Is Metacognition a Better predictor than IQ for Academic Achievement?

A Methodological and Psychometric Critique of
Gomes, Golino & Menezes (2014), *Psychology*, 5, 1095–1110

Antonio N. Weber
email: anw@anweber.org

July 6, 2026

Abstract

In their disturbingly widely cited paper, Gomes, Golino and Menezes (2014) report that a general metacognitive ability explains 20.43% of general academic achievement against 7.45% for fluid intelligence, and that a specific metacognitive ability explains 42.12% of specific (arithmetic) achievement against 1.12% for fluid intelligence, with ratios advertised as "nine times" and "15:1 at worst" in favour of metacognition. My article shows that these results are manufactured by several complementary defects, possibly resulting from bias. First, the scoring rule of the mathematics appraisal test mathematically penalizes accurate self-knowledge: a student with *zero* metacognition who simply declares certainty of success on every item is guaranteed a higher expected "metacognition" score than a perfectly self-aware student, at every accuracy level below 100% (Proposition 1); the score is, in substance, an arithmetic achievement measure. Second, the criterion for specific achievement is a parallel arithmetic test, so the headline association sits comfortably below the ceiling implied by shared accuracy alone. Third, the intelligence covariate, an obscure, short, and fluid-only battery ($\alpha = .75$) in a single-school, range-restricted sample, is reproduced to three decimal places by applying standard range-restriction and unreliability corrections to the conventional population correlation of $\rho = .50$, so the data contain no evidence against the primacy of intelligence. Model-side choices (factor scores treated as error-free, modification-index respecification, a directed path from metacognition *to* intelligence, and effect-size ratios computed between the bounds of different confidence intervals) further illustrate the poor quality of the study. All computations are stated explicitly and all empirical quantities are cited to page.

1 The paper and its claims

Gomes, Golino and Menezes [1] administered a fluid-intelligence battery, two metacognitive tests, and an arithmetic test to 684 students (grades 6–12) of a single private school in Belo Horizonte, Brazil (p. 1098), and combined these with annual grades in Mathematics, Portuguese, Geography, and History in a structural equation model. Their central estimates (p. 1103), reproduced in Table 1, ground three claims: that metacognition has incremental validity over intelligence; that in "the best of the scenarios" general metacognitive ability explains general achievement "approximately nine times more" than fluid intelligence (p. 1104); and that specific metacognitive ability outpredicts intelligence "at worst in 15:1" and prior mathematical knowledge "at worst in 18:1" for specific achievement (p. 1104).

The Critique is organized around the three load-bearing elements of the design: the scoring rule of the metacognition measure (Section 2), the relation between predictor and criterion

(Section 3), and the construction of the intelligence covariate (Section 4); Section 5 collects the model-side choices and Section 6 states what a probative design would require.

2 The scoring rule penalizes accurate self-knowledge

2.1 The rubric

The *Appraisals Ability on Mathematics Expressions* test (AAME) asks students to solve 18 arithmetic expressions and, after each, to judge their probability of success on a four-point scale (p. 1099). The published scoring rule (p. 1099) is summarized in Table 2. Its defining feature: a student who is *sure of having failed* receives 0 points *even when that judgment is correct*. The authors state the design intention explicitly:"students should be minimally able to solve the item in order to perform the evaluative process" (p. 1099), which means that the contamination of the metacognition score by task performance is not an accident of analysis, but rather a biased choice.

Metacognitive monitoring, as the field defines it, is the correspondence between confidence and performance; a valid monitoring score must reward correct self-assessments of failure as much as correct self-assessments of success. Table 2 does the opposite, and the consequences can be estimated.

2.2 Expected-score analysis

Let p denote a student's probability of solving an arithmetic item correctly, constant across the 18 items for simplicity. Fix a *reporting policy*, i.e., a rule mapping the student's (possibly absent) self-knowledge to a confidence response, and compute the expected total score $E[S]$ under Table 2.

- **Perfect self-knowledge, honest reporting.** The student is always certain and always right about himself: "sure I succeeded" on correct items (4 points), "sure I failed" on incorrect items (0 points):

$$E[S_{\text{cal}}] \;=\; 18\big[\,4p + 0\,(1-p)\big] \;=\; 72\,p. \tag{1}$$

Table 1: Variance-explained estimates reported by Gomes et al. (2014, p. 1103), with their 90% bootstrap confidence intervals. GMA = General Metacognitive Ability; SMA = Specific Metacognitive Ability; GAA/SAA = General/Specific Academic Achievement; Gf = fluid intelligence; MI = Monitoring Indicator (reading test); AI1/AI2 = appraisal indicators built from the easy/difficult arithmetic items.

