found that a latent factor representing performance on new-and-improved attention control tasks uniquely accounted for 22.1% of the variance in multitasking ability after accounting for fluid intelligence and performance on a 180-item ASVAB practice test (Fig. 2). Thus, attention control added substantially to the prediction of multitasking performance above and beyond ASVAB scores.

Not only did attention control predict multitasking above and beyond the ASVAB and fluid intelligence in latent variable analyses—its raw correlation with multitasking performance rivaled that of the ASVAB. Specifically, Martin et al. (2020) found that a composite variable representing performance on the attention control tasks correlated significantly with multitasking ability, $r(169) = .61$, 95% CI [.50, .70], $p < .001$. By comparison, scores on the ASVAB practice test correlated $r(169) = .67$, 95% CI [.60, .74], $p < .001$ with multitasking ability. One can also compare the validity of attention control to scores on the Armed Forces Qualification Test (AFQT), a composite based on the ASVAB's

[2] An interactive version of the antisaccade task can be found at https://englelab.gatech.edu/taskdemonstrations

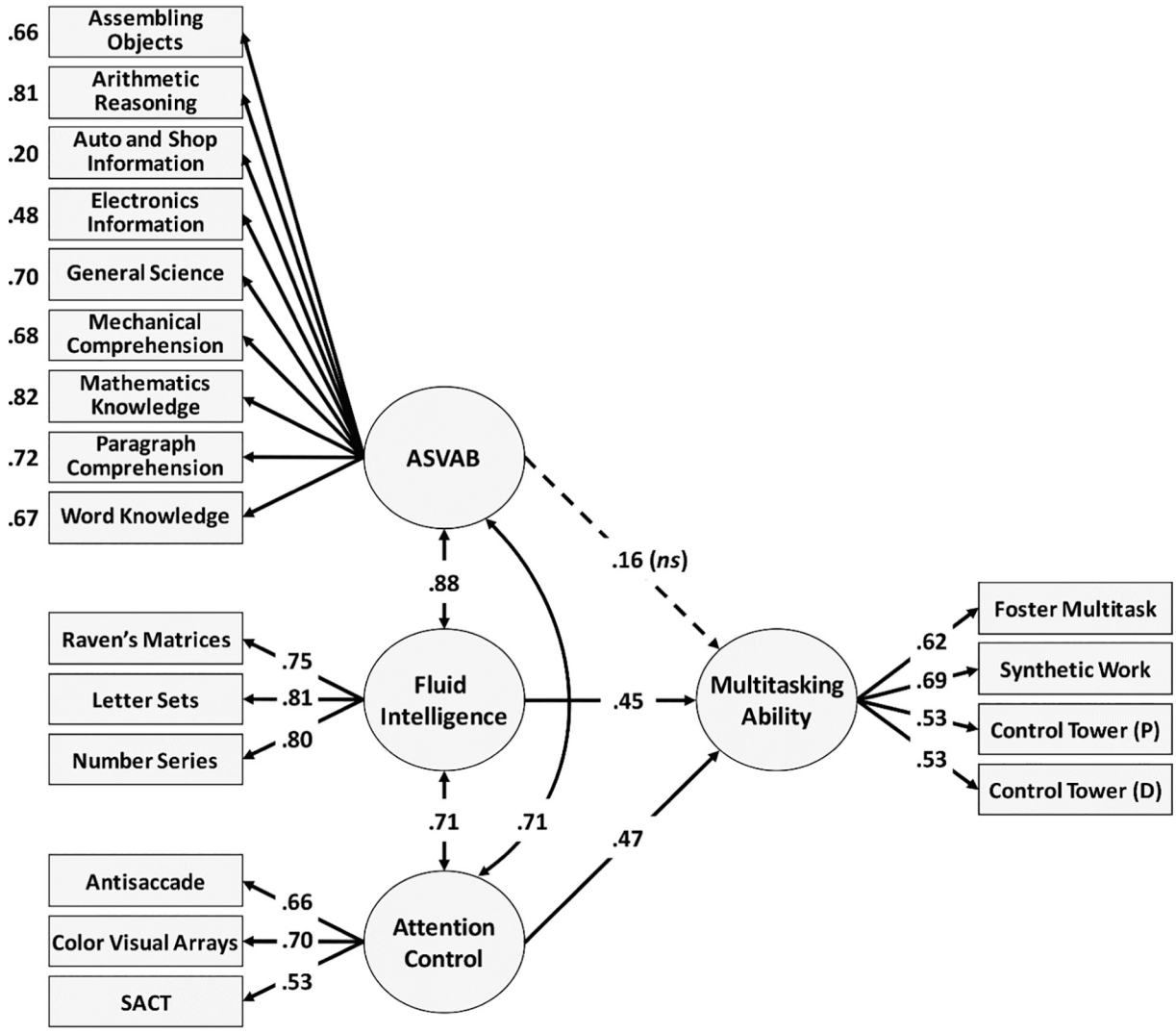

Fig. 2. In this structural equation model ($N = 171$; Martin et al., 2020), latent factors are depicted as ovals and observed measures are depicted as rectangles. Values to the side of the rectangles are factor loadings. Attention control correlated strongly with fluid intelligence (.71) and ASVAB performance (.71), as indicated by the values along the double-headed arrows. The contribution of attention control to multitasking ability (.47) was substantial and significant after accounting for the ASVAB and fluid intelligence, accounting for 22.1% of the variance in multitasking ability above and beyond the other predictors. By contrast, the contribution of ASVAB performance to the prediction of multitasking ability was non-significant (.16, *ns*) after accounting for attention control and fluid intelligence.

verbal and arithmetic subtests. In the Martin et al. (2020) data, AFQT scores correlated $r(169) = .64$, 95% CI [.55, .71], $p < .001$ with multitasking ability. None of these dependent correlations differ significantly differ from one another (all $ps > .15$).

