Protecting Statistical Reasoning from Cognitive Offloading: An AI-Resilient Framework for Malaysian Undergraduates

Khairun Nazrah Wahidin

Faculty of Education and Humanities, UNITAR International University, Petaling Jaya, Malaysia

***Corresponding author**

DOI: https://doi.org/10.47772/IJRISS.2026.102400018

Received: 18 August 2026; Accepted: 23 August 2026; Published: 25 September 2026

ABSTRACT

The utilisation of generative artificial intelligence (GenAI) by undergraduate students has exceeded that of institutional governance. Introductory statistics is particularly susceptible to this trend due to the easy delegation of its procedural tasks. It is possible for students to obtain accurate responses without engaging in the reasoning that the tasks were intended to foster. This conceptual paper examines the potential for unregulated GenAI reliance to erode the statistical reasoning of Malaysian bachelor's degree students and suggests strategies to safeguard this ability. The conceptual synthesis employed in the paper is structured. This study integrates the statistical thinking framework, cognitive load theory, cognitive offloading, self-efficacy theory, and evaluative judgement. It utilises sources from Scopus, Web of Science, ERIC, and Malaysian indexed databases, with a particular emphasis on Malaysian evidence and Malaysian Qualifications Agency guidance. Local cohorts approach statistics as a mandatory service course with documented anxiety, and national adoption is driven by perceived learning value rather than institutional facilitation, which exacerbates the concern in the Malaysian context. The primary mechanism by which GenAI use may impede three outcomes: the match between perceived and demonstrated competence, assessment validity, and statistical reasoning development is cognitive offloading. The paper suggests the AI-Resilient Statistical Reasoning (ARSR) framework in response, which is founded on three pillars: AI-transparent assessment, reasoning-first pedagogy, and calibrated AI use. It is accompanied by five testable propositions, each of which is linked to a research design, a level of analysis, and a pillar. The paper's contribution is not the individual constructs, which are well-established, but rather their statistical synthesis and conversion into falsifiable claims. Statistics educators, curriculum designers, and policymakers are provided with a structured foundation for utilising GenAI to enhance rather than replace statistical reasoning by the framework, which is in accordance with Sustainable Development Goal 4.

Keywords: cognitive offloading, generative artificial intelligence, Malaysian higher education, SDG 4 (Quality Education), statistical reasoning

INTRODUCTION

Statistics Education and Its Reform Goals

Introductory statistics courses hold a distinctive position in undergraduate education. They are compulsory for students across the social sciences, education, business, and health sciences, but they are also among the courses students approach with the greatest apprehension and the weakest sense of personal relevance (Onwuegbuzie & Wilson, 2003; Trassi et al., 2022). For several decades, reform movements in statistics education have argued that the purpose of such courses is not the mechanical execution of computations. The purpose is the development of statistical literacy, reasoning, and thinking, which involves the capacity to interrogate data, evaluate claims, reason about variation and uncertainty, and communicate evidence-based conclusions (GAISE College Report ASA Revision Committee, 2016; Koga, 2022). These competencies correspond to the data literacy and higher-order skills targeted by United Nations Sustainable Development Goal 4 (Quality Education), and particularly to Target 4.4 on equipping learners with relevant skills for employment and decent work.

The Rapid Adoption of Generative AI

The public release of large language model chatbots from late 2022 onwards has disrupted this agenda in ways the field is still working through. Generative artificial intelligence (GenAI) tools are increasingly capable of solving conventional textbook problems, interpreting statistical outputs, producing explanations of hypothesis-testing results, generating code for statistical software, and preparing polished assignment reports within a short period of time (Kasneci et al., 2023; Nikolic et al., 2023; Rudolph et al., 2023). Surveys involving undergraduate students further indicate that the use of these tools for academic purposes is already widespread and largely self-directed, with students often adopting them before formal institutional guidance has been established (Chan & Hu, 2023; Sullivan et al., 2023). A similar pattern is observable in Malaysia. In a study involving 406 Malaysian university students, Foroughi et al. (2024) found that performance expectancy, effort expectancy, hedonic motivation, and perceived learning value significantly influenced students' intentions to use ChatGPT for educational purposes. A particularly important aspect of this finding is that students' motivations are diverse and largely internally driven, suggesting that the adoption of GenAI is occurring independently of institutional encouragement. This situation gives rise to the paradox identified by Jie and Kamrozzaman (2024): although Malaysian higher education students demonstrate considerable enthusiasm for AI tools, they also face difficulties in using them effectively for learning, particularly because they may be unable to determine whether AI-generated information is accurate, relevant, or appropriate.