Path	Interpretation	Variance explained	90% CI
GMA → GAA	metacognition → school grades	20.43%	[12.60, 29.38]
Gf → GAA	intelligence → school grades	7.45%	[3.31, 12.46]
SMA → SAA	metacognition → arithmetic	42.12%	[36.00, 48.44]
MATH → SAA	prior math grade → arithmetic	3.72%	[1.99, 5.86]
Gf → SAA	intelligence → arithmetic	1.12%	[0.32, 2.34]
GMA → MI		66.10%	[49.00, 99.06]
GMA → Gf	metacognition "explains" intelligence	12.85%	[6.05, 21.25]
SMA → AI1	easy-item indicator	90.25%	[81.36, 100]
SMA → AI2	difficult-item indicator	28.09%	[22.66, 33.64]

Table 2: The AAME scoring matrix implied by the verbal rubric of Gomes et al. (2014, p. 1099). Points awarded per item as a function of the confidence judgment and the actual outcome.

Confidence judgment	Item solved correctly	Item solved incorrectly
"Sure I failed"	0	**0**
"Not sure; think I failed"	2	3
"Not sure; think I succeeded"	3	2
"Sure I succeeded"	4	1

- **Zero self-knowledge, blind overconfidence.** The student selects "sure I succeeded" on every item, ignoring all internal evidence (4 points if correct, 1 if not):

$$E[S_{\text{over}}] \;=\; 18\big[\,4p + 1\,(1-p)\big] \;=\; 18 + 54\,p. \tag{2}$$

- **Zero self-knowledge, random confidence.** The student chooses randomly among the four responses. Correct items then earn $(0+2+3+4)/4 = 2.25$ points on average and incorrect items $(0+1+3+2)/4 = 1.5$ points:

$$E[S_{\text{rand}}] \;=\; 18\big[\,2.25\,p + 1.5\,(1-p)\big] \;=\; 27 + 13.5\,p. \tag{3}$$

- **Perfect self-knowledge, optimized reporting.** A self-aware student who games the rubric, i.e. he answers "sure I succeeded" when he knows he is right (4) and "not sure; think I failed" when he knows he is wrong (3, the maximum available for an incorrect item):

$$E[S_{\text{strat}}] \;=\; 18\big[\,4p + 3\,(1-p)\big] \;=\; 54 + 18\,p. \tag{4}$$

Proposition 1 (Blind overconfidence outperforms honest calibration)**.** *For every $p < 1$, $E[S_{\text{over}}] - E[S_{\text{cal}}] = 18(1-p) > 0$: a student with no metacognition at all, i.e. one who always claims certainty of success, always obtains a higher "metacognition" score than a student with perfect, honestly reported self-knowledge. The two coincide only at $p = 1$ i.e. when they both answer every arithmetic question correctly.*

Proposition 2 (Random confidence beats calibration below 46% accuracy)**.** *Setting* (1) *equal to* (3) *gives $72p = 27 + 13.5p$, i.e. $p = 27/58.5 \approx .462$. For any student solving fewer than* 46% *of the items, answering the confidence question* at random *yields a higher expected score than perfect self-knowledge.*

Proposition 3 (Honesty is penalized even given perfect self-knowledge)**.** *Comparing* (4) *with* (1)*: $E[S_{\text{strat}}] - E[S_{\text{cal}}] = (54 + 18p) - 72p = 54(1-p) \geq 0$. The score-maximizing policy for a perfectly self-aware student is to* misreport *his certainty of failure (avoiding the* 0*-scored "sure I failed" response). A test on which the optimal strategy for the best possible metacognizer is to lie about his metacognition is not properly measuring metacognition.*

Figure 1 plots (1)–(4). Two further readings of the figure matter. First, under every policy the expected score is increasing and linear in p, with accuracy slopes of 72, 54, 18 and 13.5 points per unit of p; whatever a student's confidence policy, his score rises mechanically with his arithmetic ability, so between-student variance in the AAME total is dominated by achievement variance. Second, at any fixed p the vertical spread between "perfect" and "zero" metacognition policies is small and *non-monotone in metacognitive quality* (the honest-calibrated line lies *below* two zero-metacognition lines over most of the range). The instrument is, in mathematical substance, an arithmetic achievement test with a confidence-flavoured perturbation.