Although Martin et al. (2020) did not report group differences in performance on the attention control measures relative to the ASVAB or AFQT, we conducted secondary analyses using their data and present them here. First, we note the following caveats: the sample size for these comparisons is far too small to draw strong conclusions, comprising only 68 Black participants and 38 White participants. Second, the sample is not representative; although efforts were made to recruit a broad sample, most of the Black participants were recruited from the greater Atlanta area and most of the White participants were college students, many of whom attended the Georgia Institute of Technology. Thus, selection effects confounded race and cognitive ability, leading to much larger standardized group differences in this sample than would be observed in the general population, and indeed than is observed in the literature on group differences. Nevertheless, the data are still potentially informative for examining the *relative difference* in performance between groups on the ASVAB and attention control measures, as the same participants completed both sets of tasks.

With those caveats in mind, in the Martin et al. (2020) dataset the attention control composite reduced the group difference between White and Black participants by three-quarters of one standard deviation (i.e., a reduction of $d = 0.75$) compared to the ASVAB, and two-thirds of one standard deviation (i.e., a reduction of $d = 0.63$) compared to the AFQT. Specifically, for the attention control measures, the group difference was $d = 1.11$, 95% CI [0.75, 1.50], whereas for the ASVAB, $d = 1.86$, 95% CI [1.47, 2.33], and for the AFQT, $d = 1.74$, 95% CI [1.32, 2.23]. In other words, there were smaller group differences associated with the attention control measures than the ASVAB and AFQT, suggesting that attention control tasks could reduce adverse impact.

That said, the overall group differences in Martin et al. (2020) are far larger than those observed in meta-analyses of cognitive ability, such that all of the measures would likely result in adverse impact if used for personnel selection (recall that Bobko & Roth, 2013 found an average *d* ranging from 0.72 to 0.86). To reiterate, the participant recruitment process used by Martin et al. (2020) is the most likely explanation for these large *d* values, and is only one reason among others (e.g., the small sample size) to regard this finding as speculative. As this preliminary result is by no means conclusive, we are currently investigating whether

attention control tests can improve the prediction of job performance while reducing adverse impact using larger and more representative samples in both applied and laboratory settings. Following an extensive data collection effort, we plan to meta-analyze group differences on the attention control measures alongside other high-stakes tests to determine whether attention control tests can reduce adverse impact while yielding meta-analytic *d*s that are comparable to those typically observed in the literature (see, e.g., Bosco et al., 2015).

