



COMMENTARY

Rumors of general mental ability's demise are the next red herring

Jeffrey M. Cucina¹  and Theodore L. Hayes² 

¹Personnel Research and Assessment Division, U.S. Customs and Border Protection, Washington, DC, USA and ²Arlington, VA, USA

Corresponding author: Jeffrey M. Cucina; Email: jcucina@gmail.com

In this paper we focus on the lowered validity for general mental ability (GMA) tests by presenting: (a) a history of the range restriction correction controversy; (b) a review of validity evidence using various criteria; and (c) multiple paradoxes that arise with a lower GMA validity.

How did we get here? History of the range restriction correction controversy

Sackett et al. (2022, 2023) revisit an old issue with the General Aptitude Test Battery (GATB) studies underlying Schmidt and Hunter's (1998) .51 validity estimate. The GATB was a career guidance tool used by state unemployment offices to refer the unemployed to employers. Unlike typical selection settings, employers were not recruiting and selecting applicants based on GATB scores and there were no job-specific local applicant pools. Use of a normative SD from the entire workforce was tenable for the GATB as unemployed jobseekers, who were recently laid off, likely represented the U.S. workforce. When career counselors used the GATB to help people choose jobs, the "applicant pool" was the U.S. workforce and people wishing to enter it.

Sackett et al. (2022) critique Hunter's (1983a) use of (a) a national workforce SD instead of local job-specific applicant SDs and (b) corrections for predictive versus concurrent studies. They state these are "previously unnoticed flaws," but a National Academy of Sciences report covered the SD issue (Hartigan & Wigdor, 1989, pp.166–7). Later studies show this practice only slightly lowers applicant pool SDs (Lang et al., 2010; Sackett & Ostgaard, 1994). Schmidt et al. (2007) discussed this and cited Pearlman et al.'s (1980) finding of similar observed validities for predictive ($r = .23$) and concurrent studies ($r = .21$), suggesting similar range restriction for both. Sackett et al. (2022) also asserted Hunter's (1983a) .67 u_x value was implausibly low; but no solid explanation (e.g., educational requirements or self-sorting) of why a .67 u_x occurred exists. We reviewed jobs from the GATB studies and concluded that most were entry-level and only 4% of the studies had jobs requiring a college/advanced degree. Because people cannot estimate their GMA well (Freund & Kasten, 2012), it is unlikely they would self-sort into jobs based on GMA.

Overlooked validity evidence from other job performance criteria

Sackett et al.'s (2022, 2023) validity estimates focus on supervisory ratings, which may be deficient (SIOP *Principles*, 2018) because these may not reflect employees' job knowledge.

Note. The views expressed in this paper are those of the authors and do not necessarily reflect the views of U.S. Customs and Border Protection, the U.S. Federal Government, or any agency of the U.S. federal government. Portions of this paper were presented at the 2023 meeting of the Society for Industrial and Organizational Psychology.

© The Author(s), 2023. Published by Cambridge University Press on behalf of the Society for Industrial and Organizational Psychology.

Table 1. Meta-Analytic Validities for Intercorrelations of Workplace Criteria from Hayes et al. (2003)

Workplace criterion measure	Training	Work simulation	Supervisor rating
Training	$r_{yy} = .8$.551	.349
Work simulation	.369	$r_{yy} = .8$.349
Supervisor rating	.197	.197	$r_{yy} = .6$

Note. Workplace criterion measures include scores in training, scores for work simulations, and supervisor ratings. Values along the diagonal are default generally accepted reliability estimates for each workplace measure. Values below the reliability diagonal are uncorrected zero-order correlations between each workplace measure. Correlations above the reliability diagonal are corrected for restriction of range ($u = .8072$) and measurement error. For the observed correlation between training performance and work simulation performance, $k = 4$, $n = 1,743$, $SD = .0799$; for the observed correlation between training performance and supervisor ratings, $k = 5$, $n = 2,003$, $SD = .0813$; and for the observed correlation between work simulation performance and supervisor ratings, $k = 6$, $n = 2,991$, $SD = .0689$.

One cannot dismiss GMA as a predictor based on “low” correlations with supervisory ratings without considering the *predictor* deficiency of not including the effect of GMA on acquisition and transfer of job knowledge. Hunter (1983b) demonstrated that job knowledge was the best predictor of job performance. Huang et al.’s (2015) meta-analysis showed that GMA is the best predictor of transfer of training specifically for maximal performance. This spurs a question Sackett et al. (2023) do not address: what is the construct validity of supervisor ratings as a criterion? Fortunately, data on objective criterion measures are available in validity studies. In Table 1, we show Hayes et al.’s (2003) meta-analytic correlations indicating that supervisory ratings are somewhat distinct from work simulations and training performance. Hayes et al., (2003; Hayes & Reilly, 2002) also meta-analyzed the criterion validity of reasoning tests. As shown in Table 2, their validities for supervisory ratings were similar to Sackett et al.’s (2022, 2023). However, validities were much higher when work simulations (i.e., low-fidelity simulations of task performance and procedural knowledge) were used as criteria.

