THE JUDGMENT PREMIUM™ ASSESSMENT

Methodology & Research Foundations

A short, honest account of what the Judgment Premium Assessment measures, how it was built, what the evidence base is — and, just as importantly, what it does *not* yet claim. Written for the professional taking it, not the psychometrician reviewing it.

> **The one-paragraph version.** The JPA measures two things separately: how well you know your field, and how well you exercise judgment when AI can do the work. It uses realistic trade-off scenarios with no right answer, scoring each choice on both axes. It is grounded in thirty years of research on how people over-rely on automation. It is a *practitioner-derived* instrument on a defined validation roadmap — useful and evidence-based, but *not yet* a psychometrically validated test, and not for hiring or ranking. We'd rather tell you that plainly than overclaim.

Why it exists

Enterprises can measure two things well: whether you have technical skill, and whether you understand the technology. Neither measures the thing that now matters most — whether you can exercise professional judgment when AI produces the work product. Every major research body has flagged the same gap:

- Gartner predicts the atrophy of critical-thinking skills from GenAI use will push half of organisations to require "AI-free" skills assessments.
  > Gartner, "Strategic Predictions for 2026 and Beyond," Nov 2025.
- The World Economic Forum projects 39% of core skills will change by 2030.
  > WEF, Future of Jobs Report 2025 (1,000+ employers, 14m workers, 55 economies).
- McKinsey found high-performing organisations are nearly three times more likely to have defined processes for when AI outputs need human validation.
  > McKinsey, "Superagency in the Workplace," Jan 2025.
- The EU AI Act classifies AI in employment and operational decisions as high-risk, requiring documented human oversight — which presumes you can measure the quality of that oversight.
  > EU AI Act, 2025.

What it measures — two independent axes

The assessment scores you on two separate scales, each out of 5. They are deliberately kept independent: a high score on one says nothing about the other.

AXIS	WHAT IT CAPTURES
Domain Competency	Depth of professional expertise. For the BA edition, mapped to the IIBA BABOK® Guide and Competency Model. "Can you do the work?"
Judgment Premium	How well you exercise judgment over AI, across four disciplines (below). Universal — the same regardless of profession. "Can you tell when it's right, and own the call?"

The combination places you in one of four quadrants — **Interpreter, Augmenter, Orchestrator, Weaver** — with the **Augmenter** (high skill, low judgment) identified as the highest-risk position. That is not an opinion; it follows from the research below.

The four disciplines (the judgment axis)

DISCIPLINE	WHAT IT IS
Critical Examination	Questioning the premise of an AI output — what's wrong, missing, or quietly assumed.
Evidence Discipline	Demanding the source, baseline, and quantified evidence behind a claim.
Political Navigation	

	Reading the human reality AI can't see — who blocks, who decides, what history shapes the room.
Contextual Filtering	Filtering recommendations through what will actually work here, with these people, now.

How the questions work

The assessment uses an **ipsative** (forced-choice, self-referenced) design — a method formalised by Cattell in 1944. Every scenario involves AI producing a work product, and every option is defensible; there is no obviously correct answer. Each choice scores on both axes at once. Crucially, at least one option per scenario is an "*Augmenter trap*" — the impressive-sounding, high-skill, low-judgment choice. This is the heart of the design: someone who consistently picks the option that "sounds best" reveals a real pattern, not a lucky guess.

> Ipsative methodology: Cattell (1944); Edwards (1957); Saville & Willson (1991).

The science behind it

The instrument rests on more than thirty years of research into **automation bias** — the human tendency to over-trust automated advice.

- Automation bias was first identified in aviation decision-making.
  > Mosier & Skitka (1996); Parasuraman & Riley (1997).
- A 2025 systematic review of 35 studies found automation bias is *made worse by expertise, not reduced by it* — which is precisely why the Augmenter (high skill, low judgment) is the highest-risk quadrant.
  > AI & Society (Springer), 2025; Horowitz & Kahn (2024).
- A 2025 study of 529 participants confirmed expertise increases confidence but does *not* improve appropriate reliance on AI.
  > Taylor & Francis, Nov 2025.
- Research on "appropriate reliance" gives the framework for calibrating trust — knowing when to accept and when to question.
  > Schemmer, Kuehl, Benz, Bartos & Satzger (2023), ACM IUI '23.
- Deskilling research warns AI can lift novices to near-expert output while eroding the independent judgment that expertise is built on.
  > Frey & Weiss (2025), Harvard Data Science Review 7(2).

