Human Judgment Before Machine Output: The VERIFY Protocol for AI-Assisted Statistical Research

Jossanth Vicente

Abstract

Generative artificial intelligence (GenAI) can explain statistical concepts, propose analyses, generate code, and draft interpretations, but fluent outputs can conceal fabricated references, invalid assumptions, privacy risks, and reasoning that learners cannot defend. This conceptual review develops the VERIFY framework for responsible GenAI use in senior high school quantitative research. Drawing on human-centred AI guidance, K-12 AI education research, statistical reporting principles, and scholarship on large language models, the framework assigns GenAI a bounded role across six actions: Verify the question and data provenance; Examine measurement and design; Review assumptions and alternatives; Independently reproduce computations; Fact-check claims and citations; and Yield a transparent human-authored interpretation. The framework maps appropriate assistance across the research cycle and proposes process portfolios, oral defenses, prompt-and-output logs, and reproducible files as assessment evidence. It treats AI literacy and statistical literacy as mutually reinforcing: learners need statistical knowledge to audit AI output, while supervised AI dialogue can expose misconceptions and support explanation. Responsible integration therefore requires neither uncritical adoption nor blanket prohibition, but designs in which verification is observable and consequential.

Keywords: generative arti icial intelligence; statistical literacy; quantitative research; secondary education; academic integrity; human-centred AI

1. Introduct on

Generative artificial intelligence has altered how students formulate questions, search for explanations, write code, and produce academic prose. In quantitative research, a conversational system can suggest a t-test, calculate an effect size, interpret a p-value, or rewrite a results paragraph within seconds. These functions may lower technical barriers, but speed and fluency are not evidence of validity. A statistically incorrect answer can appear authoritative, and a plausible citation may not exist.

The educational problem is not simply whether students will encounter GenAI, but whether they can evaluate its output while retaining ownership of the research process. UNESCO advocates a human-centred approach that protects agency, inclusion, privacy, and cultural diversity (Miao & Holmes, 2023). Its student competency framework organizes AI learning around human-centred values, ethics, foundational knowledge, applications, and system design (Miao et al., 2024). These priorities are especially important for adolescents whose projects may involve classmates, school records, attitudes, or other sensitive information.

Statistical research is a productive context for responsible AI use because claims can be traced through a visible chain: question, design, measurement, data, assumptions, computation, and interpretation. When any link is weak, polished prose cannot rescue the inference. Statistical guidance likewise discourages binary thinking based solely on p-value thresholds and emphasizes context, uncertainty, and transparency (Wasserstein & Lazar, 2016; Wasserstein et al., 2019).

This article proposes VERIFY, a six-action model for integrating GenAI into senior high school quantitative research without outsourcing judgment. It asks which tasks may benefit from bounded assistance, which require direct human control, and how assessment can make responsible reasoning visible. The framework is conceptual rather than empirical; it offers a pedagogical design for classroom implementation and evaluation.

2. Conceptual Bass

2.1 Statistical literacy as claim evaluation

Statistical literacy involves more than computation. Learners must recognize how data were produced, whether variables support proposed operations, what assumptions connect a sample to a claim, and how uncertainty constrains interpretation. A model that returns correct arithmetic for the wrong design does not provide a valid analysis. AI assistance should therefore be judged by whether it strengthens or obscures this chain of reasoning.

2.2 AI literacy as calibrated trust

AI literacy includes understanding capabilities and limitations, recognizing ethical and social consequences, and deciding when human review is necessary. Large language models generate probable sequences of text rather than guaranteed truths. Outputs can vary with wording, context, model version, and system constraints. Calibrated trust means using provisional output where it is useful while independently checking claims whose failure would affect validity, privacy, or authorship.

2.3 Human agency and integrity

Agency is preserved when students can explain why a method was chosen, reproduce the analysis without an opaque answer, identify errors, and take responsibility for the final claim. Integrity is not achieved merely by declaring AI use. It requires traceable evidence of decisions and compliance with school, journal, and data-protection rules. Private or identifiable data should not be entered into an external system without explicit authorization and adequate safeguards.

3. The VERIFY Framework

3.1 Verify the question and data provenance

The learner first states the population, variables, comparison or association, and intended claim. Provenance records how observations were collected, transformed, excluded, and stored. AI may generate alternative question formulations, but it must not invent observations, fill missing values without a declared rule, or silently change the target population.

3.2 Examine measurement and design

A proposed procedure must be evaluated against measurement level and design. A single ordinal response is not automatically continuous; scores from the same student are not independent; and correlation does not establish causation. Learners may compare candidate methods with AI, but the final choice must refer to the actual instrument, sample, grouping, and timing.

3.3 Review assumptions and alternatives

An AI-generated assumption checklist can become ritualistic. Independence is primarily a design property. Linearity requires examination of the relationship. Distributional and variance assumptions should be evaluated with residuals, sample size, balance, and robustness in view. Requiring one credible alternative reveals whether the learner understands the method rather than merely recognizing its name.