McHenry et al. (1990) reported similar results from the U.S. Army’s Project A. GMA predicted a supervisory/peer job performance ratings method factor with a range restriction (but not criterion unreliability) corrected validity of .15. GMA predicted general soldiering proficiency with a validity of .65 corrected for range restriction (.47 uncorrected) and core technical proficiency with a validity of .63 corrected for range restriction (.43 uncorrected). These two criteria were combinations of job and training knowledge test scores, supervisory and peer ratings, and hands-on performance test (HOPT) scores.

Cucina et al. (2023a) meta-analyzed GMA’s prediction of HOPTs using data from four U.S. military branches. Multivariate range restriction corrections were applied in this meta-analytic database and the Project A database, a correction endorsed by a National Academy of Sciences review (Wigdor & Green, 1991) and corroborated by Held et al. (2015) and Cucina et al. (2023a). Cucina et al. (2023a) found an operational validity of .44 for the Armed Forces Qualifying Test and .55 for aptitude indices (i.e., linear combinations of cognitive tests).

Hunter (1983b) meta-analyzed the validity of GMA for supervisory ratings and HOPTs using non-GATB data. Using his data, we found weighted average validities (corrected for criterion unreliability but not range restriction) of .49 ($n = 3281$; $k = 11$) using HOPTs and .27 using supervisory ratings ($n = 3,605$; $k = 12$). Ree et al. (1994) reported a .42 meta-analytic validity for GMA with interview-based job performance measures correcting for range restriction.

Sackett et al. (2023) critiqued the use of HOPTs as a criterion stating these were measures of maximal, not typical, performance. It is unclear if this is true or why it is an issue. Validity study participants are often told that their data will be used only for research purposes, which may reduce their motivation levels from maximal to typical. McHenry et al. (1990) reported that temperament/personality had validities of .26 for core technical proficiency and .25 for general soldiering proficiency. These validities are typical for personality and are higher than the .18 validity obtained using the supervisory/peer ratings method factor criterion. This suggests that

Table 2. Meta-analytic validities for reasoning tests

Reasoning predictor	<i>k</i> & <i>N</i>	Criterion-related validity coefficient		
		$r_{obs.}$	r_{x, T_y}	ρ_{ov}
		Uncorrected	Corrected for criterion unreliability	Corrected for criterion unreliability and range restriction
Criterion: training performance				
Nonverbal	<i>k</i> = 12, <i>N</i> = 12,872	.181	.202	.248
Quantitative	<i>k</i> = 25, <i>N</i> = 3,718	.417	.466	.553
Verbal	<i>k</i> = 26, <i>N</i> = 3,487	.410	.459	.544
LBM	<i>k</i> = 18, <i>N</i> = 7,533	.456	.509	.599
Criterion: work simulation				
Nonverbal	<i>k</i> = 5, <i>N</i> = 1,229	.318	.355	.429
Quantitative	<i>k</i> = 6, <i>N</i> = 2,491	.470	.525	.616
Verbal	<i>k</i> = 9, <i>N</i> = 3,213	.379	.424	.506
LBM	<i>k</i> = 5, <i>N</i> = 2,512	.455	.509	.598
Criterion: supervisory ratings				
Nonverbal	<i>k</i> = 11, <i>N</i> = 1,939	.107	.138	.170
Quantitative	<i>k</i> = 17, <i>N</i> = 5,575	.198	.255	.313
Verbal	<i>k</i> = 18, <i>N</i> = 4,363	.143	.185	.227
LBM	<i>k</i> = 11, <i>N</i> = 4,149	.168	.217	.267

Note. All predictors were measures of reasoning (i.e., nonverbal reasoning, etc.); LBM: Logic-Based Measurement (LBM), which are tests developed using an established logical framework for reasoning items (Simpson et al., 2007); *k* = number of studies; *N* = combined sample sizes.

objective performance measures are not entirely measures of maximal performance. Further, work simulations/HOPTs are measures of core task proficiency, which is an important criterion.

Conceptual paradoxes for GMA research findings

Believing a lowered GMA validity estimate leads to five empirical paradoxes.