Together these establish the core claim: domain competency and AI-era judgment are **independent capabilities** — which is why the instrument measures them on separate axes.

How answers are kept honest

Because the assessment can be rushed or gamed, it runs several automatic validity checks. If any trigger, you still see your result, but a notice recommends a retake, and the record is tagged so it can be filtered from analysis.

- **Speed check** — flags runs faster than the scenarios can plausibly be read (they run 150–250 words; genuine consideration takes 45–90 seconds).
- **Position check** — since options are shuffled each time, always picking the same on-screen position signals not reading.
- **Consistency check** — four "anchor pairs" test the same discipline twice; large gaps flag inconsistency.
- **Structural** — the ipsative design itself resists gaming: the impressive-sounding answer is often the low-judgment one, so trying to look good produces a true signal.

What is validated — and what is not

> **This is the honest part, and it matters.** The scores attached to each option were assigned by *expert reasoning* — mapping options against the IIBA Competency Model (skill axis) and the Four Disciplines (judgment axis) — *not* yet by large-scale empirical validation with respondent data. This is normal and legitimate for an instrument at this stage of development. It also means the JPA should be read as a *structured professional diagnostic*, not a validated psychometric test. It is a mirror for development. It is not a basis for hiring, promotion, or ranking people.

The path from here to full validation is defined and public:

Phase 1 — Content validity (Q2 2026): a panel of senior (CBAP-level) practitioners independently scores each option; disagreements beyond ±1.0 are investigated and adjusted.

Phase 2 — Internal consistency (Q3 2026): with 50+ completed assessments, analyse answer distributions; any question where one option dominates (>60%) means the trade-off failed and is redesigned.

Phase 3 — Known-groups validation (Q4 2026): test whether the skill axis correctly separates certified proficiency levels (ECBA / CCBA / CBAP).

Phase 4 — Cross-profession portability (2027): confirm the four disciplines produce consistent judgment scores when the instrument is adapted to a second profession.

Phase 5 — Mode consistency (2027): confirm that someone taking both the Beginner and Advanced versions gets a comparable judgment score — the judgment measure should be independent of domain difficulty.

What makes it a first

Existing frameworks (BABOK, PMBOK, SFIA) ask "can you do the work?" AI-literacy frameworks (UNESCO, DigComp) ask "do you understand the technology?" Neither measures the capacity to exercise professional judgment when AI can produce the work product. The Judgment Premium introduces a two-axis model that measures competency and judgment independently, a quadrant identity model that names the Augmenter as the highest-risk position, and an ipsative design that measures orientation under real constraints rather than knowledge of best practice.

Selected references

Cattell, R. B. (1944). Psychological measurement: ipsative, normative, and interactive.
Mosier, K. & Skitka, L. (1996). Human decision makers and automated decision aids.
Parasuraman, R. & Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse.
Schemmer, M., Kuehl, N., Benz, C., Bartos, A. & Satzger, G. (2023). Appropriate Reliance on AI Advice. ACM IUI '23.
Horowitz, M. & Kahn, L. (2024). Bending the automation bias curve.
Frey, C. & Weiss (2025). Deskilling and AI. Harvard Data Science Review 7(2).
Automation bias systematic review (2025). AI & Society (Springer).
Kahneman, D. (2011). Thinking, Fast and Slow. · Tetlock, P. & Gardner, D. (2015). Superforecasting.

Industry: Gartner (2026); McKinsey (2025); WEF (2025); IDC (2025); EU AI Act (2025).
Domain: IIBA BABOK® Guide v3; IIBA Competency Model® v4.

Judgment Premium™, the 4C Mindset™, the Four Disciplines, and the supporting COMPLETE / FLOW / FEATS frameworks © 2026 Susan K.L. Yu · Altschool Australia. This is a plain-language summary; a fuller technical methodology is maintained separately. The JPA is a practitioner-derived instrument on the validation roadmap above and is not intended for selection, hiring, or ranking.