CAMBRIDGE
UNIVERSITY PRESS

**COMMENTARY**

# Perfect is the enemy of good enough: Putting the side effects of intelligence testing in perspective

In-Sue Oh*

Department of Human Resource Management, Fox School of Business, Temple University
*Corresponding author. Email: insue.oh@temple.edu

In their focal article, Watts et al. (2021) discuss the important yet often ignored side effects associated with many widely known organizational interventions including intelligence testing and urge industrial and organizational (I-O) psychologists to pay more attention to them. Although I agree that I-O psychologists should not merely focus on the benefits of intelligence testing (e.g., validity) but also its side effects (e.g., adverse impact), I would like to take this opportunity to put in perspective several of the side effects mentioned in the focal article, given their potential to be misleading and perpetuate misconceptions about intelligence testing.[1]

First, Watts et al. (2021) state that intelligence testing "has in some instances been associated with noteworthy side effects, such as ... illegal discrimination against minority groups" (p. X) and that "on average, organizations that used these [traditional intelligence] tests to select applicants increased their odds of discriminating against minority group members" (p. X). In this excerpt, Watts et al. use two strong words—"illegal" and "discrimination"—without sufficient qualifiers. The use of these words has the potential to elicit the wrong impression, particularly among laypeople and nonexperts, that (a) the use of a traditional (written) GMA test as a selection procedure is (likely) ruled to be illegal in court and (b) such tests are less predictive of job performance among minority groups than among the majority group. Speaking from a legal standpoint, such demonstrations in court often rely on cumulative research findings rather than on local validation studies plagued with sampling error and other methodological artifacts (Schmidt, 2009). Therefore, GMA tests are (and should be) quite defensible in court given the extensive meta-analytic evidence that such tests are highly predictive of job performance, and, relevantly, their validity does not differ across major ethnic groups (e.g., Roth et al., 2014; Schmidt, 1988). Also, GMA tests are not predictively biased against minority groups, meaning that both minority and majority applicants with the same GMA test scores have practically the same level of later job performance (Berry & Zhao, 2015).[2] As Schmidt (2009) noted, employers have been winning more and more such suits since the mid-1980s, and there are simply fewer such suits in recent years; "currently, less than 1% of employment-related lawsuits are challenges

[1]Given that the term intelligence is not widely used in personnel selection, I will use another term, general mental ability (GMA), interchangeably with intelligence.

[2] Cumulative research has shown that GMA test scores, in general, slightly overpredict minority (in particular, Black) applicants' job performance. However, a recent study by Berry et al. (2020) has reported that "cognitive ability tests can be expected to underestimate Hispanic American job applicants' job performance by a small to moderate amount much of the time" (p. 537).

to selection tests or other hiring procedures" (p. 12). In summary, there is ample research evidence that GMA tests, if properly used following professional guidelines such as the Society for Industrial and Organizational Psychology's (SIOP; 2018) *Principles*, are predictively fair or unbiased and, as such, a legally defensible selection procedure in most, if not all, cases (Schmidt, 2009).

Second, what is referred to as the side effects of GMA tests in the focal article appears to be the well-known concept of disparate or adverse impact. However, adverse impact is not necessarily a real (vs. potential) legal risk or an illegal side effect. Cumulative research shows that members of some minority groups have lower average scores on GMA tests than members of the majority group (Roth et al., 2017). If applicants are selected based solely on their GMA test scores in a top-down manner, this can lead to lower hiring rates on average for lower scoring minority groups. This was well known even in the early 1980s. Government agencies such as the Equal Employment Opportunity Commission started to refer to these lower hiring rates as "adverse impact," which is rarely a real legal risk in the case of GMA tests as discussed above. (As Schmidt [2009] noted, "the term adverse impact is deceptive, because it implies that the GMA tests *create* the difference in test scores, when in fact the tests only *measure* real preexisting differences in mental skills" [p. 12].) Additionally, not using written GMA tests does not necessarily remove the possibility of adverse impact as long as cognitively loaded constructs (e.g., GMA and job-relevant knowledge, skills, and specific abilities) are measured using other selection procedures (e.g., job interviews; Oh, 2013). Then, what is a good enough, though not perfect, and lawful solution to adverse impact or the so-called "validity–diversity dilemma"? The answer, fortunately, lies well within the realm of currently accepted best personnel selection practices. Specifically, GMA tests can be supplemented with valid noncognitive predictors that show little, if any, adverse impact (e.g., conscientiousness). This selection battery is not only predictive of later job performance but also helps increase diversity by reducing the adverse impact associated with GMA tests (Ployhart & Holtz, 2008, p. 168). In recent years, some employers have also attempted balancing the biobjectives of performance (validity) and diversity using Pareto-optimal weighting methods borrowed from the multiple-objective optimization literature in engineering and economics (De Corte et al., 2021; Rupp et al., 2020).[3] In summary, the adverse impact associated with intelligence testing is not an intractable risk, as most employers (can) measure valid noncognitive traits (e.g., conscientiousness) in addition to GMA in hiring without incurring too much additional cost.