The Challenge for Statistics Education

The challenge facing statistics educators at the bachelor's degree level can therefore be expressed clearly. Students are now able to obtain accurate answers, interpretations, and written reports without necessarily undertaking the cognitive processes required for statistical reasoning. When such reasoning is delegated to a technological tool, the resulting grade may no longer provide reliable evidence of the student's actual competence. This creates both a learning concern, because skills that are not actively practised are unlikely to develop, and an assessment concern, because scores may lose validity as indicators of students' capabilities (Dawson et al., 2024). The issue may also have implications for equity, particularly because the proposed mechanism is influenced by affective factors. Students who begin a statistics course with a weaker quantitative foundation and higher anxiety tend to rely heavily on external tools rather than engaging directly with the reasoning process (Maat et al., 2022; Onwuegbuzie & Wilson, 2003). Because these characteristics are not evenly distributed across the student population, the negative effects of replacing reasoning with technological assistance are unlikely to be shared equally. This unequal distribution of educational cost is difficult to reconcile with Sustainable Development Goal 4's commitment to inclusive, equitable and high-quality education. The policy implications arising from this concern are considered further in the discussion section.

The Malaysian Context

The Malaysian context adds further urgency to this concern. Statistics is a compulsory service course in Malaysian universities taught across disciplines, while local students have been found to experience substantial statistics anxiety. A study conducted by Maat et al. (2022) reported that most university students experienced moderate levels of statistics anxiety and this anxiety was a significant predictor of academic achievement. At the regulatory level, the Malaysian Qualifications Agency (MQA) responded relatively early to challenges introduced by GenAI through Advisory Note No. 2/2023. This advisory note urges higher education providers to encourage prudent, responsible and integrity-based use of generative AI while ensuring that graduates continue to achieve the intended knowledge and skills specified in their programmes (Malaysian Qualifications Agency, 2023). Despite this clear guidance, discipline-specific explanations of how statistics educators should translate these expectations into teaching and assessment practices remain underdeveloped.

Research Gap and Objectives

While the literature on GenAI in higher education is expanding fast (Crompton & Burke, 2023; Lodge et al., 2023), relatively limited attention has been given to its relationship with established constructs and research traditions of statistics education. The existing discussions tend to emphasize either the broad educational affordances of GenAI or its misconduct implications. What remains underdeveloped, however, is a theoretically grounded explanation of how reliance on GenAI may influence the development of statistical reasoning, alongside a framework that translates this explanation into implications for pedagogy, assessment, and policy. Hence, this conceptual paper addresses this gap through three objectives. First, it synthesises the literature on statistical literacy, statistics anxiety, and GenAI in higher education within the context of undergraduate statistics education in Malaysia. Second, it identifies cognitive offloading as the central mechanism through which GenAI use may contribute to the decline of statistical reasoning. Third, it proposes the AI-Resilient Statistical Reasoning (ARSR) framework and formulates testable propositions to guide future empirical research. The paper's contribution is not the individual constructs it draws on, several of which are established, but their statistics-specific synthesis: it connects cognitive offloading, evaluative judgement, and validity-centred assessment into a single model for undergraduate statistics, and it converts that model into five falsifiable propositions with an explicit level of analysis and research design for each.

