Revisiting Assessment Validity for Predicting Overall Job Performance

Paul Sackett
University of Minnesota

Part 1: The “Received Wisdom” re Selection Procedure Validity

- **Question: How do we “know what we know” about validity?**
- **Answer: Meta-analyses that compile validity data across studies**
  - Collect studies; compute the average
  - If possible, correct for restriction of range and unreliability in the criterion
  - Compute new average of these corrected validities

The “punchline”: Schmidt and Hunter (1998)

Table 1

Predictive Validity for Overall Job Performance of General Mental Ability (GMA) Scores Combined With a Second Predictor Using (Standardized) Multiple Regression

Personnel measures	Validity (r)	Multiple R	Gain in validity from adding supplement
GMA tests[a]	.51		
Work sample tests[b]	.54	.63	.12
Integrity tests[c]	.41	.65	.14
Conscientiousness tests[d]	.31	.60	.09
Employment interviews (structured)[e]	.51	.63	.12
Employment interviews (unstructured)[f]	.38	.55	.04
Job knowledge tests[g]	.48	.58	.07
Job tryout procedure[h]	.44	.58	.07
Peer ratings[i]	.49	.58	.07
T & E behavioral consistency method[j]	.45	.58	.07
Reference checks[k]	.26	.57	.06
Job experience (years)[l]	.18	.54	.03
Biographical data measures[m]	.35	.52	.01
Assessment centers[n]	.37	.53	.02
T & E point method[o]	.11	.52	.01
Years of education[p]	.10	.52	.01
Interests[q]	.10	.52	.01
Graphology[r]	.02	.51	.00
Age[s]	−.01	.51	.00

Are the Schmidt and Hunter results credible?

Table 1

Predictive Validity for Overall Job Performance of General Mental Ability (GMA) Scores Combined With a Second Predictor Using (Standardized) Multiple Regression

Did not report uncorrected validity

Personnel measures	Validity (r)	Multiple R	Gain in validity from adding supplement
GMA tests[a]	.51		
Work sample tests[b]	.54	.63	.12
Integrity tests[c]	.41	.65	.14
Conscientiousness tests[d]	.31	.60	.09
Employment interviews (structured)[e]	.51	.63	.12
Employment interviews (unstructured)[f]	.38	.55	.04
Job knowledge tests[g]	.48	.58	.07
Job tryout procedure[h]	.44	.58	.07
Peer ratings[i]	.49	.58	.07
T & E behavioral consistency method[j]	.45	.58	.07
Reference checks[k]	.26	.57	.06
Job experience (years)[l]	.18	.54	.03
Biographical data measures[m]	.35	.52	.01
Assessment centers[n]	.37	.53	.02
T & E point method[o]	.11	.52	.01
Years of education[p]	.10	.52	.01
Interests[q]	.10	.52	.01
Graphology[r]	.02	.51	.00
Age[s]	−.01	.51	.00

Uncorrected = .25

Into the weeds: how does meta-analysis work?

Ideally, for each study you have:

- **Validity coefficient**,
- Estimate of the amount of **range restriction** in the study, and
- Estimate of **criterion reliability** for each study

But range restriction and reliability information is typically only available for a **small subset of studies**

Schmidt and Hunter's clever solution: compute average range restriction and criterion reliability for studies that report it…

Then correct average validity for average level of range restriction and average level of unreliability

Does this make sense?

- Yes – **if** assumptions are met
- **Key assumption** - studies providing range restriction and unreliability information are a random or representative sample of all the collected validity studies
- My colleagues and I argue that the **assumption does not hold**, and the Schmidt-Hunter corrections are **substantially inflated**.

Setting the Stage

Predictive Validation

- Test applicants
- Gather performance after sufficient time on the job
- Range restriction will commonly have large effects

Concurrent Validation

- Test incumbents
- Gather performance simultaneously
- Range restriction will generally be smaller as predictor of interest not used in hiring.