5. Conclusion

A crucial goal for the future of intelligence research is to develop more equitable high-stakes tests for the occupational sector. While tests of attention control may have the potential to reduce adverse impact by shifting the focus of high-stakes tests away from acculturated knowledge, they are certainly no panacea. Specifically, some high-stakes tests, such as those used for professional credentialing (e.g., the Uniform Bar Exam for attorneys), assess examinees' domain-specific knowledge because it facilitates performance in the domain. In these cases, removing tests of domain-specific knowledge would not be warranted, especially when no job-relevant alternative measures demonstrate less adverse impact (Society for Industrial and Organizational Psychology, 2018). However, where ability or aptitude are of concern, shifting the focus or augmenting selection tests with constructs such as working memory capacity, fluid intelligence, and especially attention control may provide a route to more equitable testing. Indeed, acquired knowledge and cognitive abilities such as attention control likely capture unique (i.e., potentially incremental) variance in job performance; one is not necessarily a replacement for the other. Finally, though a promising avenue, group differences on high-stakes tests may not be totally ameliorated by focusing exclusively on tests themselves; there is still much work to be done addressing the systemic and historical inequities that continue to affect society. Nevertheless, psychological scientists should strive to use their expertise to reduce societal inequities wherever possible.

Funding

This work was supported by Office of Naval Research Grants N00173-20-2-C003 and N00173-20-P-0135 to Randall W. Engle.

Declaration of Competing Interest

None.

References

Association for Psychological Science. (2020). *APS statement on diversity, equity, & inclusion*. December 7. Association for Psychological Science https://www.psychologicalscience.org/diversity-equity-inclusion.

ASVAB Enlistment Testing Program. (2020a). *What is the ASVAB?*. August 13. Retrieved January 20, 2021, from https://www.officialasvab.com/.

ASVAB Enlistment Testing Program. (2020b). *Fairness information*. August 13. Retrieved January 22, 2021, from https://www.officialasvab.com/researchers/fairness-information/.

ASVAB Enlistment Testing Program. (2020c). *ASVAB subtests*. February 10. Retrieved February 03, 2021, from https://www.officialasvab.com/counselors-educators/subtests/.

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences, 21*, 327–336.

Binet, A., & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11*, 191–244.

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on black–white mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology, 66*, 91–126.

Bosco, F., Allen, D. G., & Singh, K. (2015). Executive attention: An alternative perspective on general mental ability, performance, and subgroup differences. *Personnel Psychology, 68*, 859–898.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*, 371–399.

Burgoyne, A. P., & Engle, R. W. (2020). Attention control: A cornerstone of higher-order cognition. *Current Directions in Psychological Science, 29*, 624–630.

Burgoyne, A. P., Hambrick, D. Z., & Altmann, E. M. (2020). Incremental validity of placekeeping as a predictor of multitasking. *Psychological Research*, 1–14. https://doi.org/10.1007/s00426-020-01348-7.

Burgoyne, A. P., Mashburn, C., Tsukahara, J. S., Hambrick, Z., & Engle, R. W. (2021). *Understanding the relationship between rationality and intelligence: A latent-variable approach*. February 8. Retrieved from psyarxiv.com/ns9ky.

Burgoyne, A. P., Tsukahara, J. S., Draheim, C., & Engle, R. W. (2020). Differential and experimental approaches to studying intelligence in humans and non-human animals. *Learning and Motivation, 72*, 101689.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124.

Collins, M. W., & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology, 93*, 463–471.

De Dreu, C. K., & West, M. A. (2001). Minority dissent and team innovation: The importance of participation in decision making. *Journal of Applied Psychology, 86*, 1191–1201.

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2020). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General, 150*, 242–275.

Earles, J. A., & Ree, M. J. (1992). The predictive validity of the ASVAB for training grades. *Educational and Psychological Measurement, 52*, 721–725.

Engle, R. W. (2018). Working memory and executive attention: A revisit. *Perspectives on Psychological Science, 13*, 190–193.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In , *vol. 44. The psychology of learning and motivation: Advances in research and theory* (pp. 145–199).

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*, 309–331.

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research, 18*, 1279–1296.

Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2005). The role of working memory in higher-level cognition: Domain-specific versus domain-general perspectives. In R. J. Sternberg, & J. E. Pretz (Eds.), *Cognition and intelligence: Identifying the mechanisms of the mind* (pp. 104–121). Cambridge University Press: Cambridge, UK.