LITERATURE REVIEW

Statistical Literacy, Reasoning, and Thinking

Statistics education research distinguishes three interrelated learning goals: statistical literacy, statistical reasoning, and statistical thinking. Statistical literacy, as conceptualised by Gal (2002) and reinforced in more recent research, refers to the competencies individuals need to function as informed and critical consumers of statistical messages (Friedrich et al., 2024). It involves the capacity to interpret, question, and discuss statistical information encountered in everyday and professional life. It rests on both a relevant knowledge base and the disposition that Gal (2002) terms a "critical stance." Statistical reasoning concerns how individuals work with and make sense of statistical ideas. This includes connecting data to the context in which they were produced, reasoning about the shape and spread of distributions, and understanding the importance of sampling variability (Garfield & Ben-Zvi, 2008; Gok & Goldstone, 2024). Statistical thinking, the most advanced of the three learning goals, concerns the habits of mind associated with professional statistical practice. These include recognising the need for data, attending to variation, and engaging in the interrogative cycle that characterises the empirical enquiry process (Wild & Pfannkuch, 1999). Assessment research indicates that these outcomes do not develop automatically through the completion of an introductory statistics course. Students may complete a first statistics course while retaining misconceptions about sampling variability, confidence intervals, and significance testing, even when they are able to perform the associated procedures correctly (delMas et al., 2007; Gok & Goldstone, 2024). This gap between procedural performance and conceptual understanding predates GenAI, but it appears to represent the underlying vulnerability that GenAI now exposes.

Statistics Anxiety and the Affective Landscape

A second strand of research is a reminder that the statistics classroom is an affective environment before it is a cognitive one. Statistics anxiety is common among social science undergraduates and is negatively associated with achievement in statistics, while also being positively related to academic procrastination (Dodeen & Alharballeh, 2024). In the Malaysian context, Maat et al. (2022) administered the Statistics Anxiety Rating Scale to 199 university students and found that 69.3% reported moderate statistical anxiety, with anxiety contributing to the prediction of achievement. Anxious students are also those most likely to welcome a tool that promises to remove the discomfort of statistical work. GenAI therefore appears to interact with a pre-existing motivational vulnerability. For the anxious or low-efficacy student, delegation to the machine is not only convenient but also emotionally protective, which may make habitual offloading more likely among the students who can least afford it.

Generative AI in Higher Education

The broader higher education literature documents both the promise and the unease surrounding GenAI. Kasneci et al. (2023) identify substantial opportunities, including personalised explanation, feedback at scale, and support for instructors, alongside risks of overreliance, bias, and hallucinated content. Students themselves report valuing GenAI for personalised support and efficiency, while also expressing concern about accuracy, academic integrity, and the effect on their own skill development (Chan & Hu, 2023; Sullivan et al., 2023). Jie and Kamrozzaman (2024) found that students face challenges spanning overdependence, uncertainty about the reliability of AI outputs, and difficulty integrating AI assistance with genuine learning. Reviews of artificial intelligence in higher education note that educators and educational theory have been largely absent from the technical literature, which has left pedagogy to trail deployment (Crompton & Burke, 2023; Lodge et al., 2023). A systematic scoping review of large language models in education concludes that practical and ethical challenges, including reliability and equity of access, remain unresolved even as adoption accelerates (Yan et al., 2024).

Genai in the Statistics Classroom: Affordances and Risks

Applied to the statistics classroom, the affordances are real. GenAI can act as an always-available tutor that re-explains sampling distributions in several ways, generates additional practice problems, assists with statistical software syntax, and offers feedback on draft interpretations. These capabilities align with long-standing reform goals of accessible explanation and reduced computational burden (GAISE College Report ASA Revision Committee, 2016). The risks, however, appear structural rather than incidental. First, the type of task that dominates conventional statistics assessment, namely the well-structured procedural item with a determinate answer, is solved by current GenAI systems with high reliability, so unsupervised assessment built on such items no longer measures student capability. The strongest evidence for this comes from the multi-institutional benchmarking conducted by Nikolic et al. (2023), in which GenAI performance was evaluated against authentic university assessment items and was found to compromise the integrity of unsupervised tasks. Commentary on academic integrity in the disciplines reaches a similar conclusion from classroom experience rather than from systematic benchmarking (Cotton et al., 2024). It should be noted that this benchmarking was conducted in engineering education, and that no equivalent exercise has yet been reported for statistics. The inference from one discipline to the other appears defensible, since both assess mainly through structured quantitative problems with determinate solutions, and since the item types most at risk are common to both. It nevertheless remains an inference, and its limits are taken up in the Limitations section. Second, GenAI output in statistics is fluent but not reliably correct. Fabricated references, misapplied procedures, and subtly wrong interpretations occur (Kasneci et al., 2023; Rudolph et al., 2023), and detecting them requires the statistical reasoning the student may not have developed. This creates a difficult position in which the students most dependent on the tool are least equipped to check it. Third, because GenAI produces confident prose, it may mask conceptual misconceptions that classroom discussion would otherwise surface, which deprives instructors of useful diagnostic information.