Restricted Test Standard Deviation as a Function of Selection Method

	Selection Ratio	
	0.50	0.10
Select on *x*	**0.60**	**0.41**
Select on *z*		
r_{zx}		
0.9	**.69**	**.56**
0.8	**.77**	**.68**
0.7	**.83**	**.76**
0.6	**.88**	**.83**
0.5	**.92**	**.89**
0.4	**.95**	**.93**
0.3	**.97**	**.96**
0.2	**.99**	**.98**
0.1	**1.00**	**1.00**

The key insights

- We can usually get the needed info for a range restriction correction **only from predictive studies**
  - Need to know range of scores in the applicant pool to make a correction
- On average, about **80% of validity studies are concurrent**
- Standard meta-analysis applies the correction factor derived from the **20% predictive studies** to the **80% concurrent studies** as well
- So we routinely **violate the assumption** that studies with range restriction info are a random sample
- Result: a **radical overcorrection**. We have overstated validity by a large margin for several decades

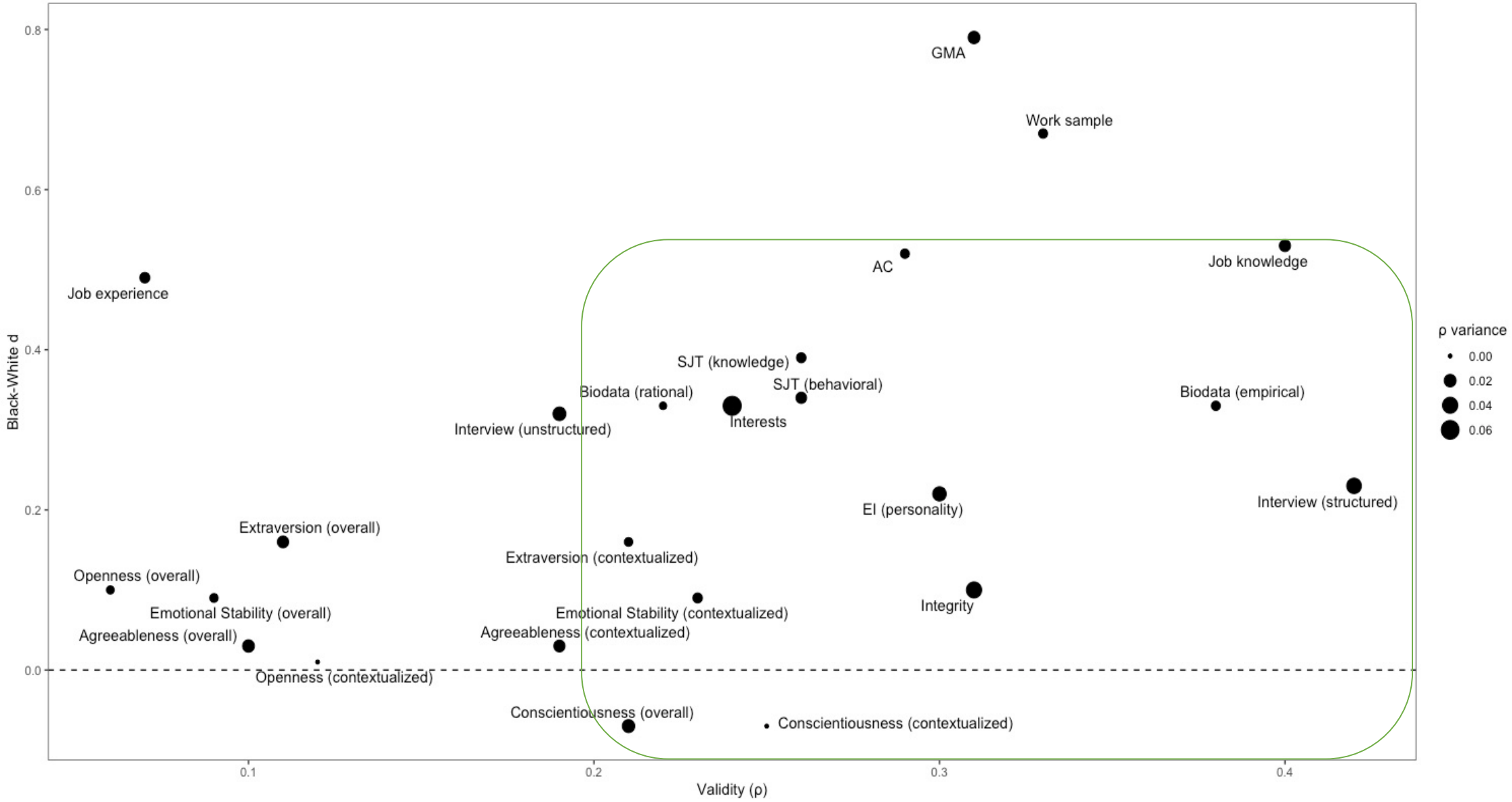
So: we started over and revisited existing meta-analyses

