BEYOND SELF-REPORT: EVALUATING AUTHENTIC AI LITERACY IN A CONTINUING EDUCATION CONTEXT

MARIKA MASCITTI [1]
University of Bologna
marika.mascitti@unibo.it

Paola Rigoni
University of Bologna
paola.rigoni2@unibo.it

Chiara Panciroli
University of Bologna
chiara.panciroli@unibo.it

Double Blind Peer Review

Citation
MASCITTI, M., Rigoni, P., & Panciroli, C. (2026). Beyond Self-Report: Evaluating Authentic AI Literacy in a Continuing Education Context. *Italian Journal of Health Education, Sport and Inclusive Didactics*, *10*(4). https://doi.org/10.32043/gsd.v10i4.1872

Doi: https://doi.org/10.32043/gsd.v10i4.1872

Copyright notice:
© 2026 The Author(s). This is an open access, peer-reviewed article published by Cares S.r.l. – REN Press and distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

renpress.it/index.php/gsdj
ISSN: 3103-5744

ABSTRACT
Research on AI literacy has predominantly relied on self-report instruments (Lintner, 2024), which correlate weakly with objective performance-based measures (Zhang et al., 2026). This study examines how AI literacy can be reliably assessed among adults engaged in continuing education. Participants (N = 141) completed the AILQ (Ng et al., 2024), then performed an authentic task based on a real-world employer branding scenario. Performance was assessed using a four-dimension analytic rubric. A Disposition-Performance Gap (DPG) index quantified the misalignment between self-perceived disposition and demonstrated performance: 68.8% of participants were aligned, while 27.7% showed a Disposition Surplus. This pattern was strongly stratified by competence: all participants at the lowest performance level overestimated their competence, whereas proficient participants were almost entirely aligned ($\chi^2(4) = 78.413$, p < .001). Findings support performance-based assessment as a necessary complement to self-report instruments.

KEYWORDS
AI literacy; Performance-based assessment; Continuing Education; DPG.

Received 2026-07-09
Accepted 2026-09-26
Published 2026-10-30

[1] This paper is the result of the joint work of the authors. Specifically, Author 1 wrote §§2-2.3, §§3-3.4; Author 2 wrote §1.1, §1.2 and §4, and Author 3 wrote §1. The section §5 and §6 was jointly developed by all the authors.

1. Introduction

The diffusion of large language models such as ChatGPT and Gemini have moved from research curiosities into everyday instruments of knowledge work for professional and in the organizational contexts. Nationally representative survey evidence collected in the United States found that roughly four in ten adults had already used generative AI within little more than a year and a half of ChatGPT's public release, with more than a quarter of employed respondents reporting use of these tools specifically for their jobs (Bick et al., 2024). Marketing and HR departments draft campaign copy and screen candidate materials with AI support, consultants prototype deliverables in minutes, managers delegate first-draft cognitive labor to systems whose internal logic remains largely opaque. This is no longer a marginal phenomenon confined to technologically advanced firms, it is a pervasive feature of contemporary business practice. This diffusion has outstripped the organizational and individual capacity to develop a correspondingly mature understanding of what it means to use these tools well. Adoption has been driven less by structured training than by accessibility, and generative AI tools produce outputs that are immediately usable, fluent, and superficially convincing. This combination of low entry barriers and high apparent output quality creates a particular risk, one that the human factors literature had already documented well before generative AI under the label of automation bias, the tendency of operators to over-rely on automated systems and fail to detect the errors they introduce, even when possessing relevant domain expertise (Goddard et al., 2012). When a tool produces text or images that read as coherent and complete, the natural inference is that the underlying content is also accurate. Yet fluency is, at best, a weak proxy for substantive quality. Field-experimental evidence makes this point with particular force (Dell'Acqua et al., 2023) showing that AI assistance substantially improved performance on tasks within what they term the «jagged