Paradox 1: GMA predicts firearms proficiency better than it does overall job performance

There are numerous cognitive decisions and activities associated with job performance. Hunt and Madhyastha (2012) identified a large GMA-based factor in job analysis data from O*NET. A job analysis of 105 Federal government jobs identified 42 core tasks and the competency with the best linkage to those tasks was reasoning, which Carroll (1993, p. 196) stated is “at or near the core of what is ordinarily meant by intelligence” (Pollack et al., 1999; Simpson et al., 2007). Yet, in Cucina et al.’s (2023b) largest dataset, reasoning had an operational validity of .268 (*n* = 14,892) in predicting how well individuals could aim and shoot a handgun at a stationary target. It is paradoxical that GMA had higher validity for a largely psychomotor task than for overall job performance if all that matters in performance are supervisory ratings.

Paradox 2: Similarly corrected GMA validities for training performance are corroborated

Schmidt and Hunter's (1998) .56 validity estimate for GMA tests with training performance used the GATB studies and the same range restriction correction that Sackett et al. (2022, 2023) critiqued. However, studies using other correction processes yielded similar validities, including results in Table 2. Brown et al. (2006) reported a .546 validity for GMA across 10 Navy training schools ($n = 26,097$). Welsh et al., (1990, p. 36) reported a range restriction corrected validity of .44 for the AFQT with final school grades ($n = 224,048$ cases)¹.

Paradox 3: There is no free lunch—GMA test proxies can be g-loaded

Sackett et al. (2023) report a .40 validity for job knowledge tests compared to the .23 validity for GMA tests. However, job knowledge tests are *g*-loaded. Using Hunter's (1983b) GMA-job knowledge correlations, we computed an average correlation of .50 ($n = 3,372$; $k = 11$). This size correlation is typical of those between the ASVAB and GATB subtests. Paradoxically, an employer eschewing GMA in favor of job knowledge is unwittingly testing applicants' GMA.

Paradox 4: GMA's validity is decreasing yet job complexity is increasing

The validity of GMA tests has purportedly decreased .51 to .31 to .23. Many of the jobs in the GATB validity studies were manufacturing and medium complexity jobs. The number of U.S. employees in manufacturing has decreased significantly since the 1970s (Gascon, 2022). Today's U.S. economy is more focused on knowledge work and there is an increased use of technology in blue collar jobs. This should lead to higher job complexity and higher GMA validities because job complexity moderates GMA's validity (Schmidt & Hunter, 1998). This is paradoxical as the validity of GMA is decreasing yet job complexity is increasing in the U.S.

Paradox 5: The folly of selecting for contextual performance but expecting proficiency

Sackett et al. (2023) state that contextual criteria are not as predictable by GMA as are task-based criteria. Although supervisory ratings are easily obtained and provide an aura of independent authoritative judgment, they are clouded by social/organizational factors (Murphy & Cleveland, 1995), impacted by supervisors' opportunity to observe performance (MacLane et al., 2020), and biased by other factors (e.g., subjectivity, criterion deficiency, poorly developed rating scales; Courtney-Hays et al., 2011). Practices in how ratings are collected impact criterion-related validity (Grubb, 2011). Sackett et al. (2023) caution against comparing meta-analytic validities as "we do not have clear understanding of the specific components underlying performance ratings"; this is good advice when selecting for "non-cognitive skills" but paradoxical (or worse) when expecting people to be trainable and develop task proficiency.

Conclusion

Criticisms of the range restriction correction procedure for the GATB validity studies are not new. The correction procedure could be tenable for the original use of GATB scores which was to predict how well individuals representative of the U.S. workforce would perform in different jobs. Making an inferential leap to local selection settings may result in lower validities when supervisory ratings serve as criteria, however, validities using objective job performance measures (i.e., work samples, work simulations, and HOPTs) are still near .51.

¹Using Pearlman et al.'s (1980) .80 estimate to correct for criterion unreliability yields an operational validity of .49.