Third, as discussed in their focal article and above, "a number of approaches have been proposed with the goal of mitigating the side effects of traditional intelligence tests (e.g., score banding, alternative measurement methods) . . . some of these approaches have demonstrated partial success (Ployhart & Holtz, 2008), but none have been found to eliminate all concerns of bias" (p. X). However, this statement is only half true, because they do not mention perhaps the most potent solution to this issue, within-group norming, which can equalize minority and nonminority hiring rates and thus eliminate adverse impact. However, this solution is not without legal barriers. As discussed in Sackett and Wilk (1994, p. 929), although within-group norming was recommended as *the* solution to the adverse impact associated with the use of intelligence testing as a selection procedure by a National Academy of Sciences (NAS) committee (Hartigan & Wigdor, 1989), the 1991 Civil Rights Act made such score adjustments "an unlawful practice for an employer," leaving hiring managers with more questions than answers.

Fourth, Watts et al. (2021) state that "although this approach [traditional assessments of *g*] to intelligence testing has been criticized on a number of fronts (e.g., Schneider & Newman, 2015), we focus on the use of *g* because such tests continue to be used frequently in employee selection systems" (p. X, bracket added). There are two problems with this statement. First, Watts et al. ignore the cumulative research findings that show the use of specific abilities instead of GMA

---

[3] According to Rupp et al. (2020), "Pareto–optimal weighting is similar to regression weighting in that it also seeks 'optimized' composite scores. However, it differs from regression weighting in that it aims to optimize two (or more) outcomes simultaneously" (p. 249).

is not generally recommended, particularly in predicting job and training performance (both broad). Specifically, the validity of GMA tests for job and training performance is generally higher than that of specific aptitudes—even when specific aptitudes are chosen to match the most important aspects of job performance (i.e., spatial perception for the pilot job). Second, another important aspect of GMA is that it is the broadest of all cognitive abilities. Narrower abilities, such as verbal, quantitative, and spatial abilities, and specific aptitude measures also predict job performance, but this is largely attributed to their partially measuring GMA. In other words, when a test of verbal ability predicts job or training performance, it is the GMA part of that test ($g$)—not much of the verbal part—that primarily does the predicting, thus "not much more than $g$" (Ree & Earles, 1991; Ree et al., 1994). Moreover, it is hard to believe that Schneider and Newman's (2015) article is cited as an example to illustrate that "intelligence testing has been criticized on a number of fronts." My understanding of this review paper is that researchers and practitioners should make better use of specific aptitudes, as they can be as predictive as GMA in predicting "specific" (vs. broad) performance dimensions. For example, Schneider and Newman (2015) stated the following as one of the most important implications from their review: "Past findings that show specific abilities do not offer strong unique prediction of general criteria (see Ree & Carretta, 2002) are not surprising (i.e., they are consistent with the compatibility principle), but such results constitute insufficient grounding from which to dismiss the validity of specific abilities. Specific abilities should predict specific job performance criteria, not overall job performance" (p. 25).

Fifth, Watts et al. (2011) state, "for some of these examples (e.g., traditional intelligence testing), I-O psychologists have played a critical role in shedding much light on side effects" (p. X). This is largely true, but not entirely. As discussed above, although the NAS recommended within-group norming as the most scientific solution to the validity–diversity dilemma, the 1991 Civil Rights Act determined it as an unlawful practice for nonscientific reasons (Sackett & Wilks, 1994). However, there was no major protest to this decision from I-O psychologists. In a poignant (in my opinion) *TIP* article, Schmidt (2006) wrote,

> Other sciences and professions—medicine, biology, engineering—have done a much better job on this. When lawyers, courts, other organizations, or the media appear to endorse false ideas, these groups launch vigorous public educational campaigns. . . . I-O psychologists have produced no such response. (p. 26)

Thus, a real side effect of intelligence testing not mentioned in Watts et al. is political risk.

Sixth, based on a survey of 5,000 Society for Human Resource Management members whose title was at the manager level and above, Rynes et al. (2002) found that 72% of the 959 respondents did not know or believe that GMA tests are more predictive of job and training performance than most noncognitive selection procedures such as conscientiousness measures (see Schmidt & Hunter, 1998). This problem is not limited to the United States but widely observed in other countries (e.g., Tenhiälä et al., 2016). Evidently, another real serious side effect of intelligence testing seems to be ignorance or disbelief of the enormous research evidence regarding the validity and utility of GMA tests among human resources (HR) managers.