- My great team: Charlene Zhang, Chris Berry, and Filip Lievens
  - Published in JAP: Sackett, Zhang, Berry, & Lievens, 2022)
- **Revisited meta-analyses** of a broad range of predictors
- Made **appropriate corrections** if we **had** the needed data
- Made **no correction** if we **lacked** the needed data
  - We argue for being **conservative**: don't correct if you do not have the needed data
  - And with most studies concurrent, at most we underestimate by a couple correlation points

We re-estimated validity. And added in subgroup (Black-White) mean differences

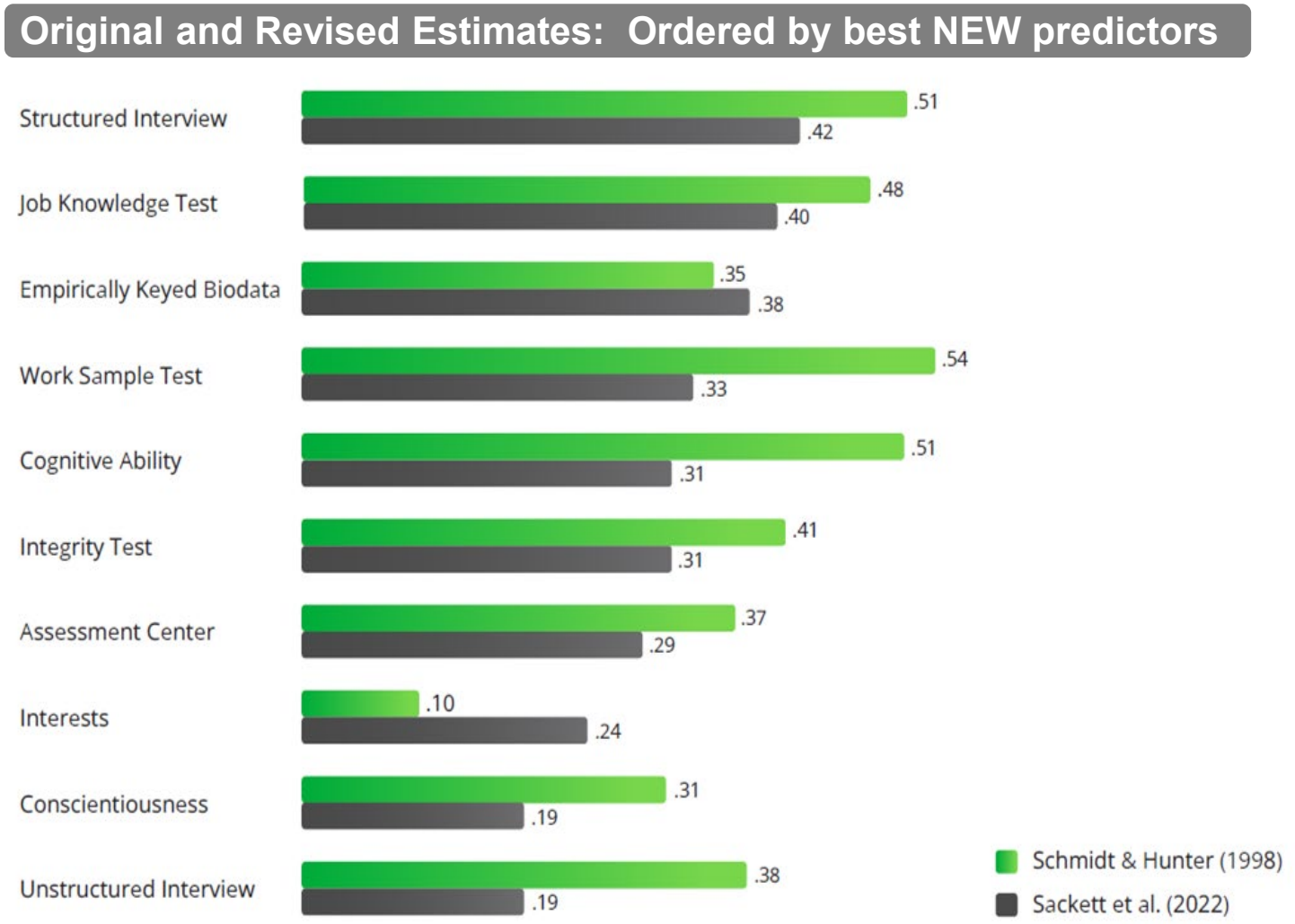
Predictor	Schmidt & Hunter (1998) Validity Estimate	Current Validity Estimate (ρ)	B-W d
Employment interviews – structured	0.51	0.42	0.23
Job knowledge tests	0.48	0.40	0.53
Empirically keyed biodata	0.35	0.38	0.33
Work sample tests	0.54	0.33	0.67
Cognitive ability tests	0.51	0.31	0.79
Integrity tests	0.41	0.31	0.10
Personality-based EI		0.30	0.22
Assessment centers	0.37	0.29	0.52
SJT – knowledge		0.26	0.39
SJT - behavioral tendency		0.26	0.34
Conscientiousness – contextualized		0.25	-0.07
Interests	0.10	0.24	0.33
Emotional Stability – contextualized		0.23	0.09
Ability-based EI		0.22	
Rationally keyed biodata		0.22	0.33
Extraversion – contextualized		0.21	0.16
Conscientiousness- overall	0.31	0.21	-0.07
Employment interviews – unstructured	0.38	0.19	0.32