technological frontier» of current AI capability, yet just as reliably degraded performance on tasks outside that frontier, because consultants could not reliably distinguish, from fluency alone, which side of the frontier a task fell on. Critically, recent experimental work testing whether self-reported AI literacy protects against this kind of uncritical deference found that it does not: participants' self-reported AI literacy showed no significant protective effect against automation bias when exposed to deliberately faulty AI advice, leading the authors to conclude that «familiarity alone is not sufficient to ensure healthy skepticism» (Wingerter et al., 2025). This finding anticipates, in miniature, the central empirical puzzle motivating the present study: even where individuals believe themselves equipped to evaluate AI output, that belief offers no guarantee of actual evaluative competence. Resolving this puzzle requires first clarifying what AI literacy itself is taken to mean, since the term AI literacy has been used with considerable definitional variability, reflecting both the youth of the field and the speed at which its object of study has changed. One influential conceptualization grounds itself in an analogy to digital literacy, identifying four constructs: awareness, usage, evaluation, and ethics, validated through a twelve-item AI Literacy Scale (AILS) across two independent samples using a rigorous multi-step process (Wang et al., 2022). The resulting four-factor structure clearly outperformed one, two, and three-factor alternatives in model fit, and the authors explicitly distinguish AI literacy from general digital literacy on substantive grounds: people often interact with AI through social rather than instrumental logic, meaning a person can be highly digitally literate while remaining poorly equipped to evaluate generative AI outputs. A more recent framework is the Affective, Behavioural, Cognitive, and Ethical (ABCE) model (Ng et al., 2024), which explicitly incorporates non-cognitive dimensions (confidence, motivation, engagement, and collaboration) alongside the conceptual and ethical competencies central to earlier formulations. The associated AI Literacy Questionnaire (AILQ) was validated through expert judgment, qualitative interviews, a pilot study, and first- and second-order confirmatory factor analysis,

demonstrating good model fit and high internal consistency (Cronbach's alpha exceeding 0.90). The ABCE model was validated on a sample of 363 secondary school students in Hong Kong, aged 12 to 17, within a structured educational program (Ng et al., 2024). Like the AILQ, it relies entirely on self-report measures: all four dimensions, including the cognitive dimension, are assessed by asking participants to rate their own perceived competence rather than by requiring them to demonstrate that competence through performance tasks. As a result, the model provides a well-validated framework for understanding how people perceive their affective, behavioral, cognitive, and ethical engagement with AI, but not necessarily how accurately these perceptions reflect their actual abilities.

1.1 Self-report bias

The ABCE/AILQ and the AILS are not exceptions. In validating the Meta AI Literacy Scale (MAILS), Koch et al. (2024) note that most AI literacy instruments developed in recent years rely primarily on self-report measures. While these scales have generally received support for their factorial structure, their validity has rarely been examined beyond this initial level of psychometric evaluation. Although the validation of the MAILS was methodologically rigorous (the authors tested a nomological network of theoretically predicted relationships across multiple studies), most of the evidence relied on correlations between MAILS scores and other self-report instruments designed to measure similar constructs. This approach has an important limitation: when all instruments are based on self-report, strong correlations do not necessarily indicate that they capture actual AI competence. Instead, they may simply reflect a common tendency for people to evaluate themselves in similar ways, that is, a shared response bias.

This limitation becomes evident in the only study in which the MAILS was compared with an independently developed, peer-reviewed objective knowledge test (Koch et al., 2024). Only the overall MAILS score, the Know & Understand dimension, and Emotion Regulation showed significant correlations with objective performance,

and these relationships were relatively weak. By contrast, several dimensions that are arguably most relevant for competent AI use (Use AI, Ethics, Detect AI, and Problem Solving) showed no significant relationship with objectively measured knowledge.

The same issue emerges even more clearly in the work of Zhang et al. (2026). Rather than comparing one self-report scale with another, the authors developed two parallel instruments based on the same theoretical framework: a Self-Report (SR) measure and an Objective-Based (OB) performance test, which they administered to 288 Taiwanese K–12 teachers. Because both instruments assess the same underlying construct using different measurement approaches, this design avoids the circularity inherent in comparing two self-report questionnaires. Both instruments demonstrated strong psychometric properties. The objective test was validated using Rasch modeling (KR-20=0.862), while the self-report questionnaire showed excellent fit in confirmatory factor analysis (α=0.889). Nevertheless, correlations between objectively measured competence and self-reported competence were consistently weak (r=0.07–0.24). The authors therefore conclude that self-reported and objectively measured AI literacy cannot be used interchangeably. Latent profile analysis further illustrates why this finding matters at the individual level. The population was not composed simply of people who tended to overestimate their abilities. Instead, six distinct profiles emerged: some participants genuinely overestimated their competence, others underestimated it (the largest group), others showed close agreement between perceived and demonstrated competence, and two additional groups clustered around the average with only minor discrepancies. The key implication is that the error in self-report measures is not systematic or uniformly directional. There is no single overestimation bias that can simply be corrected statistically. Rather, different individuals misjudge their competence in different ways: some overestimate it, some underestimate it, and others assess it accurately. Consequently, a self-report score alone provides limited information about a person's actual AI competence

and does not reliably indicate which of these profiles an individual belongs to.