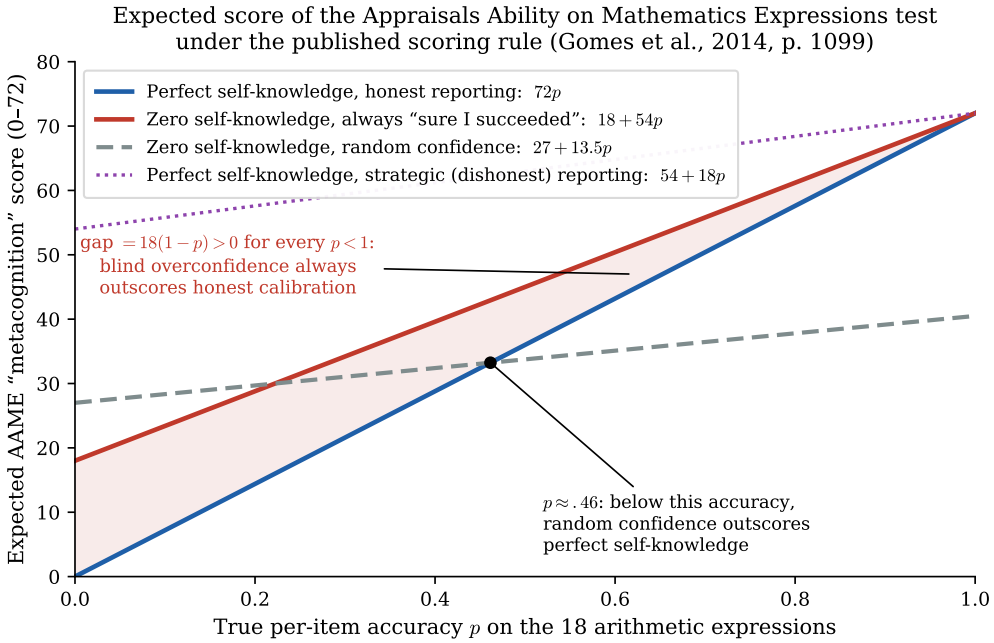

Figure 1: Expected AAME score as a function of true item accuracy p under four reporting policies (equations (1)–(4)). The shaded band is the dominance gap of Proposition 1. Figure computed by the author from the scoring rule on p. 1099 of [1].

The paper's own factor results corroborate this diagnosis: latent Specific Metacognitive Ability explains 90.25% of the indicator built from the *easy* arithmetic items but only 28.09% of the indicator built from the difficult ones (Table 1; Figure 2b). Easy items are where confidence responses sit at ceiling and residual score variance reduces to accuracy variance; the "metacognition" factor lives where metacognition supposedly varies least.

3 The criterion is almost the same test as the predictor

What does the contaminated score predict? "Specific Academic Achievement" is measured by the *Arithmetic Expressions Test*: 18 arithmetic expressions scored pass/fail, with Cronbach's $\alpha = .88$ (p. 1100); imagine my shock upon finding out about this. The appraisal test (the one used to estimate Specific Metacognitive Ability) itself consists in the solving of 18 arithmetic expressions (p. 1099), with $\alpha = .86$ (p. 1100). The paper never states that the two item sets are distinct, and the appraisal instructions ("After solving the item, the respondents must evaluate their probability of success," p. 1099) describe a single solve-then-judge procedure. On my most natural reading the criterion scores are the solving outcomes in the predictor's own administration, a part–whole relation; on the most charitable reading they are parallel forms of the same type of test taken in the same session.