Same info in a graph



Which predictors show biggest change in the validity estimate?

- Work samples: ⇩.21
- Cognitive ability: ⇩.20
- Unstructured interview: ⇩.19
- Interests: ⇧.14
- Conscientiousness: ⇩.12
- Integrity tests: ⇩.10
- Structured interview: ⇩.09



Job-specific measures top the list

- Structured interview
- Job knowledge tests
- Empirically-keyed biodata
- Work samples
- Assessment centers (using updated validity estimate)

Followed by

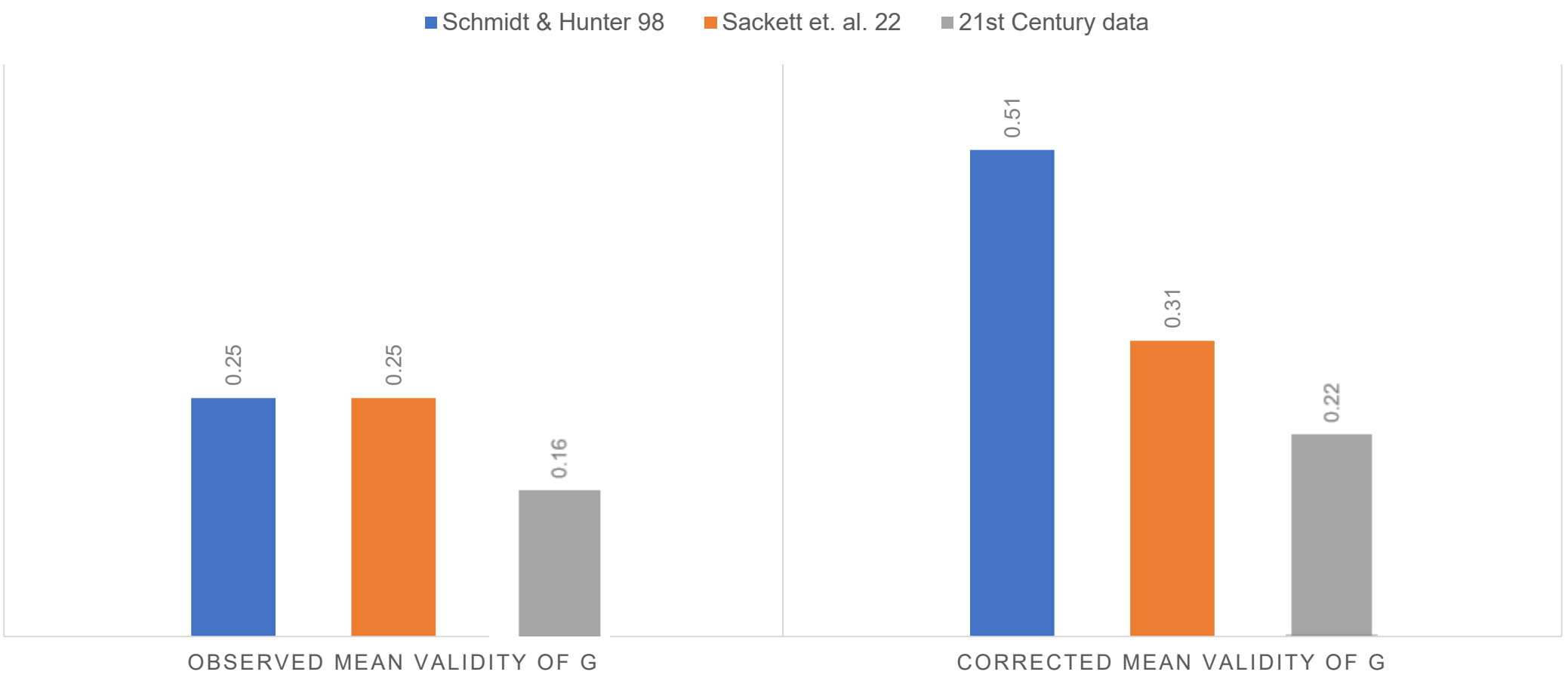
- Cognitive ability
- Integrity tests
- Personality-based emotional intelligence

Part 2: More current info on cognitive ability validity

Sackett, Demeke, Bazian, Griebe, Priest, and Kuncel (2024):

- Schmidt and Hunter's analysis uses **data >50 years old**
- Examined **21st century studies** - between 2000 and 2021
- 153 validation samples (40k+ participants) (published and consultant provided)

