

Does Multi-Tool Verification Work? Automation Bias and Attitude– Behavior Gaps in GenAI-Assisted Data Analysis

Revised version, 17 October 2026

Research Overview

Using locked task submissions as the outcome, this study finds that multi-tool verification did not protect business undergraduates from adopting a wrong GenAI answer, and that a survey-based blunder measure more than doubled the blunder count, producing an apparent multi-tool paradox.

Purpose of Study

Existing AI literacy frameworks embed two assumptions about reducing GenAI over-reliance: that declarative knowledge of AI limitations is protective (Bućinca et al., 2021; Chiu et al., 2024; Lintner, 2024), and that cross-checking with multiple AI tools improves output reliability, a strategy widely recommended in AI literacy guidance (Southworth et al., 2023; Allen & Kendeou, 2024; Ng et al., 2021). Both assume that knowing translates into doing, yet AI literacy is still measured mainly through self-report (Lintner, 2024). Using behavioral data, this study asks whether prior knowledge (RQ1) or attitude scales (ABI, DQV; RQ2) predict task blunders, whether multi-tool verification reduces or increases error rates (RQ3), and whether survey-based blunder measurement reflects actual task behavior (RQ4).

Research Methods

Participants were 169 undergraduate business students (Classes B, D, E) at a private university in Taiwan, surveyed in Week 2 of Spring 2026, before any AI literacy intervention. After eight exclusions ($n = 7$ missing Quiz01; $n = 1$ timestamp anomaly), 161 formed the analytic sample. Students uploaded a simulated sales dataset with embedded traps (duplicate IDs, blank-padded strings, negative gross margins) to their chosen GenAI tool and reported the average gross margin. A wrong AI answer of 33.3% appeared in students' screenshots; the simple average of the raw records is 28.1%, and the correct answer after cleaning and quantity weighting is approximately 9.97%.

The same decision was measured twice. The behavioral blunder was submitting 33.3% as the final answer in the task portal, which was locked before the survey. The self-reported blunder was choosing 33.3% on a single-choice survey item (Q24; options 33.3%, 5.1%, or not among the options). Both measures were available for 126 students. The survey also recorded

tool use, multi-tool verification, the Automation Bias Index (ABI; 4 items, $M = 2.92$, $SD = 0.63$, $\alpha = .52$; Parasuraman & Manzey, 2010) and the Data Quality Verification scale (DQV; 3 items, $M = 3.74$, $SD = 0.66$, $\alpha = .38$; Redman, 1998; Doshi-Velez & Kim, 2017); Quiz01 (6 items, $M = 4.96$, $SD = 0.93$) was merged by ID. Agreement was tested with McNemar's exact test and Cohen's κ ; tool contrasts used Fisher's exact tests, since six behavioral blunders cannot support multivariable models. The original logistic models are retained as survey-item analyses.

Results and Discussion

Finding 1: Self-report did not match behavior (RQ4). Among 126 paired students, the survey item flagged 13 blunders (10.3%); the locked submissions showed 6 (4.8%). Positive predictive value was 38.5% (5/13; 95% CI [13.9%, 68.4%]) and negative predictive value 99.1%; errors ran one way (8 false positives vs. 1 false negative; McNemar exact $p = .039$; $\kappa = .49$). Of the eight false positives, six had submitted 28–29% and two the correct answer (about 10%); neither value was an option, consistent with option cueing. A source-monitoring account could not be tested.

Finding 2: Multi-tool verification did not protect (RQ3). On the survey item, multi-tool users appeared to blunder more (17.9% vs. 8.2%; full survey model, $N = 161$: 18.9% vs. 6.5%, $OR = 2.84$, $p = .072$). In actual submissions the difference was small and non-significant (7.1% vs. 4.1%, Fisher $p = .61$); four of the five multi-tool survey blunders were false positives. The apparent paradox is therefore largely a measurement artifact: multi-tool checking is source diversification, not verification, which requires independent computation. In the paired sample, Gemini users blundered more on both measures (Fisher's exact tests: survey 16.1% vs. 4.7%, $p = .043$; submissions 8.1% vs. 1.6%, $p = .11$; five of six behavioral blunders): consistent but underpowered.

Finding 3: Knowledge and attitudes did not predict blunders (RQ1–RQ2). Quiz01 ($OR = 1.22$, $p = .556$), ABI ($OR = 1.09$, $p = .858$) and DQV ($OR = 1.58$, $p = .323$) did not predict survey-item blunders; with six behavioral events these models could not be re-estimated, so the knowledge–behavior and attitude–behavior gaps remain exploratory.

The results argue for measuring decisions rather than recollections, avoiding survey options that display the AI's output, and teaching verification habits (traceability, reproducibility, verifiability) instead of relying on declarative knowledge.

Keywords: *automation bias; generative AI; multi-tool verification; self-report validity; AI literacy education*

Note on Revision

This version replaces the abstract submitted on 13 April 2026, which described Findings 1–3 as task outcomes although the models used the survey item. It also corrects the paired sample (126, not 125) and relabels 99.1% as the negative predictive value, not specificity.