Previous experience with AI tools appears to reduce the most extreme forms of mismatch between perceived and actual competence, but it does not eliminate the problem of overestimation. Stronger evidence comes from two large-scale experiments in which participants used ChatGPT to solve LSAT-style logical reasoning problems (Fernandes et al., 2026). Across the two studies, involving nearly 700 participants, individuals completed the same reasoning tasks both without assistance and with ChatGPT support. As expected, AI assistance significantly improved objective performance. However, it produced an even larger increase in participants' estimates of their own performance. In other words, although participants performed better with AI assistance, they overestimated their performance to a much greater extent than those who completed the same tasks without AI. Importantly, this pattern persisted even when participants received financial incentives for estimating their performance as accurately as possible. The persistence of overestimation despite monetary rewards suggests that the effect reflects a genuine metacognitive bias rather than a simple lack of motivation or care.

Perhaps the most striking finding concerns the distribution of this overestimation. Under normal conditions, the well-established Dunning-Kruger effect predicts that lower-performing individuals tend to overestimate their abilities more than higher-performing individuals. Under ChatGPT assistance, however, this pattern disappeared. AI support reduced performance differences across individuals with different ability levels, but it did not improve their ability to evaluate their own performance accurately. As a result, both higher and lower-performing participants displayed similarly elevated levels of overconfidence. The same study also reports a finding with direct implications for AI literacy measurement. Using the self-report SNAIL instrument, they found that participants reporting higher levels of Technical Understanding were, paradoxically, those who showed the greatest overestimation of their own performance. Similarly, the Critical Appraisal and Practical Application

dimensions were associated with greater confidence in one's answers, but not with a better ability to distinguish correct from incorrect responses.

The authors interpret these findings through the concept of the illusion of explanatory depth. Familiarity with AI systems appears to create the impression that one understands how they work far better than is the case. Therefore, experience with AI increases confidence in one's own abilities without producing a corresponding improvement in actual evaluative competence

1.2 Research gap and research questions

To date, no AI literacy instrument has been developed and validated specifically for adult professionals working in organizational settings, even though these are the individuals who routinely use AI tools to support business decisions and organizational processes. Existing research has focused primarily on students, teachers, or online samples drawn from the general population, while most available instruments continue to rely exclusively on self-report questionnaires. As discussed in the previous sections, this approach has an important limitation: perceived competence does not necessarily correspond to demonstrated competence.

The present pilot study was designed to address both gaps within a population of adult business professionals.

Therefore, two questions guided the research: how do business professionals assess their own ability to use, interpret and engage with AI, and does it differ from an externally evaluated performance? How can an authentic task be designed to operationalize the ABCE dimensions of AI literacy as observable performance indicators in adult professionals?

To answer the first question, in this study we used the validated ABCE framework (Ng et al., 2024) within a population of adult business professionals. The ABCE framework was selected as the theoretical foundation because it conceptualizes AI

literacy as a multidimensional construct encompassing Affect, Behaviour, Cognition, and Ethics, providing a more comprehensive account of competent AI use than frameworks that focus primarily on cognitive skills or consumer-oriented technology use. Furthermore, the framework is particularly well suited to the target population of this study, students enrolled in a Master's programme in Advanced Digital Technologies, as it captures the broad range of competencies required by future professionals who are expected not only to use AI effectively, but also to critically evaluate its implications and apply it responsibly in professional contexts.

To answer the second question, the present study assesses AI literacy through an authentic task, designed according to the principles of authentic assessment (Gulikers et al., 2004). A performance assessment rubric, with top-down codebook, was used to operationalize the dimensions of the ABCE framework. Participants were asked to evaluate AI-generated communication materials in a scenario representative of activities commonly performed by professionals in marketing, human resources, and corporate communication, moving beyond the abstract reasoning tasks that have characterized previous research. This methodology combines self-report and performance-based evidence to overcome the limitations of AI literacy research based solely on self-assessment. The study was conducted with professionals enrolled in the Digital for Business Joint Master programme[2] in Advanced Digital Technologies for Business (EQF 6), a European continuing education initiative accredited by ASIIN, and jointly delivered by the University of Bologna, NOVA School of Business and Economics, Linköping University, and the National College of Ireland. This population was deliberately selected because it consists of working professionals who have chosen to deepen their expertise in digital transformation and artificial intelligence while occupying roles in organizational settings. As such, it provides an appropriate context for developing and validating AI literacy measures intended for professionals.

[2] The project ID is DIGITAL-2021-SKILLS-01-01-101084013

2. Method - Marika

Data was collected first via the AI Literacy Questionnaire (AILQ) via an online survey, providing self-reported measures across the ABCE dimensions; a week later, the same participants completed an authentic task designed to elicit observable AI literacy competence in terms of critical analysis skills of AI-generated content. Open-ended responses collected were analysed through a deductive coding approach in which categories are defined a priori: in this case, by the dimensions and level descriptors of an analytic rubric. Rubric scores were then converted into numerical values through quantizing (Sandelowski et al., 2009) and entered the statistical analyses. The quantitative gap between self-report and performance was then quantified through a Disposition-Performance Gap (DPG).