Part 2: More current info on cognitive ability validity

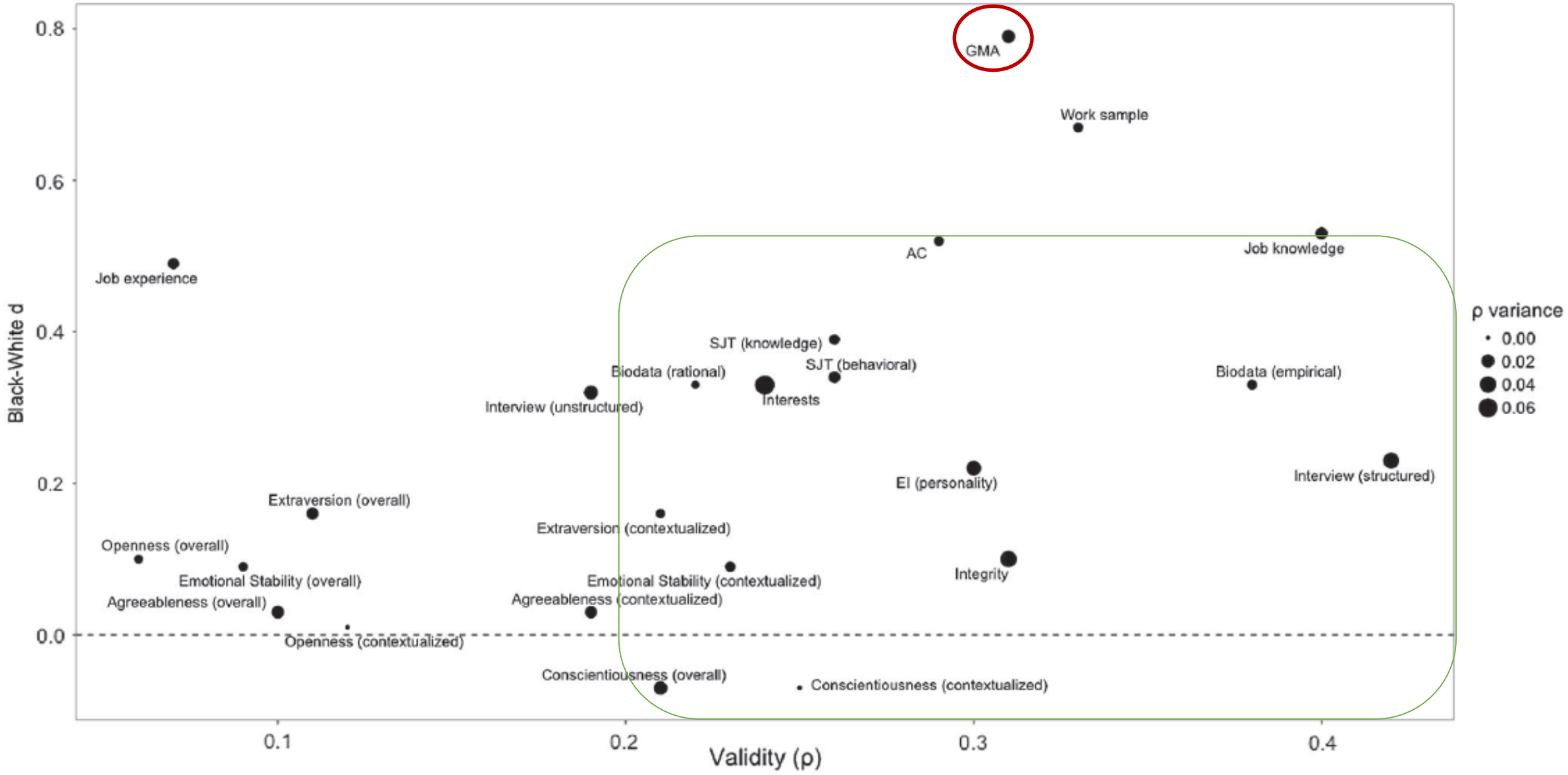


Two more new meta-analyses

- Chris Nye: validity studies from 1990-2012
- He get a mean value of .23
- Piers Steel: post-1990 studies using the Wonderlic Personnel Test
- He gets mean of .16 and .20 for two subsets of studies