Disciplinary guidance is itself changing, and its current state is worth recording accurately, because it bears on the argument in two ways. The GAISE College Report is undergoing its second major revision, released in stages for community consultation from 2025 (American Statistical Association, 2025). Two features of that draft are relevant here. First, its recommendations on technology now extend explicitly to artificial intelligence, and endorse the “appropriate and vetted use” of AI to support students’ understanding of data. The wording is significant, because it presumes rather than supplies the student capacity to vet, and therefore locates the difficulty at the same point this paper identifies. Second, the draft’s dedicated section on assessment remained under development at the time of writing. The discipline has therefore indicated that AI belongs in the statistics classroom before issuing guidance on how assessment should respond, and it is this interval that the present paper addresses. Because the revision is a draft in open consultation rather than an endorsed standard, the ASA-endorsed 2016 report is cited throughout this paper as the authoritative statement of reform goals. The draft is noted only where it indicates the direction of travel.

The Malaysian Higher Education Context

The dynamics described above take a specific institutional shape in Malaysia. GenAI acceptance research on Malaysian students indicates that uptake is driven less by institutional facilitation than by perceived performance benefits and learning value (Foroughi et al., 2024). This suggests that students will use these tools with or without formal support, and that governance must therefore work through curriculum and assessment rather than through access control. Regional evidence from elsewhere in Southeast Asia is instructive, although it must be read as comparative rather than local. In a survey of 1,000 college students in the Philippines, Asio (2024), publishing in a Malaysian journal, found that AI literacy, AI self-efficacy, and AI self-competence varied systematically by college and year level and were interrelated. This suggests that the capacity to use AI well is itself an educational outcome linked to learner confidence, rather than a uniform property of a digital-native generation. Whether the same pattern holds among Malaysian undergraduates has not been established, and the present framework treats it as an open question rather than as settled local evidence. Recent Malaysian research further documents a gap between enthusiasm for AI and the pedagogical support needed to use it well. Saravanan and Kamrozzaman (2025) found near-universal AI adoption among urban Malaysian teachers alongside insufficient professional training and institutional support, and called for TPACK-based professional development to close the gap. At the policy level, Advisory Note No. 2/2023 of the Malaysian Qualifications Agency positions responsible GenAI use as a quality assurance matter for all Malaysian higher education providers (Malaysian Qualifications Agency, 2023). What national guidance necessarily leaves open is the discipline-level translation, which concerns which competencies in a statistics course must remain demonstrably human, and how course design can protect them. The framework developed in this paper is intended to supply that translation. Three features make it appropriate to the Malaysian setting in particular rather than merely illustrated by it. First, statistics is a compulsory service course carrying documented, moderate-to-high anxiety in local cohorts (Maat et al., 2022). Second, national adoption is driven by perceived learning value rather than institutional facilitation, so access control is not a workable lever (Foroughi et al., 2024). Third, a national quality-assurance instrument already exists in Advisory Note No. 2/2023, into which discipline-level competencies can be written. The framework is designed to operate through these conditions rather than around them.

METHODOLOGY

As a conceptual paper, this work advances an argument by integrating existing theory and evidence rather than by collecting new data. Its warrant therefore depends on how clearly the synthesis is described. Literature was identified through structured searches of Scopus, Web of Science, and ERIC. These were supplemented by targeted searching of Malaysian indexed sources through MyCite and of the published advisories of the Malaysian Qualifications Agency. Search strings combined terms for the technology (generative artificial intelligence, ChatGPT, large language models), for the learning constructs (statistical literacy, statistical reasoning, statistical thinking, statistics anxiety, cognitive offloading, self-regulated learning), and for the setting (higher education, undergraduate, Malaysia). Reference lists of the retrieved reviews were also searched backwards to recover foundational sources that keyword searching alone does not reliably surface.