2.1 Participants

The present study was conducted in February 2026. 141 students of the Master's in Advanced Digital Technology for Business, with a mean age of 44.2 years (male 74,5%), voluntarily completed both assessments. Descriptives of the participants are illustrated in Annex 1.

2.2 Instruments

2.2.1 AI Literacy Questionnaire (AILQ)

The AILQ (Ng et al., 2024) operationalises the ABCE framework (Affective, AL; Behavioural, BL; Cognitive, CL, and Ethical learning, EL) through 32 items on a five-point Likert scale (1=strongly disagree, 5=strongly agree). According to the author, the *AL* comprises four factors capturing learners' subjectively experienced attitudes toward AI: intrinsic motivation, self-efficacy, career interest, and confidence in learning AI; *BL* comprises behavioural commitment to pursuing AI-related learning

goals, and collaboration aimed at building relationships that support such learning; *CL* comprises three factors organized along a progression of thinking skills, from lower-order (knowing and understanding AI) through mid-order (using and applying AI) to higher-order skills (evaluating and creating with AI); and *EL* includes reliability, safety, privacy, responsibility, transparency, awareness, and social good, intended to foster a positive disposition toward AI.

A Socio-Technical Awareness (STA) dimension was added to the original ABCE structure, defined as the ability to recognize that AI outputs are not neutral artifacts but are shaped by training data, model architecture, and system default settings. The *STA* dimension comprises a bifactorial structure: sociotechnical responsibility and critical non-neutrality awareness. The score for each AILQ dimension is computed as the arithmetic mean of its constituent items, preserving the 1-5 interpretive range and ensuring comparability across dimensions of varying length (Tab. 1).

Table. 1. Descriptive statistics for the survey dimensions

Dimension	***n items***	***M***	***SD***	***Skewness***	***α***
AL	10	4.19	0.51	-0.41	.872
BL	8	3.94	0.60	-0.40	.858
CL	6	3.77	0.70	-0.46	.841
EL	8	4.49	0.41	-0.56	.732
STA_F1	5	4.12	0.59	-0.45	.750
STA_F2	4	3.52	0.95	-0.32	.830

2.2.2 Authentic task

The authentic task was embedded as part of the educational programme. The scenario presents participants with a realistic professional situation: as

communication or HR professionals, they are asked to evaluate two AI-generated employer branding campaigns produced by different large language model systems (namely, ChatGPT-GPT 5.2 and Gemini 2.5 Flash-Imagen3), both generated from an identical, generic prompt: "Generate a text and an image for an employer branding campaign". The use of a single, underspecified prompt across both systems is a deliberate design choice: it replicates a common real-world condition in which generative AI tools are used without domain-specific guidance, thereby maximising the likelihood of systematic biases in the output. Fig. 1 shows the outputs participants had to analyse.

The activity structure includes 5-points Likert-scale items assessing perceived effectiveness of the prompt, text, and image, and likelihood of professional use, as well as open-ended reflective questions in which participants are asked to articulate their reasoning across all four components of the task.

Fig. 1 Image and Text outputs generated from the prompt: "Generate a text and an image for an employer branding campaign"



[ChatGPT Output]: We are a dynamic and fast-growing company made up of young, motivated, and results-driven professionals. We believe in merit, individual responsibility, and the ability to excel. We reward those who demonstrate natural leadership, autonomy, and the ability to make rapid decisions. We do not value bureaucracy or people who need rigid rules: we prefer individuals who are independent, self-confident, and action-oriented. Our team operates in a stimulating, competitive, and high-performance environment where we reward talent and the willingness to step up and challenge oneself. We are looking for flexible and resilient individuals who are ready to work under pressure to achieve ambitious goals. We offer rapid growth opportunities for those who demonstrate commitment, dedication, and the ability to adapt to a constantly evolving environment.

[Gemini Output]: Headline: We Don't Hire Employees. We Build Leaders. Who We Are: We are a dynamic, fast-growing powerhouse fueled by one thing: excellence. We don't just meet targets; we redefine them. Our team is a collective of young, hungry professionals who aren't afraid to take risks, make decisions, and own their results.



Our Culture: Merit over Bureaucracy. At [Company Name], we have no room for red tape or rigid hand-holding. We believe in individual responsibility and radical autonomy. Here, talent isn't just noticed—it's rewarded. If you are a self-confident high-performer who prefers action over over-analysis, you've found your tribe.

What We Look For: Resilience & Grit: You thrive under pressure and see challenges as fuel for growth. Result-Orientation: You are obsessed with the "win" and don't stop until the goal is conquered. Decisive Leadership: You don't wait for permission; you take initiative and make rapid, data-driven decisions. Unapologetic Ambition: You want a career trajectory that moves as fast as you do.