Either way the association is guaranteed in advance. Both totals are functions of the same underlying accuracy p, so their correlation is limited only by measurement error. By the classical Spearman attenuation identity [2], the maximum observable correlation between two fallible measures of a common quantity is

$$r_{\max} = \sqrt{r_{xx}\, r_{yy}} = \sqrt{.86 \times .88} \approx .870. \tag{5}$$

The paper's standardized path from Specific Metacognitive Ability to Specific Academic

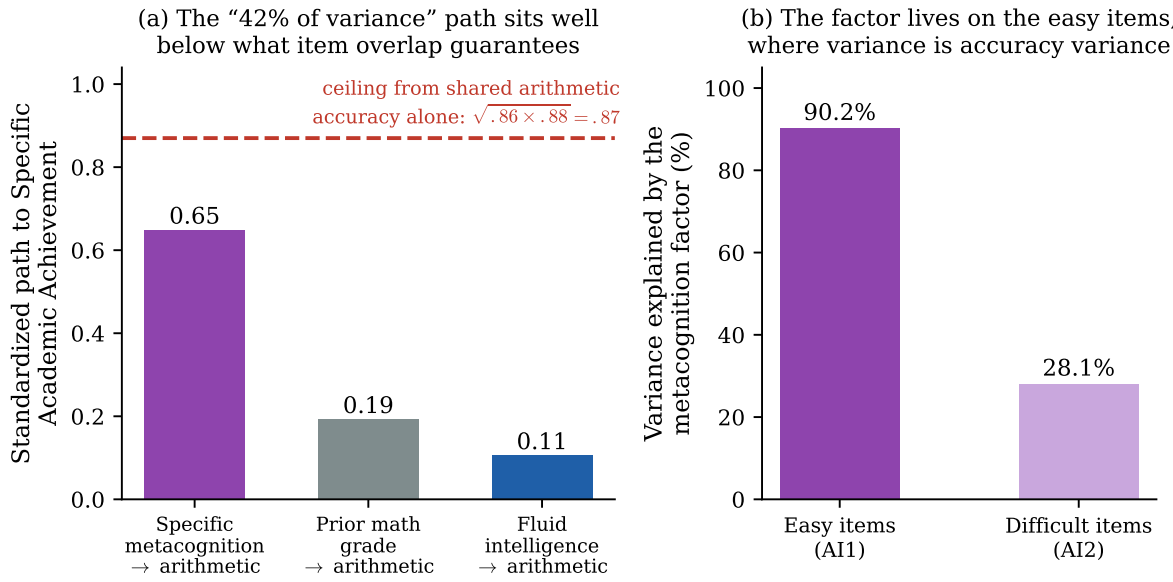

Figure 2: (a) The specific-achievement path against the attenuation ceiling of equation (5): the celebrated association is most likely explicable by shared arithmetic accuracy. (b) Variance in the two appraisal indicators explained by the metacognition factor (p. 1103): the factor is anchored in the easy items, where score variance is accuracy variance.

Achievement is $\sqrt{.4212} \approx .649$, which is comfortably inside the range that shared arithmetic accuracy alone produces, with no metacognitive contribution required at all (Figure 2a).[1] The headline claims about predictive power, i.e. metacognition 42.12%, intelligence 1.12%, prior mathematics grades 3.72% (p. 1103), therefore translates as: *performance on 18 arithmetic expressions predicts performance on similar 18 arithmetic expressions better than an abstract-reasoning test or last year's grades do.*

The "general" side has the same disease in milder form. The Reading Monitoring Test (find nine planted contradictions in a text (p. 1098))is an error-detection task, classically confounded with reading comprehension and verbal ability: a reader who misses a contradiction may simply not have understood the passage [5, 6]. The authors themselves mention this critique of the Markman paradigm on p. 1099 before adopting the paradigm anyway, and the test's reliability is a modest $\alpha = .63$ (p. 1100). The General Metacognitive Ability latent, assembled from a verbal-comprehension-loaded reading task and an arithmetic-performance-loaded appraisal task, is operationally a verbal-plus-numerical scholastic composite. That such a composite predicts school grades in Portuguese, Mathematics, History and Geography is unremarkable, and uninformative about metacognition.