Two timeframes were applied. Foundational sources establishing the constructs were admitted without date restriction. These include the statistical literacy, reasoning, and thinking traditions, cognitive load theory, self-efficacy theory, and cognitive offloading. The theoretical apparatus predates the technology and is drawn upon here in its established form. Sources concerning generative AI in education were restricted to 2022 onwards, corresponding to the public release of large language model chatbots, with priority given to peer-reviewed empirical work and to systematic reviews rather than commentary. Items were retained where they contributed to at least one of three functions: characterising the endangered construct, evidencing the offloading mechanism, or informing the pedagogical, assessment, or policy response. Where sources converged, preference was given to those reporting experimental or longitudinal designs over those reporting perceptions. This distinction is preserved rather than flattened when the evidence is weighed in the analysis that follows. The empirical base on generative AI and learning outcomes is at present young, heterogeneous in design, and heavily weighted towards perception studies, which makes the aggregation a systematic review presumes difficult to support. The purpose here is also explanatory rather than summative, since the aim is to specify a mechanism and to derive testable propositions from it. Conceptual synthesis appears better suited to that purpose (Jaakkola, 2020). The consequence, acknowledged again in the Limitations section, is that what follows carries the status of a research agenda rather than of findings.

Problem Conceptualisation and Theoretical Foundations

Cognitive Offloading as the Central Mechanism

This paper reads the problem through the lens of cognitive offloading, which refers to the use of bodily action or external tools to reduce the mental demands of a task by shifting part of its information processing out of the head (Risko & Gilbert, 2016). On this account, offloading is a basic feature of human cognition rather than a modern failing. The important question is not whether students offload, but what they offload. Using statistical software to remove computational burden is a productive offload, because it frees working memory for higher-order processing. The concern arises when what is delegated is not the burden surrounding the target cognition but the target cognition itself. Recent empirical evidence supports this concern. A large survey study found a significant negative correlation between frequent AI tool use and critical thinking, mediated by cognitive offloading, with the effect most pronounced among younger users (Gerlich, 2025). A systematic review concluded that over-reliance on AI dialogue systems is associated with reduced decision-making and analytical thinking among students (Zhai et al., 2024). Experimental evidence indicates that when AI assistance is withdrawn, students who had relied on it fail to sustain the quality of work it had supported, which suggests dependence rather than internalised capability (Darvishi et al., 2024). Most directly relevant here, a randomised experiment found that learners supported by ChatGPT improved their immediate task performance but showed no advantage in knowledge gain or transfer, and engaged less in the self-regulated learning processes the authors term "metacognitive laziness" (Fan et al., 2025). The tool appears to have improved the product while circumventing the learning process. Although these findings converge, they do not carry the same evidential weight. Gerlich's (2025) findings are correlational and therefore cannot exclude the possibility that individuals with weaker critical-thinking skills are simply more likely to rely on AI. By contrast, the experimental designs employed by Darvishi et al. (2024) and Fan et al. (2025) provide stronger grounds for causal inference. Considered together, however, the studies identify a similar mechanism from different methodological perspectives. Such convergence is consistent with the view that the observed effect arises from cognitive offloading itself rather than from an artefact of self-reported behaviour. The implications for learning follow from a central premise of statistics education research: statistical reasoning develops through sustained, effortful engagement with contextualised data and participation in the interrogative cycle of empirical enquiry (Garfield & Ben-Zvi, 2008; Wild & Pfannkuch, 1999). Every task delegated to GenAI represents a reasoning opportunity that the learner does not undertake. Accumulated across a degree programme, this pattern may produce graduates who possess the formal credential of quantitative competence without the corresponding capacity for independent statistical reasoning.

The Calculator Objection and Its Limits

The argument developed thus far raises an obvious objection that warrants direct consideration. Every substantial technology introduced into quantitative education has attracted a similar prediction of cognitive decline. Calculators were expected to destroy arithmetic fluency, and statistical packages were expected to produce a generation able to run a regression without understanding one. Neither prediction was borne out in the strong form in which it was made. Statistics education came in time to regard computational software as helpful, and reform guidance explicitly endorses technology that reduces computational burden so that conceptual understanding can be given more attention (GAISE College Report ASA Revision Committee, 2016). Why, then, should generative AI be treated differently? Three differences appear to carry the answer. The first concerns the scope of what is delegated. A statistical package executes a procedure that the user must first select, having framed a question, chosen a design, and judged the assumptions. The tool operates downstream of the reasoning and leaves that reasoning intact. GenAI accepts the problem in natural language and returns the framing, the selection, the execution, and the interpretation together. This differentiation corresponds with cognitive load theory's classification of extraneous and germane processing (Sweller, 2011). Calculators primarily diminish superfluous cognitive exertion, whereas GenAI tools, if not judiciously constrained, could supplant the cognitive processes essential for substantive learning.