Why Join Us? Forget the slow, predictable corporate ladder. We offer a high-octane environment where professional evolution is accelerated by merit, not seniority. If you are ready to work hard, play for the win, and see your impact recognized in real-time, your journey starts here.

[Company Name]: Drive. Analyze. Conquer.

2.2.3 Performance Assessment Rubric

Performance on the authentic task is assessed through a four-dimension analytic rubric, scored on an ordinal scale from 0 to 3 per dimension, yielding a total score range of 0-12. The four dimensions and their behavioural descriptors are:

a) *Prompt Evaluation,* assessing the capacity to identify structural limitations and missing elements in the prompt, and to connect them to the characteristics of the generated output;

b) *Text Analysis,* assessing the capacity to identify and critically argue the presence of specific bias categories in the AI-generated text, with reference to their professional implications;

c) *Image Analysis,* assessing the capacity to identify representational choices, visual bias patterns, and AI-generated artefacts in the accompanying image;

d) *Socio-Technical Awareness,* assessing the capacity to articulate how cultural values, ideological assumptions, and design choices are encoded in AI outputs, and to recognise that the training data underlying large language models reflects structurally biased representational patterns, reflecting structurally biased representational patterns rooted in dominant Western-centric, male-centred, ageist, and ableist cultural assumptions.

Score 0 indicates complete absence of the dimension's elements in the response; scores 1-3 correspond to progressively more articulated and evidence-based performance, anchored to empirically documented bias categories. The descriptor anchors were developed from a body of literature on bias in large language model outputs, including: ageism (Kamruzzaman et al., 2023; Martens et al., 2025); ableism and neurodiversity bias (Glazko et al., 2024; Phutane et al., 2025); gender bias in text-to-image systems (Girrbach et al., 2025; Naik & Nushi, 2023); racial representation (Bender et al., 2021); toxic rhetoric and cultural bias (Gehman et al., 2020; Muñoz-Ortiz et al., 2024; Tao et al., 2024); absence of inclusive values and diversity washing (Beckert & Koch, 2025); legal risk under EU anti-discrimination law (Council of the European Union, 2000).

The 0-12 range defined three interpretive thresholds:

a) Naive (0-4);

b) Developing (5-8);

c) Proficient (9-12).

The final score was calculated as the average of the ratings assigned through a double blind peer review of the open-ended questions included in this study, conducted by two researchers. Cohen's κ values, computed on a 25-response subsample (17.7% of the total), ranged from .471 (socio-technical awareness) to

.593 (text analysis), indicating moderate agreement across all four dimensions (Landis & Koch, 1977). All values reached statistical significance ($p < .001$).

2.2.4 Disposition-Performance Gap

To examine the relationship between self-perceived and demonstrated AI literacy, a Disposition-Performance Gap (DPG) index was computed. The rationale for relating the two instruments rests on their complementary status as proxies of the same underlying object of interest: the AILQ provides an internal proxy of AI literacy competence, filtered through the participant's self-representation, whereas the rubric provides an external, criterion-referenced evaluation of that competence as demonstrated on a concrete task. Comparing the two therefore addresses the extent to which self-representation converges with externally assessed performance, that neither instrument can answer alone. At the same time, the two proxies operationalize conceptually distinct constructs: as stated by Ng et al. (2024), the AILQ captures a general self-perceived disposition towards AI (motivation, self-efficacy, behavioural commitment, ethical values), whereas the rubric captures a task-specific critical analysis competence. For this reason, the DPG is not framed as a measure of metacognitive calibration (Eva & Regehr, 2005), but as the misalignment between a general disposition and a specific performance. The perceived disposition score was computed as the mean of the four AILQ dimensions (AL, BL, CL, EL) and linearly rescaled to the rubric metric, mapping the 1-5 range onto 0-1 scale:

$$\textit{Perceived competence score (normalized)} = (\textit{Mean_ABCE} - 1) \times 3$$

$$\textit{DPG_ABCE} = \textit{Perceived competence score (normalized)} - \textit{Rubric_mean}$$

where Mean_ABCE is the unweighted mean of the four ABCE dimension scores, and Rubric_mean is the mean rubric score across both stimuli (S1 and S2). An analogous formula applies for DPG_ABCE_STA, using the mean of all six dimensions (AL, BL,

CL, EL, STA_F1, STA_F2). DPG_ABCE is treated as the primary measure, as it relies exclusively on the originally validated AILQ dimensions; DPG_ABCE_STA serves as a sensitivity check.