References

- Brown, K.G., Le, H., & Schmidt, F.L. (2006). Specific aptitude theory revisited: Is there incremental validity for training performance? *International Journal of Selection and Assessment*, *14*, 87–10.
- Carroll, J.B. (1993). *Human cognitive abilities*. Cambridge University Press.
- Courtney-Hays, J.M., Carswell, J.J., Cucina, J.M., Melcher, K.M., & Vassar, A. (2011). *Variety is the spice of validation: Moving beyond “traditional” criteria*. Panel discussion at the 26th annual SIOP Conference, Chicago, IL.
- Cucina, J.M., Burtneck, S.K., De la Flor, M.E., Walmsley, P.T., & Wilson, K.J. (2023a). *Meta-analytic validity of cognitive ability for hands-on military job performance*. Poster presented at the 38th annual SIOP Conference, Boston, MA.
- Cucina, J.M., Wilson, K.J., Hayes, T.L., Walmsley, P.T., & Votraw, L.M. (2023b). Is there a g in gunslinger? Cognitive predictors of firearms proficiency. *Intelligence*, *99*, 101768.
- Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*, 296–321.
- Gascon, C.S. (2022, July 13). *Labor constraints remain greatest challenge for resurgent manufacturing sector*. *The Regional Economist*. Federal Reserve Bank of St. Louis.
- Grubb, A.D. (2011) Promotional assessment at the FBI: How the search for a high-tech solution led to a high-fidelity low-tech simulation. In S. Adler & N.T. Tippins (Eds.), *Technology-Enhanced Assessment of Talent* (pp. 338–354). Jossey-Bass.
- Hartigan, J.A & Wigdor, A.K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. National Academy Press.
- Hayes, T.L., McElreath, J., & Reilly, S.M. (2003). *The criterion-related validity of logic-based measurement and reasoning tests in public sector merit-based selection systems* (Report Number 03-01). U.S. Department of Homeland Security.
- Hayes, T.L., & Reilly, S.M. (2002, April). The criterion-related validity of logic-based measurement tests: The SIOP conference paper. In T.L. Hayes (Chair), *The validity of logic-based measurement for selection and promotion decisions*. Symposium conducted at the 17th annual SIOP Conference, Toronto, Canada.
- Held, J.D., Carretta, T.R., Johnson, J.W., & McCloy, R.A. (2015). *Technical guidance for conducting ASVAB validation/standards studies in the U.S. Navy* (Technical Report NPRST-TR-15-2). Navy Personnel Research, Studies, and Technology.
- Huang, J.L., Blume, B.D., Ford, J.K., & Baldwin, T.T. (2015). A tale of two transfers: Disentangling maximum and typical transfer and their respective predictors. *Journal of Business and Psychology*, *30*(4), 709–732.
- Hunt, E., & Madhyastha, T.M. (2012). Cognitive demands of the workplace. *Journal of Neuroscience, Psychology, and Economics*, *5*, 18–37.
- Hunter, J.E. (1983a). *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. (Technical Report No. USES-TRR-45). U.S. Department of Labor.
- Hunter, J.E. (1983b). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedick, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257–266). Erlbaum.
- Lang, J.W.B., Kersting, M., & Hulsheger, U.R. (2010). Range shrinkage of cognitive ability test scores in applicant pools for German governmental jobs: Implications for range restriction corrections. *International Journal of Selection and Assessment*, *18*, 321–328.
- MacLane, C.N., Cucina, J.M., Busciglio, H.H., & Su, C. (2020). Supervisory opportunity to observe moderates criterion-related validity estimates. *International Journal of Selection and Assessment*, *28*, 55–67.
- McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, M.A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, *43*, 335–354.
- Murphy, K.R., & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage Publications, Inc.
- Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, *65*, 373–406.
- Pollack, L., Simons, C., & Patel, R. (1999). *Federal professional and administrative occupations: An application of the Multipurpose Occupational Systems Analysis Inventory–Closed-ended (MOSAIC)*. Office of Personnel Management.
- Ree, M.J., Earles, J.A., & Teachout, M.S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology*, *79*, 518–524.
- Sackett, P.R., & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology*, *79*, 680–684.
- Sackett, P.R., Zhang, C., Berry, C.M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*, 2040–2068.
- Sackett, P.R., Zhang, C., Berry, C.M., & Lievens, F. (2023). Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *16*(3), 283–300. doi: [10.1017/iop.2023.24](https://doi.org/10.1017/iop.2023.24)
- 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.
- Schmidt, F.L., Le, H., Oh, I.-S., & Shaffer, J. (2007). General mental ability, job performance, and red herrings: Responses to Osterman, Hauser, and Schmitt. *Academy of Management Perspectives*, *21*, 64–76.

- Simpson, R., Nester, M.A., & Palmer, E.** (2007). *The validity of logic-based tests*. Paper presented at the annual meeting of the International Public Management Association for Human Resources Assessment Council (IPMAAC), St. Louis, MO.
- Welsh, J.R., Jr., Kucinkas, S.K., & Curran, L.T.** (1990). *Armed Services Vocational Battery (ASVAB): Integrative review of validity studies* (Report No. AFH R L-TR-90-22). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.
- Wigdor, A.K., & Green, B.F., Jr.**, (1991). *Performance assessment for the workplace, Vol. 1; Vol. 2: Technical issues*. National Academy Press.