The second distinction is reasoning visibility. Statistics software produces fixed outputs that students must interpret. A regression table cannot explain its meaning, so the student must interpret it. GenAI can produce fluent interpretations, but linguistic fluency does not guarantee conceptual accuracy. These explanations may seem complete and convincing, so students may not realise they need more reasoning or verification. An incomplete, awkward, or incorrect interpretation would reveal valuable diagnostic evidence, but instructors lose it. A calculator won't hide a mistake, but a confident paragraph may.

The third distinction is empirical. There is no historical precedent for comparing with calculators. Earlier explorations of calculator use were based on rather little empirical evidence. The present discussion is grounded in a much wider evidence base. In line with this view, one study found that students who used AI assistance were unable to sustain the quality of their work once that support was removed (Darvishi et al., 2024). A comparable pattern has been found for ChatGTP. It improved immediate task performance but did not lead to corresponding increases in knowledge or transfer (Fan et al., 2025). These results suggest something the calculator analogy doesn't quite capture.

These concerns do not go far enough to justify banning GenAI, and the proposed framework does not recommend doing so. The value of the objection is not to say that the more pessimistic argument is less persuasive, but to show the condition under which it is less persuasive. Technologies that support reasoning are more likely to reduce unnecessary mental effort, rather than replace the thinking that learners need to do for themselves. That's been the traditional role of calculators. If GenAI is prudently and appropriately limited in its deployment, it may play a similar role. This view is supported by evidence from tutor-configured applications where positive learning outcomes have been reported when GenAI is used to explain, question and prompt students rather than simply generate answers (Kasneci et al., 2023). The real problem is not the technology but the instructional design with which it is used. This is the basic proposition of the framework developed in the next section.

The Assessment Validity Problem

The second face of the problem is assessment validity. An assessment is valid to the extent that scores support inferences about the intended construct. Recent scholarship argues that in the GenAI era, validity rather than cheating should be the organising concern of assessment reform (Dawson et al., 2024). When students can complete take-home statistics tasks using GenAI, scores reflect access to and skill with the tool rather than statistical reasoning, and the inference from grade to competence breaks down. Detection-based responses are unreliable and adversarial, and the literature increasingly favours redesign over policing, which means constructing assessment so that the construct cannot be demonstrated without genuine student cognition (Chiu, 2024; Cotton et al., 2024; Rudolph et al., 2023). Statistics education is comparatively well positioned for such redesign, because its constructs can be probed through personalised data, oral defence of analytical decisions, critique of flawed analyses, and in-class reasoning tasks. These formats nevertheless remain the exception in large undergraduate service courses, where marking workloads favour standardised procedural items.

Theoretical Foundations

Four theoretical lenses anchor the framework developed below. First, Wild and Pfannkuch's (1999) statistical thinking framework specifies the dimensions of statistical thinking, including the investigative cycle, consideration of variation, the interrogative cycle, and dispositions such as scepticism. It defines the construct at risk and identifies the components students need in order to check GenAI output. Statistical thinking is therefore both the object the framework seeks to protect and the capability that makes safe GenAI use possible.

Second, cognitive load theory holds that learning occurs when working memory resources are devoted to processing that builds schemas in long-term memory (Sweller, 2011). The theory helps explain why GenAI is both promising and risky. Offloading extraneous load, such as arithmetic and syntax errors, is beneficial, whereas offloading germane processing, such as selecting procedures and reasoning from output to conclusion, removes the engagement through which statistical schemas form. GenAI does not distinguish between the two, so load allocation becomes a design responsibility of the instructor and a self-regulation responsibility of the student.