Three categories were defined using a priori cut-offs of ±3, corresponding to one rubric level: *Disposition Surplus* (DPG > +3), in which self-perceived general disposition exceeds demonstrated task-specific performance; *Alignment* (−3 ≤ DPG ≤ +3); and *Performance Surplus* (DPG < −3), in which demonstrated performance exceeds self-perceived disposition.

2.3 Data analysis

Descriptive statistics were computed for all AILQ dimensions and for rubric scores by dimension and stimulus. Paired-samples t-tests examined differences between S1 and S2 rubric scores across dimensions. Performance-level distributions (Disposition Surplus/Alignment/Performance Surplus) were computed from the mean rubric score across stimuli. DPG distributions were examined both continuously (mean, SD, range) and categorically (frequency per level). The association between performance level and DPG category was tested using Pearson's chi-square. Pearson correlations were computed between AILQ dimension scores and rubric total score.

3. Results

3.1 Self-Reported AI Literacy (AILQ)

As shown in Tab. 1, all dimension means fall above the scale midpoint (3.0), with EL recording the highest value (M=4.49, SD=0.41) and STA-F2 the lowest (M=3.52, SD=0.95). EL also displays the most restricted range (3.50-5.00), indicative of a partial ceiling effect: no participant rated ethical attitudes below 3.50, suggesting near-universal endorsement of general AI ethics principles. STA-F2 exhibits the highest variability (SD=0.95), reflecting heterogeneous positions on the non-neutrality of algorithms and digital platforms. All six distributions are negatively

skewed, with kurtosis values approximating zero, consistent with approximate normality.

3.2 Performance Assessment (Rubric)

Descriptive statistics for rubric scores by dimension and stimulus, along with paired t-test results for the S1-S2 comparison, are presented in Tab. 2. The mean total score across both stimuli is M=7.93 (SD = 2.89), positioning the sample at the upper boundary of the Developing level and approaching the Proficient threshold (≥9). The full range (0-12) is observed in each dimension and for both stimuli, indicating that the rubric effectively discriminates across all performance levels without floor or ceiling effects.

Table. 2. Rubric descriptive statistics and S1 vs S2 paired t-tests

Dimension	***S1 M***	***SD***	***S2 M***	***SD***	***N***	***t (df=140)***	***p***
Prompt evaluation	2.02	0.96	2.02	0.96	141	—	n.a.
Text analysis	1.98	0.85	1.99	0.85	141	-0.35	.728
Image analysis	2.03	0.82	2.06	0.77	141	-0.47	.636
Socio-technical awareness	1.91	0.99	1.87	0.96	141	0.76	.450
Total score (0-12)	7.93	3.07	7.93	2.84	141		

Paired-samples t-tests reveal no statistically significant differences between S1 and S2 on any dimension, confirming the two stimuli as equivalent assessment instruments despite their different rhetorical and visual characteristics. Performance-level distributions are presented in Tab. 3. Approximately half the sample (49.6%, n = 70) achieves the Proficient level on the mean score, with 36.9% (n = 52) at the Developing level and 13.5% (n = 19) at the Naive level. Distributions are highly consistent across S1 and S2.

Table. 3. Distribution of participants by performance level (Naive / Developing / Proficient)

Level	*S1 n*	*S1 %*	*S2 n*	*S2 %*	*Mean n (%)*
Naive (0-4)	23	16.3	18	12.8	19 (13.5)
Developing (5-8)	51	36.2	55	39.0	52 (36.9)
Proficient (9-12)	67	47.5	68	48.2	70 (49.6)
Total	141	100	141	100	141 (100)

3.3 Disposition-Performance Gap (DPG)

Descriptive statistics and category distributions for both DPG versions are presented in Tab. 4. Both versions yield positive mean values (DPG_ABCE: M = +1.36, SD = 2.96; DPG_ABCE_STA: M = +1.08, SD = 2.93), indicating a slight tendency toward *Disposition Surplus* at the sample level. In both cases, however, the mean falls within the *Alignment* zone (±3). No participant reaches the theoretical extremes of the scale (±12), and the observed range is substantially narrower than the theoretical range (approximately −5 to +8 for both versions).

Table 4: DPG descriptive statistics and category distributions

DPG version	*M*	*SD*	*Min*	*Max*	*Perf. Surplus / Align. / Disp. Surplus*
DPG_ABCE	+1.36	2.96	−5.11	+8.02	3.5% / 68.8% / 27.7%
DPG_ABCE_STA	+1.08	2.93	−5.08	+8.38	5.0% / 68.1% / 27.0%

The two DPG versions produce nearly identical category distributions (Performance Surplus: 3.5% vs 5.0%; Alignment: 68.8% vs 68.1%; Disposition Surplus: 27.7% vs 27.0%).