Last, it is also important to recognize the side effects associated with *not* using, rather than using, GMA tests as a major selection procedure. A well-controlled, natural quasi-experiment (e.g., U.S. Steel Plant at Fairless Hill, PA) introduced in Schmidt (2009) highlights the substantial economic loss associated with hampering the use of GMA tests as a selection procedure (e.g., increased training time and costs; decreased productivity). This point is not limited to intelligence testing but applicable to other organizational interventions.

In conclusion, although I agree with the general theme of the focal article, I think they fail to put in perspective the sheer complexity around the side effects of intelligence testing. In short: Perfect

is the enemy of good enough! This seems to be the perfect phrase for a GMA test because its benefits, such as high validity and utility, outweigh its side effects, whether real or not. However, pseudo, fake, and voodoo science (to be clear, not intended for the focal article) seem to focus unevenly more on its (potential) side effects and argue that unless perfect, a GMA test should probably not be used as a selection procedure, thus doing more harm than good. All of us, as I-O psychologists, need be beware of this doctrine, as it only pushes us further away from rather than closer to the truth.

# References

Berry, C. M., & Zhao, P. (2015). Addressing criticisms of existing predictive bias research: cognitive ability test scores still overpredict African Americans' job performance. *Journal of Applied Psychology*, **100**(1), 162–179. https://doi.org/10.1037/a0037615

Berry, C. M., Zhao, P., Batarse, J. C., & Reddock, C. (2020). Revisiting predictive bias of cognitive ability tests against Hispanic American job applicants. *Personnel Psychology*, **73**(3), 517–542. https://doi.org/10.1111/peps.12378

De Corte, W., Lievens, F., & Sackett, P. R. (2021). A comprehensive examination of the cross-validity of pareto-optimal versus fixed-weight selection systems in the biobjective selection context. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0000927

Hartigan, J. A., & Wigdor, A. K. (Eds.) (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery*. National Academy of Sciences Press.

Oh, I.-S. (2013). Adverse impact is unlikely to be eliminated as long as cognitively loaded constructs are assessed. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **6**(4), 506–508. https://doi.org/10.1111/iops.12092

Ployhart, R. E., & 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. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Ree, M. J., & Carretta, T. R. (2002). g2K. *Human Performance*, **15**, 3–23. https://doi.org/10.1080/08959285.2002.9668081

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology*, **44**(2), 321–332. https://doi.org/10.1111/j.1744-6570.1991.tb00961.x

Ree, M. J., Earles, J. A., & Teachout, M. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology*, **79**(4), 518–524. https://doi.org/10.1037/0021-9010.79.4.518

Roth, P. L., Le, H., Oh, I.-S., Van Iddekinge, C. H., Buster, M. A., Robbins, S. B., & Campion, M. A. (2014). Differential validity for cognitive ability tests in employment and educational settings: Not much more than range restriction? *Journal of Applied Psychology*, **99**(1), 1–20. https://doi.org/10.1037/a0034377

Roth, P. L., Van Iddekinge, C. H., DeOrtentiis, P. S., Hackney, K. J., Zhang, L., & Buster, M. A. (2017). Hispanic and Asian performance on selection procedures: A narrative and meta-analytic review of 12 common predictors. *Journal of Applied Psychology*, **102**(8), 1178–1202. https://doi.org/10.1037/apl0000195

Rupp, D. E., Song, Q. C., & Strah, N. (2020). Addressing the so-called validity–diversity trade-off: Exploring the practicalities and legal defensibility of Pareto-optimization for reducing adverse impact within personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **13**(2), 246–271. https://doi.org/10.1017/iop.2020.19

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practices. *Human Resource Management*, **41**(2), 149–174. https://doi.org/10.1002/hrm.10029

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, **49**(11), 929–954. https://doi.org/10.1037/0003-066X.49.11.929

Schmidt, F. L. (1988). The problem of group differences in ability scores in employment selection. *Journal of Vocational Behavior*, **33**(3), 272–292. https://doi.org/10.1016/0001-8791(88)90040-1

Schmidt, F. L. (2006). The orphan area for meta-analysis: Personnel selection. *The Industrial/Organizational Psychologist*, **44**(2), 25–28.

Schmidt, F. L. (2009). Select on intelligence. In Locke, E.A. (Ed.), *Principles of organizational behavior* (2nd ed., pp. 3–17). Wiley.

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**(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schneider, W. J., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review*, **25**, 12–27. https://doi.org/10.1016/j.hrmr.2014.09.004

Society for Industrial and Organizational Psychology (SIOP). (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). Society for Industrial and Organizational Psychology.

**Tenhiälä, A., Giluk, T. L., Kepes, S., Simon, C., Oh, I.-S., & Kim, S.** (2016). The research-practice gap in human resource management: A cross-cultural study. *Human Resource Management*, **55**(2), 179–200. https://doi.org/10.1002/hrm.21656

**Watts, L. L., Gray, B. E., & Medeiros, K. E.** (2021). Side effects associated with organizational interventions: A perspective. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **15**(1), 76–94.