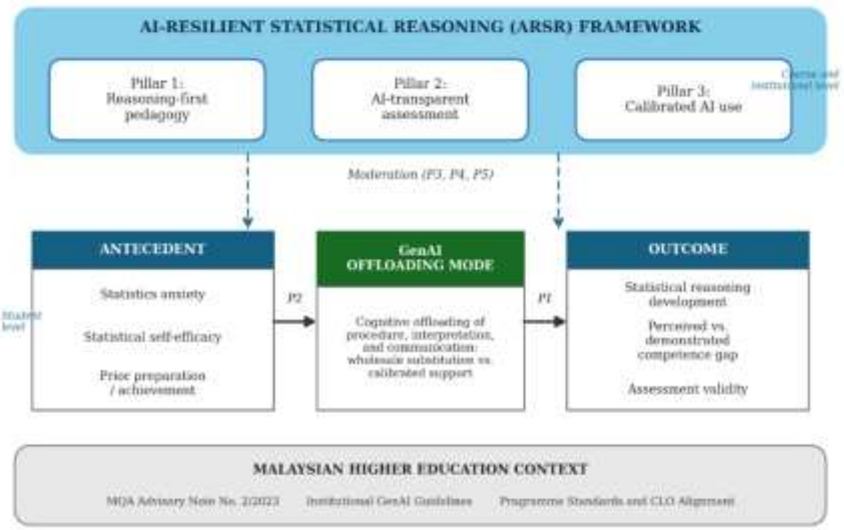
Third, self-efficacy theory supplies the motivational mechanism (Bandura, 1997). Wholesale offloading deprives students of the mastery experiences from which statistical self-efficacy grows, and this may create a reinforcing loop in which low efficacy motivates offloading, offloading prevents mastery, and absent mastery keeps efficacy low. Perceived competence may nevertheless inflate even as actual competence stagnates, because polished GenAI-assisted products provide misleading performance feedback.

Fourth, and connecting the preceding three to the pedagogical response, Bearman and Ajjawi's (2023) account of learning to work with the black box supplies the construct of evaluative judgement. This refers to the capacity to appraise the quality of work, one's own and that produced by others or by machines, against disciplinary standards. Their argument is that pedagogy for a world containing opaque and fallible artificial intelligence can rest on neither prohibition nor trust, and must instead develop the learner's capacity to judge. In the statistical case this converges with Wild and Pfannkuch (1999) from a different direction. What the statistical thinking framework supplies as a disposition, namely scepticism within the interrogative cycle, evaluative judgement supplies as an appraisal capability directed at an external producer of analysis. Taken together, the two specify what the auditing competence referred to throughout this paper consists of. They also explain why the framework treats the ability to interrogate GenAI output as a learning outcome rather than as an incidental by-product of subject mastery.

The Proposed Conceptual Framework

Drawing the preceding analysis together, this paper proposes the AI-Resilient Statistical Reasoning (ARSR) framework. What distinguishes the framework from broader AI-in-education models is that it is specified for measurement: each pillar is tied to a proposition, and its central learner-side construct, calibrated AI use, is defined through three measurable dimensions rather than left as a disposition. The framework's premise is that the goal of undergraduate statistics education in the GenAI era is not to exclude the technology, which is neither feasible nor desirable. The goal is to develop statistical reasoning that is resilient to it, meaning reasoning that students can perform without the tool, and that enables them to direct, interrogate, and check the tool when they use it. This premise follows from the account of evaluative judgement set out above (Bearman & Ajjawi, 2023), since reasoning that cannot be performed unaided cannot be used to appraise what the tool returns. Figure 1 presents the framework as a conceptual diagram.

Figure 1. The AI-Resilient Statistical Reasoning (ARSR) conceptual framework



Note. Solid arrows denote the hypothesised direct relationships tested by Propositions 1 and 2, while dashed arrows denote the moderating influence of the three ARSR pillars tested by Propositions 3 to 5. The framework operates at two levels. The antecedent, offloading, and outcome chain is a student-level mechanism, whereas the three pillars act as moderators at the course and institutional levels. The framework is conditioned within the Malaysian higher education policy context, anchored by Advisory Note No. 2/2023 of the Malaysian Qualifications Agency. As observed in Figure 1, the framework models a mediating chain. Affective and cognitive antecedents, namely statistics anxiety, statistical self-efficacy, and prior preparation, shape the mode of GenAI cognitive offloading a student adopts, which ranges from wholesale substitution to calibrated support. This in turn shapes three outcomes: the development of statistical reasoning, the gap between perceived and demonstrated competence, and the validity of assessment. The three ARSR pillars operate as moderators at the course and institutional levels. Although they cannot remove the factors that lead to offloading, they may guide students towards more appropriate GenAI use and thereby protect the intended outcomes.