Figure 2

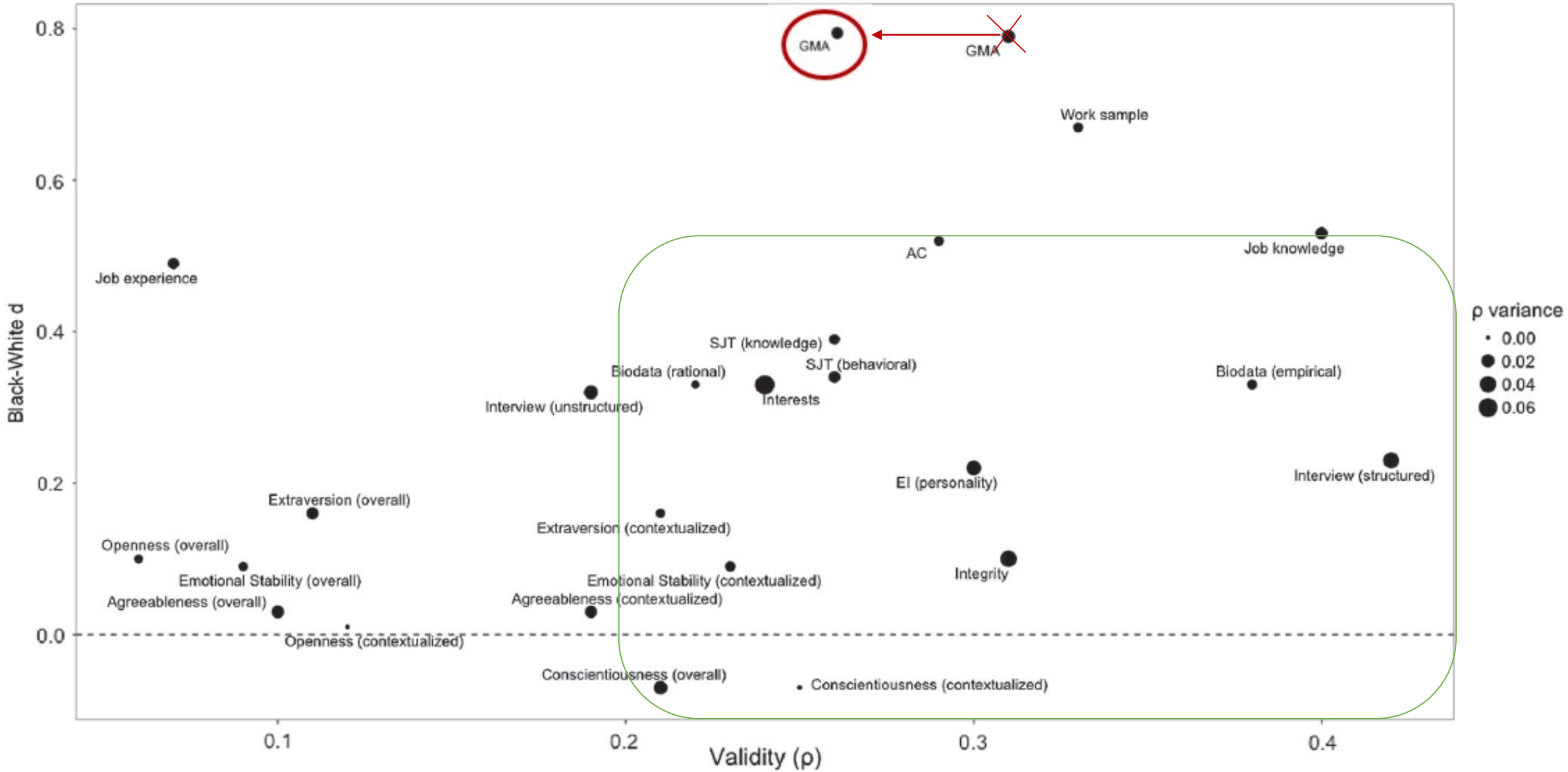
A Visual Summary of Common Selection Procedures' Validity, Validity Variance, and, Black–White d



Note. GMA = General Mental Ability. AC = Assessment Center. SJT = Situational Judgment Test. EI = Emotional Intelligence.

Figure 2

A Visual Summary of Common Selection Procedures' Validity, Validity Variance, and, Black–White d



Note. GMA = General Mental Ability. AC = Assessment Center. SJT = Situational Judgment Test. EI = Emotional Intelligence.