The cross-tabulation of performance level by DPG_ABCE category (Tab. 5) reveals a highly structured pattern.

Table 5: Cross-tabulation of performance level by DPG_ABCE category

Performance level	*Perf. Surplus*	*Alignm.*	*Disp. Surplus*	*Total*
Naive (0-4)	0 (0.0%)	0 (0.0%)	19 (100%)	19
Developing (5-8)	0 (0.0%)	33 (63.5%)	19 (36.5%)	52
Proficient (9-12)	5 (7.1%)	64 (91.4%)	1 (1.4%)	70
Total	5	97	39	141

All 19 participants at the Naive performance level fall in the Disposition Surplus category (100%). Among Developing-level participants, 63.5% are Aligned and 36.5% are in the Disposition Surplus category; no Developing-level participant is in the Performance Surplus category. Among Proficient-level participants, 91.4% are Aligned, 7.1% (n = 5) are in the Performance Surplus zone, and only 1.4% (n = 1) is in the Disposition Surplus zone. Pearson's chi-square test confirms a statistically significant association between performance level and DPG category ($\chi^2(4) = 78.41$, $p < .001$).

4. Discussion

The central idea of the Artificial Intelligence Literacy Questionnaire (AILQ) is that being AI literate involves more than simply understanding AI from a technical perspective; it also requires developing appropriate attitudes, behaviours, and ethical awareness in the use of these technologies.

The AILQ questionnaire reveals that the sample is highly motivated and ethically oriented toward AI, yet less collaborative in its learning practices and less confident in advanced cognitive operations such as creation and system design. In the rubric the socio-technical awareness dimension consistently yields the lowest mean scores in both S1 (M = 1.91, SD = 0.99) and S2 (M = 1.87, SD = 0.96), indicating that the capacity to articulate the socio-technical embeddedness of AI outputs is the least consolidated competence in this sample.

Regarding the DPG, at the aggregate level the sample appears well calibrated: the mean score falls within the a priori tolerance zone, and approximately two-thirds of participants fall in the Alignment zone. However, this aggregate pattern conceals a threshold-dependent structure that becomes apparent only when results are disaggregated by performance level (Tab. 5). Every participant classified as Naive on the authentic task falls in the Disposition surplus zone. Among Developing-level participants, overestimate their disposition relative to their external assessment, and none falls in the Performance surplus zone. Among Proficient-level participants, the pattern reverses: Disposition surplus is nearly absent (1.4%), while a small subgroup (7.1%) underestimates their disposition relative to their external assessment. This indicates that discrepancy between the dispositions and the performance is not a uniform feature of the sample: those with a Naive performance are also those who most consistently overestimate their AI literacy, while Proficient participants assess their own AI literacy comparatively accurately.
This finding extends, in a professional setting, the evidence reported by Fernandes et al. (2026) and Zhang et al. (2026), according to which self-assessment errors are not uniform across individuals and the risk of miscalibration is not randomly distributed across the population. Practically, this implies that self-reported confidence alone cannot identify which individuals are most likely to over-rely on AI output: demonstrated performance must be measured directly.

5. Conclusion

The relevance of this study in stressing the idea that business professionals should develop awareness not only of their technical skills, but also of their ability to analyse the ethical, political, and socio-technical implications that characterize the use of AI. AI is not a neutral tool; its outputs reflect design choices, training data, and implicit cultural values that go beyond the purely technical dimension. In this sense, the findings of this study support the need for educational pathways aimed at fostering metacognition and critical reflection on the use of AI,

identifying/highlighting a gap between the perception of one's own dispositions and the actual skill to capture its complexity. The DPG therefore proves to be an agile tool for monitoring these abilities, moving beyond reliance on self-reported questionnaires by complementing them with authentic tasks that provide a more accurate frame of the training needs of business professionals, in relation to the use of an advanced and generative technology such as AI.

6. Limitations and future directions

This study has several limitations. First, the sample consisted of professionals enrolled in a single Digital for Business Master's programme, limiting the generalisability of the findings to other professional contexts. Second, the cross-sectional design does not allow changes in AI literacy over time or causal relationships to be examined. Third, although the authentic task provides a more performance-based assessment than self-report measures alone, it was based on a specific scenario and scoring rubric, which may not capture the full range of AI literacy skills across different contexts or AI systems. Nevertheless, the architecture of the task separates components that can be held constant across contexts from components that must be re-specified. Three elements are context-dependent: the professional artefact requested in the prompt (e.g., a job advertisement or candidate summary in human resources, where disability bias in GPT-based screening and hiring scenarios has been documented; see Glazko et al., 2024; Phutane et al., 2025), the set of bias categories, and the applicable normative framework. Each adaptation would require renewed expert review of the bias categories and rubric descriptors, as well as new inter-rater reliability estimates, since agreement levels for analytic rubrics cannot be assumed to transfer to new stimuli. Where outputs are text-only, the Image Analysis dimension cannot be scored, and total scores and DPG values would not be comparable with those reported here. Future research should replicate the study with larger and more diverse samples, adopt longitudinal designs, and validate the DPG using a broader range of authentic tasks and generative AI applications.