4 The intelligence estimate was handicapped

Against this scholastic composite, "intelligence" was estimated with a single reduced fluid-intelligence test with which I was not familiar and rightly so, given that it is indeed an obscure test: the only references I could find about it were in articles produced by authors of the presently critiqued metacognition study; most concerningly, the test itself was co-developed by one of the co-authors [14]. The reduced test consists of 27 items, Cronbach's $\alpha = .75$ (falling

[1]The .649 is a path coefficient estimated with fluid intelligence and prior mathematics grades also in the equation; the corresponding zero-order association is, if anything, larger, which only strengthens the point.

to .69 in the authors' own re-analysis), with eight items loading below .40 and one at .05 (p. 1100). Most notably, this test lacked any item assessing crystallized intelligence, or verbal factor, and no general factor was extracted from multiple broad abilities, despite a criterion (grades in language- and knowledge-heavy subjects) that loads heavily on crystallized ability. Having excluded verbal ability from the intelligence side of the model, the design then lets the reading test carry that same variance under the label "metacognition." The single-school sample (p. 1098; acknowledged pp. 1106–1107) compounds the handicap: school selection restricts the spread of intelligence, and restriction of a predictor's spread obviously shrinks its correlations.

The paper's own results should be symptomatic of the poor quality and bias of this test: Fluid intelligence explains 7.45% of general achievement (p. 1103), yet the introduction cites the established finding that intelligence explains 25–50% of achievement variance (p. 1096), and Deary and colleagues' five-year study of some 70,000 English schoolchildren, which was cited on that same page found a latent correlation of about .81 between intelligence at age 11 and national examination results at 16, i.e. 66% of variance at the latent level [4]. When a benchmark underperforms the literature by a factor of three to nine, the defect likely lies in the measurement.

This can be made quantitative. The reported 7.45% corresponds to a standardized coefficient of $\sqrt{.0745} \approx .273$. Start from the conventional population correlation $\rho = .50$ (the midpoint of the paper's own cited range) and apply the two artifacts in sequence.

1: range restriction. Thorndike's Case II formula [3] gives the correlation observable when the predictor's standard deviation is reduced by selection, and though I'm no statistician, this seems to me like the appropriate instrument. Writing $u = \sigma_{\text{restricted}}/\sigma_{\text{population}}$,

$$r_{\text{restr}} \;=\; \frac{u\,\rho}{\sqrt{u^2\rho^2 + 1 - \rho^2}} \;=\; \frac{0.60 \times 0.50}{\sqrt{0.36 \times 0.25 + 0.75}} \;=\; \frac{0.30}{\sqrt{0.84}} \;\approx\; 0.327, \tag{6}$$

taking $u = 0.60$ as plausible degree of selection for a private school (an assumption, this isn't an estimate from their data).

2: unreliability. Attenuation for predictor unreliability [2] multiplies the correlation by the square root of the measure's reliability:

$$r_{\text{obs}} \;=\; r_{\text{restr}} \times \sqrt{r_{xx}} \;=\; 0.327 \times \sqrt{.75} \;\approx\; 0.283 \qquad (\text{or } 0.327 \times \sqrt{.69} \approx 0.272). \tag{7}$$

The observed .273 is recovered almost exactly (Figure 3). The conclusion is not that $\rho = .50$ is so proven, since u was assumed, but rather that the paper's data are probably consistent with the standard intelligence–achievement correlation after proper adjustment and therefore contain little evidence against the primacy of intelligence. In other words, they show what $\rho \approx .50$ looks like after passing through a short, noisy, fluid-only, and dubious test in a range-restricted sample (It's important to stress again that $\alpha = .75$ is their best case scenario for reliability, as the paper itself reports $\alpha = .69$ for this reduced version of their test) in this case. The same logic is the covariate-side threat: a control variable measured with error cannot absorb its own variance, and the unabsorbed remainder strengthens whatever correlated predictor stays in the model [7], in this case, the scholastic composite which the authors label as metacognition.