Speculation as to the drop in validity

Are **different jobs** studied?	• Yes – old data relied heavily on **manufacturing** jobs • New data involve jobs with **larger interpersonal and teamwork** demands
Are **criteria different**?	• We suspect so: "overall performance" likely incorporates effectiveness in these interpersonal and teamwork domains
Have **applicant pools** narrowed?	• Yes – evidence of narrower applicant pool SDs, relative to overall workplace SD

Speculation as to the drop in validity

Are newer tests of lower quality?	• Get same validity estimate with “legacy” tests and newer tests
Do studies by consultants get better results?	• No; in fact, consultant studies get slightly lower mean validity
Are results biased due to studies focusing on incremental validity of new predictors over cognitive ability?	• No; in fact studies that that do not find incremental validity over cognitive ability show slightly higher mean validity

Part 3: New insights into the validity/diversity dilemma

- Berry, Lievens, Zhang, and Sackett (2024) examined the validity of **composites** of predictors
- Updated the Bobko/Roth classic matrix; with six predictors:
  - Cognitive ability
  - Conscientiousness
  - SJT
  - Structured interview
  - Integrity
  - Biodata
- Updated **validity** for each, and **intercorrelation** among each

With **old** values:

composites that did not include cognitive ability were on average **.10-.20 points lower** in validity

With **new** values:

composites that did not include cognitive were on average **equal in validity** to those that did not

So much less tension between validity and diversity than we'd thought

Part 4: Settings in which high validity is found for cognitive ability

- Cognitive ability has been found a strong predictor of performance in educational and training settings as well as technical proficiency
  - e.g., lots of large-N studies of relationships between military's AFQT and training school grades and hands-on performance tests, with validities in .40's and .50's
- No argument here: settings with a large post-hire training/learning component are prime candidates for heavier reliance on cognitive ability

Part 4: Settings in which high validity is found for cognitive ability (continued)

- Technical proficiency reflects **maximum performance**: Short term performance attainable with effort maximized
- Military also measures effort, reflecting **typical performance** (e.g., task persistence, diligence)
  - This correlates .28 with Technical Proficiency

So the big question: do you want to predict “can do” vs. “will do”?

A new view of cognitive ability

Old: high validity makes it a key predictor, despite large subgroup differences

New: validity is solid, but it does not stand out. The large subgroup differences make its use harder to justify

BUT: in terms of “off the shelf” predictors, it remains one of our most valid predictors

AND: here we are predicting overall job performance. Cognitive ability is also used to predict training performance, or narrower facets of overall performance.

Part 5: New insights into the Hunter validity estimate of .51

- Ones and Viswesvaran (2023) call Sackett et al. (2022) “mere speculation”, and endorse relying on “facts”
  - Call for continuing to use the Hunter .51 value based solely on GATB studies
- Problem 1 with relying on Hunter:
  - Hunter used an “early” set of GATB studies. His work predated a “later” set of several hundred GATB studies, for which I have the raw data.
  - These data produce a) lower observed validity (.21 vs. .25), and a u ratio of .96: minimal range restriction
  - Ones and Vish ignore these studies

Part 5: New insights into the Hunter validity estimate of .51

- Problem 2 with relying on Hunter
  - This ignores the three contemporary meta analyses we've just discussed

- Problem 3: Overlooked evidence of a file drawer problem.
  - Bemis (1968) reports that GATB studies not obtaining successful findings were set aside and not "published"
  - Bemis writes that about 10% of studies were set aside
  - With small N's it is reasonable that a good number of studies would produce low validity estimates
  - Recall that the Hunter GATB studies had mean observed validity of .25; newer GATB studies have mean observed validity of .21.

How consequential is the historic overestimation of cognitive ability validity?

- For decades we have given cognitive ability a dominant role in a great many selection systems based on the belief that its validity was so high that we needed to use it, despite high adverse impact
- We now see we can get comparable validity from a composite of predictors that does not include cognitive ability
- The uncomfortable question: over the decades how many job candidates from protected groups have been screened out based on systems that overweighted cognitive ability?

Bottom line

- **Our predictors are useful, but less so than the literature has reported**
- **But composites of multiple predictors still fare quite well**
- **The claim that cognitive ability is our best predictor needs reconsideration**
- **Less tension between validity and diversity than we envisioned in the past**