References

Beckert, J., & Koch, T. (2025). Corporate perspectives on diversity: Engagement, communication motives, and the Diversity-Washing Dilemma. *Journal of Marketing Communications*, *31*(5), 538–555. https://doi.org/10.1080/13527266.2025.2470346

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bick, A., Blandin, A., & Deming, D. J. (2024). *The rapid adoption of generative AI* (NBER Working Paper No. 32966). National Bureau of Economic Research. https://doi.org/10.3386/w32966

Council of the European Union. (2000). *Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation*. Official Journal of the European Communities, L 303, 16–22. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality* (Harvard Business School Working Paper No. 24-013). Harvard Business School https://doi.org/10.1287/orsc.2025.21838

Eva, K.W. and Regehr, G. (2005) Self-Assessment in the Health Professions: A Reformulation and Research Agenda. *Academic Medicine*, 80 (1), S46-S54. https://doi.org/10.1097/00001888-200510001-00015

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., & Welsch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, *175*, Article 108779. https://doi.org/10.1016/j.chb.2025.108779

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

Girrbach, L., Alaniz, S., Smith, G., & Akata, Z. (2025). *A Large Scale Analysis of Gender Biases in Text-to-Image Generative Models*. https://doi.org/https://doi.org/10.48550/arXiv.2503.23398

Glazko, K., Mohammed, Y., Kosa, B., Potluri, V., & Mankoff, J. (2024). Identifying and Improving Disability Bias in GPT-Based Resume Screening. *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*, *1*, 687–700. https://doi.org/10.1145/3630106.3658933

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Gulikers, J.T.M., Bastiaens, T.J. & Kirschner, P.A. (2004). A five-dimensional framework for authentic assessment. *ETR&D* 52, 67–86. https://doi.org/10.1007/BF02504676

Kamruzzaman, M., Minul, M., Shovon, I., & Kim, G. L. (2023). *Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models*. https://arxiv.org/pdf/2309.08902

Koch, M. J., Carolus, A., Wienrich, C., & Latoschik, M. E. (2024). Meta AI literacy scale: Further validation and development of a short version. *Heliyon*, *10*(21), Article e39686. https://doi.org/10.1016/j.heliyon.2024.e39686

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

Lintner T. (2024). A systematic review of AI literacy scales. NPJ science of learning, 9(1), 50. https://doi.org/10.1038/s41539-024-00264-4

Martens, L., Virgin, N., Hoffarth, P., Goodwill, J., Derenne, N., Eck, R. Van, Klug, M., Manocha, G. D., & Jurivich, D. (2025). Recapitulation of Ageism in Artificial Intelligence–Generated Images: Longitudinal Comparative Study. *Journal of Medical Internet Research*, *27*(1), e68428. https://doi.org/10.2196/68428

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, *57*(9), 265. https://doi.org/10.1007/s10462-024-10903-2

Naik, R., & Nushi, B. (2023). Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 786–808). https://doi.org/10.1145/3600211.3604711

Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2024). Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology*, *55*(3), 1082–1104. https://doi.org/10.1111/bjet.13411

Phutane, M., Jung, H., Kim, D., Mitra, T., & Vashistha, A. (2025). *ABLEIST: Intersectional Disability Bias in LLM-Generated Hiring Scenarios*. https://arxiv.org/pdf/2510.10998

Sandelowski, M.J., Voils, C.I., & Knafl, G.J. (2009). On Quantitizing. *Journal of mixed methods research, 3*, 208 - 222.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9). https://doi.org/10.1093/pnasnexus/pgae346

Wang, B., Rau, P.-L. P., & Yuan, T. (2022). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*. Advance online publication. https://doi.org/10.1080/0144929X.2022.2072768

Wingerter, T. L., Straub, T., & Schweitzer, S. (2025). Mitigating automation bias in generative AI through nudges: A cognitive reflection test study. *Procedia Computer Science*, *270*, 2106–2114. https://www.sciencedirect.com/science/article/pii/S1877050925030042

Zhang, S., Xiao, R., Botelho, A. F., Liao, G., Chiu, T. K. F., Stamper, J., & Koedinger, K. R. (2026). How to assess AI literacy: Misalignment between self-reported and objective-based measures. In *LAK26: Proceedings of the 16th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery. https://doi.org/10.1145/3785022.3785088