3.4 Independently reproduce computations

Every reported value should be reproducible from a visible analysis file. AI-generated code must be read, run, and checked; formulas should be tested on small cases calculated independently. Verification includes variable labels, missing-data handling, sample size, degrees of freedom, and rounding. A correct-looking paragraph is not evidence that computation occurred.

3.5 Fact-check claims and citations

An AI-generated citation remains unverified until the source is opened and matched to the claim. Verification covers author, year, title, venue, DOI or stable location, and the proposition supported. Sources should be read rather than cited from snippets. This counters bibliographic fluency that creates an appearance of scholarship without evidence.

3.6 Yield a human interpretation

The final interpretation states what was estimated, how uncertain it is, and why the magnitude matters. Exact p-values, confidence intervals, effect sizes, and descriptives should be reported where appropriate. Very small p-values are written $p < .001$, not $p = 0.000$. Causal language must match the design, and failure to reject a null hypothesis is not proof of no effect.

4. Bounded Uses Across the Research Cycle

During question development, AI may expose ambiguity or suggest alternative wording; the researcher retains control of feasibility, ethics, and the intended claim. In literature work, AI may suggest search terms or compare sources that have been supplied, but the researcher must retrieve, read, verify, and cite each source. For instruments, AI can critique clarity or flag double-barrelled items, while expert review, piloting, validity evidence, and permissions remain necessary.

For data management, AI may draft a codebook or syntax using nonsensitive examples. Identifiable data require authorization and safeguards, and all transformations must be validated. For analysis, AI may explain methods, draft code, and propose diagnostics; the researcher selects the method, runs the file, checks output, and preserves reproducibility. For writing, AI may improve organization from verified notes, but the named author owns the argument, checks every fact, follows disclosure rules, and approves the final text.

The appropriate boundary depends on consequence and verifiability. Low-consequence brainstorming may tolerate provisional output. Decisions affecting privacy, inclusion, analysis, attribution, or claims require stronger independent evidence. The same action may be acceptable in one context and inappropriate in another.

5. Assessment for Observable Verfication

A final paper alone can hide whether a student understands the analysis. Assessment should include process evidence: an approved question, design memo, variable dictionary, authorized data or synthetic substitute, analysis file, diagnostics, selected AI interactions, source-verification notes, and revision history. The purpose is traceability rather than surveillance.

Oral defense is especially valuable. Brief questions—Why is this test appropriate? What changes if observations are paired? Which assumption is consequential? What does the confidence interval permit?—reveal whether the student can defend the work. Process evidence should be graded for reasoning, not prompt sophistication.

Teachers also need task-level policies. “AI allowed” is too vague. Assignments should specify permitted functions, prohibited data, disclosure expectations, verification evidence, and consequences for fabricated sources or undisclosed substitution of authorship. Policies should address unequal access and provide a non-AI route that assesses the same outcomes.

6. Implementation Model for Schools

A staged model can align increasing AI freedom with demonstrated competence. At the foundation stage, students analyze small teacher-provided data and verify calculations manually. At the guided stage, they use AI to explain or critique methods while completing a verification log. At the independent stage, approved tools may support a project, but the learner submits reproducible evidence and defends the analysis. Advancement depends on judgment rather than tool familiarity.

Teacher learning should mirror the framework. Educators need opportunities to test model variability, inspect fabricated citations, identify privacy hazards, and design process-focused assessments. School leaders should define approved platforms, age-appropriate safeguards, data-retention expectations, accessibility provisions, and escalation routes. Governance and pedagogy must develop together.

7. Research Agenda

Future studies could compare VERIFY-guided instruction with ordinary software instruction on method selection, error detection, transfer to unfamiliar data, and interpretation quality. Process research could examine which verification actions students omit and how feedback changes behavior. Equity studies should test whether verification requirements reduce or amplify differences in access, language proficiency, disability support, and prior statistical knowledge. Because tools change rapidly, evaluations should report model, version or access date, settings, prompt context, and task conditions.

8. Concluson

Generative AI can support statistical learning only when answer generation remains subordinate to evidence evaluation. VERIFY places responsibility at six points: provenance, design, assumptions, reproducibility, factual accuracy, and interpretation. It treats AI output as neither forbidden nor authoritative; trust is conditional on visible verification. This approach links AI literacy with statistical literacy while protecting the central educational outcome: a learner who can explain, reproduce, and defend a quantitative claim.

References

Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lee, S. J., Kwon, K., & Lee, J. (2024). A systematic review of AI education in K–12 classrooms from 2018 to 2023. Computers and Education: Artificial Intelligence, 6, 100211. https://doi.org/10.1016/j.caeai.2024.100211

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Miao, F., Shiohira, K., & Lao, N. (2024). AI competency framework for students. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000391105

Sullivan, G. M., & Artino, A. R., Jr. (2013). Analyzing and interpreting data from Likert-type scales. Journal of Graduate Medical Education, 5(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values. The American Statistician, 70(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “p < 0.05.” The American Statistician, 73(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913