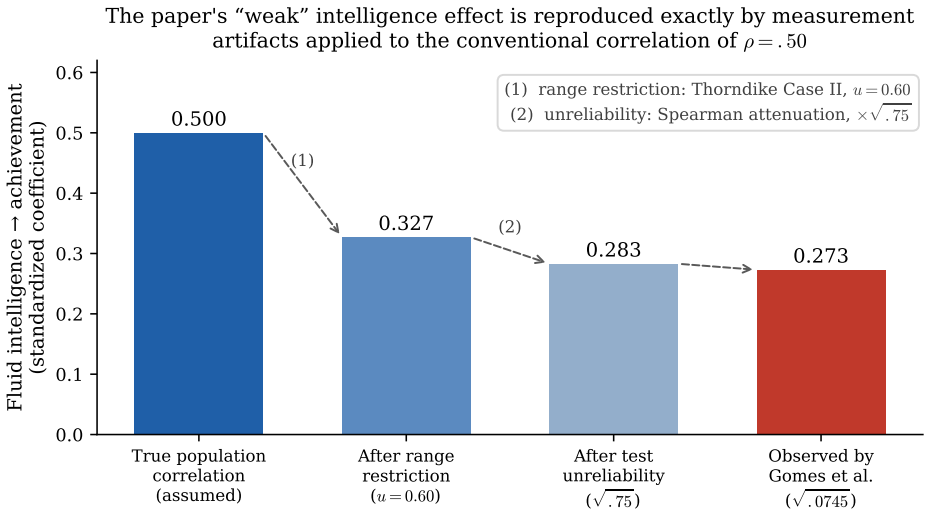

Figure 3: Reconstruction of the paper's fluid-intelligence coefficient from probable artifacts: the conventional $\rho = .50$ (which they themselves cite as the standard), passed through Thorndike's Case II range-restriction formula (6) with an assumed selection ratio $u = 0.60$ and the Spearman attenuation formula (7) with the reported reliability $\alpha = .75$, lands on the observed value to within 0.01.

5 Model-side choices that allocate the shared variance to metacognition

Four analytic decisions then tilt whatever variance remains.

Factor scores treated as error-free. The authors claim factor scores "take into account only the true score and eliminate both the error and specific variance" (p. 1101). This is incorrect as factor scores are fallible estimates subject to factor indeterminacy [8, 9]. This means that entering them as single observed variables, as opposed to latent variables, completely loses the disattenuation that a full latent-variable specification would have provided, i.e. the least reliable measure in the model, which is the intelligence, takes the hit. battery.

Modification-index respecification. The final model was reached by adding paths suggested by modification indices (a Mathematics–Portuguese covariance; direct paths from fluid intelligence to Mathematics and History), improving fit from $\chi^2 = 103.31$ ($df = 22$) to $\chi^2 = 51.18$ ($df = 19$), CFI = 1.00 (p. 1103). Post-hoc respecification of this kind capitalizes on sample-specific chance, and a near-saturated model with CFI = 1.00 cannot then be offered as confirmation.

A directed path from metacognition to intelligence. The model routes GMA → Gf, with metacognition "explaining" 12.85% of intelligence (p. 1103), a pure specification choice in cross-sectional data that credits the shared variance of the two constructs to metacognition by construction. The discussion states the mechanism with disarming obliviousness: when the metacognition trait enters, "it extracts considerably the intelligence's variance" (p. 1105). That sentence describes how collinear variance was allocated under an arbitrary model, not a genuine finding.

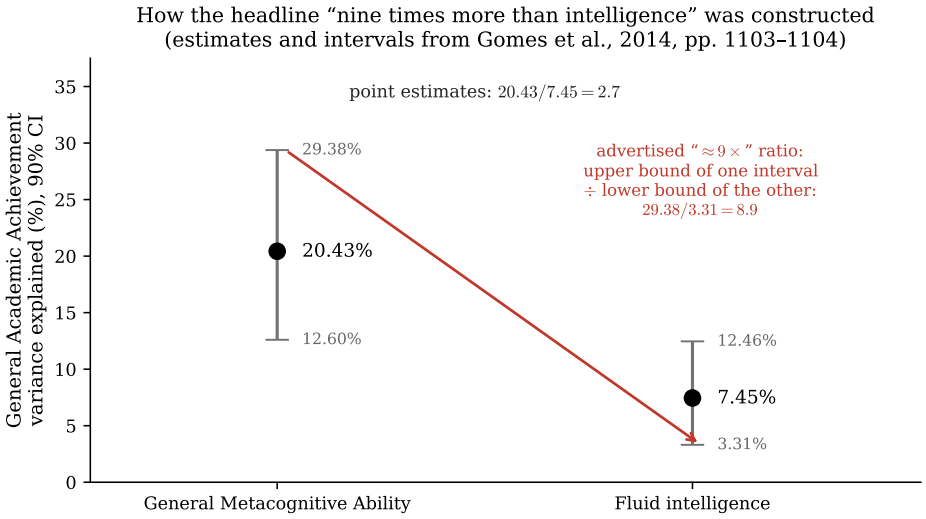

Figure 4: The construction of the headline ratio. Point estimates and 90% intervals from p. 1103 of [1]; the “≈ 9×” claim (p. 1104) divides the upper bound of one interval by the lower bound of the other.