Hambrick, D. Z., Rench, T. A., Poposki, E. M., Darowski, E. S., Roland, D., Bearden, R. M., … Brou, R. (2011). The relationship between the ASVAB and multitasking in navy sailors: A process-specific approach. *Military Psychology, 23*, 365–380.

Hicks, K. L., Harrison, T. L., & Engle, R. W. (2015). Wonderlic, working memory capacity, and fluid intelligence. *Intelligence, 50*, 186–195.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148–161.

van der Linden, W. J. (2000). In C. A. Glas (Ed.), *Computerized adaptive testing: Theory and practice*. Springer Science & Business Media.

Martin, J., Mashburn, C. A., & Engle, R. W. (2020). Improving the validity of the armed service vocational aptitude battery with measures of attention control. *Journal of Applied Research in Memory and Cognition, 9*, 323–335.

Martin, N., & LaPort, K. (2017). *Traditional cognitive Testing got you down? Get your G.A.M.E. on with Aon!*. April 13. Retrieved February 03, 2021, from https://www.aon.com/human-capital-consulting/talent/talent-rewards-performance-leadership-assessment-b/gameon-blog.jsp.

Matthews, T. D., & Lassiter, K. S. (2007). What does the wonderlic personnel test measure? *Psychological Reports, 100*, 707–712.

McKinnon, J. (2002). The black population in the United States: March 2002. In *Current population reports* (pp. P20–541). US Census Bureau.

McLeod, P. L., Lobel, S. A., & Cox, T. H., Jr. (1996). Ethnic diversity and creativity in small groups. *Small Group Research, 27*, 248–264.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*, 89–111.

Moses, M. S. (2010). Moral and instrumental rationales for affirmative action in five national contexts. *Educational Researcher, 39*, 211–228.

National Center for O*NET Development. (2020). *Browse by O*NET data. O*NET online*. Retrieved January 26, 2021, from https://www.onetonline.org/find/descriptor/result/1.A.1.g.2.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., … Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Office of Federal Contract Compliance Programs. (1993). *Federal contract compliance manual*. Washington, DC Department of Labor: Employment Standards Administration, Office of Federal Contract Compliance Programs (SUDOC No. L 36.8: C 76/993).

Ones, D. S., & Dilchert, S. (2004). *Practical vs. general intelligence in predicting success in work and educational settings*. Paper presented at the University of Amsterdam. October.

Outtz, J. L., & Newman, D. A. (2011). A theory of adverse impact. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 53–94). Routledge: New York.

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G., & Stankov, L. (2000). The armed services vocational aptitude battery (ASVAB): Little more than acculturated learning (Gc)!? *Learning and Individual Differences, 12*, 81–103.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83–92.

Roth, P. L., Bobko, P., & Switzer, F. S., III (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology, 91*, 507–522.

Ryan, C. L., & Siebens, J. (2012). Educational attainment in the United States: 2009. Population characteristics. In *Current population reports* (pp. P20–566). US Census Bureau.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*, 215–227.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin, 135*, 1–22.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56*, 302–318.

Schmidt, F. L., & 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*, 262–274.

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures* (Fifth Edition).

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67*, 415–437.

Sternberg, R. J., & Wagner, R. K. (1993). The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2*, 1–4.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Tsukahara, J. S., Harrison, T. L., Draheim, C., Martin, J. D., & Engle, R. W. (2020). Attention control: The missing link between sensory discrimination and intelligence. *Attention, Perception, & Psychophysics, 82*, 3445–3478.

Uniform Guidelines On Employee Selection Procedures. (1978). *29 CFR § 1607*.

Verive, J. M., & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence, 23*, 15–32.

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D. R. (1992). *Sensitivity and fairness of the armed services vocational aptitude battery (ASVAB) technical composites*. Defense Manpower Data Center, Department of Defense: Personnel Testing Division. https://www.officialasvab.com/wp-content/uploads/2019/08/AS92009_Sensitivity_Fairness_of_ASVAB_Tech_Composites.pdf.

Wonderlic, E. F. (2007). *Wonderlic personnel test-revised: Manual*. Los Angeles, CA: Western Psychological Services.

Zedeck, S. (2011). Adverse impact: History and evolution. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 3–28). Routledge: New York.