Ratios computed between the bounds of different intervals. The advertised “approximately nine times” (p. 1104) is obtained by dividing the *upper* bound of metacognition’s 90% interval by the *lower* bound of intelligence’s: 29.38/3.31 = 8.9. The point estimates give 20.43/7.45 = 2.7 (Figure 4). The “15:1” and “18:1” figures for specific achievement are the same construction (36.00/2.34 = 15.4; 36.00/1.99 = 18.1). Comparing the top of one confidence interval with the bottom of another is not a legitimate effect-size comparison under any convention; it suggests sloppy bias by the authors, eager to prove a dubious narrative.

6 What would count as evidence, and the verdict

Set beside the think-aloud literature of the Veenman group [10, 11] and the accuracy-measure tradition, this paper is the cleanest published specimen of a recurring recipe for “metacognition beats intelligence” results: (i) the metacognition measure is taken on or immediately beside the criterion task while intelligence is measured remotely; (ii) the metacognition score is contaminated with raw performance, and here the contamination is written into the rubric itself (Section 2); (iii) the intelligence covariate is short, unreliable, fluid-only, and administered in a range-restricted sample (Section 4); (iv) the model is tuned after the fact and the directional choices favour metacognition (Section 5); and (v) personality, motivation and prior achievement are absent or accounted for (note that even prior mathematics grades were permitted only 3.72% here, because the model routed the shared variance through metacognition first).

The paper is also instructive for what it *fixed*. It repaired exactly the two weaknesses it identified in the earlier think-aloud studies, i.e., judge-based scoring and small samples (p. 1097), while leaving construct contamination and covariate quality untouched, and it thereby produced the most inflated metacognition-to-intelligence ratio in the literature. A probative design would need to satisfy three conditions simultaneously: a criterion measured independently of the metacognition assessment occasion and item pool; an intelligence covariate measured at full strength (multiple broad factors, adequate reliability, unrestricted or restriction-corrected sample); and a metacognition score mathematically decoupled from raw accuracy (e.g. signal-detection-theoretic sensitivity measures, with their own reliability problems honestly reported).

As far as I am aware, no study claiming metacognitive superiority over intelligence has yet met all three at once. Where all three are approximated, the unique predictive contribution of metacognition settles into the 1–3% of criterion variance in my private assessment implied by the meta-analytic evidence [12], i.e.,an order of magnitude below the ratios advertised here.

Note on sources and computation. All empirical quantities attributed to [1] are cited to page above. All figures were computed by me from the equations and estimates stated in the text; no data beyond the published estimates were used. The selection ratio $u = 0.60$ in Section 4 is explicitly an assumption; every other input is taken from the paper.

References

[1] Gomes, C. M. A., Golino, H. F., & Menezes, I. G. (2014). Predicting school achievement rather than intelligence: Does metacognition matter? *Psychology*, 5, 1095–1110.

[2] Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.

[3] Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

[4] Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35, 13–21.

[5] Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development*, 50, 643–655.

[6] McCormick, C. B. (2003). Metacognition and learning. In I. B. Weiner et al. (Eds.), *Handbook of Psychology: Educational Psychology* (pp. 79–102). Hoboken, NJ: Wiley.

[7] Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11(3), e0152719.

[8] Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6, 430–450.

[9] DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20), 1–11.

[10] Veenman, M. V. J., & Elshout, J. J. (1991). Intellectual ability and working method as predictors of novice learning. *Learning and Instruction*, 1, 303–317.

[11] Veenman, M. V. J., Elshout, J. J., & Meijer, J. (1997). The generality vs. domain-specificity of metacognitive skills in novice learning across domains. *Learning and Instruction*, 7, 187–209.

[12] Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, 13, 179–212.

[13] Colom, R., & Flores-Mendoza, C. (2007). Intelligence predicts scholastic achievement irrespective of SES factors: Evidence from Brazil. *Intelligence*, 35, 243–251.

[14] Gomes, C. M. A., & Borges, O. N. (2009). Qualidades Psicométricas do Conjunto de Testes de Inteligência Fluida. Avaliação Psicológica, 8, 17-32