One feature of the framework requires explicit comment, because it may otherwise be read as an inconsistency. The three outcomes differ in nature. The development of statistical reasoning and the gap between perceived and demonstrated competence relate to individual students. Assessment validity is not. It is a property of an assessment instrument and of the inferences drawn from it within a course, and it is degraded not by any single student's offloading but by the aggregate availability of GenAI to a cohort. The framework is therefore multilevel in nature. Student-level offloading affects individual outcomes, while its cumulative effect across a cohort may weaken a course-level property. This distinction has direct methodological implications. Propositions 1, 2, 3, and 5 are student-level claims that can be tested within a single course or across pooled cohorts. Proposition 4 is a course-level claim requiring either between-course comparison or a multilevel design in which students are nested within courses. Treating it as a student-level correlation would confound the unit of analysis with the construct. The pillars are likewise not student attributes. Reasoning-first pedagogy and AI-transparent assessment are course-design decisions, and only calibrated AI use resides in the learner. Figure 1 represents the three as a single moderating band for compactness, but they are better read as operating at two distinct levels.

The first pillar, reasoning-first pedagogy, requires that course design allocate protected, tool-free engagement to the germane processes identified by cognitive load theory. These are procedure selection, interpretation of output, reasoning about variation, and evaluation of assumptions. Class time shifts from the demonstration of procedures, which GenAI has made routine, towards structured argumentation about data, error analysis, and critique of statistical claims. This is consistent with long-standing reform recommendations (GAISE College Report ASA Revision Committee, 2016). GenAI is used specifically as a tool for explanation and practice generation, while its known limitations in statistics are taught as part of the course. In this way, its weaknesses become learning material that supports the interrogative cycle.

The second pillar, AI-transparent assessment, involves redesigning assessment tasks so that their validity is maintained even when GenAI tools are readily available (Chiu, 2024; Dawson et al., 2024). This pillar comprises four main assessment approaches. First, personalised or student-generated datasets reduce the usefulness of generic responses by making them visibly inconsistent with the specific data under analysis. Second, secured synchronous components, including invigilated reasoning tasks and viva-style defences of submitted analyses, provide direct evidence of individual competence. Third, process-visible artefacts, such as analysis logs, decision journals, and documented prompts, allow assessment to focus on how effectively students direct, evaluate, and verify the tool's output. Fourth, critique-based tasks require students to identify and explain errors in GenAI-generated analyses, thereby directly assessing the evaluative and auditing capabilities that the framework seeks to develop.

The third pillar, calibrated AI use, focuses on the learner's metacognitive and self-regulatory capacity to determine when cognitive offloading supports learning and when it replaces the reasoning that learning requires. This capacity can be developed through explicit instruction on cognitive offloading, staged disclosure practices that make AI use open to discussion rather than concealed, and reflective activities that require students to compare their independent reasoning with GenAI-generated responses. The pillar is also grounded in self-efficacy theory, as calibrated AI use preserves opportunities for mastery experiences and provides more accurate feedback on the learner's own performance. In doing so, it helps protect the cycle

ISSN: 2454-6186 | DOI: 10.47772/IJRISS | INCOEd 2026 | Special Issue | Volume X Issue XXIV September 2026

through which self-efficacy is developed, a cycle that may be disrupted when students rely on GenAI to perform the reasoning on their behalf (Bandura, 1997).

Because Proposition 3 relies on this construct, calibrated AI use requires an explicit operational definition rather than a purely descriptive account. It is conceptualised here as comprising three measurable dimensions. The first is monitoring accuracy, defined as the degree of correspondence between a student's confidence in analytical judgement and the actual correctness of that judgement. This dimension may be assessed through item-level confidence ratings paired with performance scores and expressed using calibration or bias indices commonly applied in metacognitive research. The second facet is offloading discrimination, which refers to a student's ability to decide when GenAI reduces unnecessary effort and when it replaces essential thinking. It can be measured using scenario-based tasks in which students judge whether GenAI use is appropriate for different types of activities, or by examining their recorded prompts against the reasoning steps that were delegated to the tool. Thirdly, disclosure fidelity, defined as the degree of alignment between a student's reported AI use and the use evidenced in process artefacts. This dimension serves both as an indicator of academic integrity and as a measure of whether students have genuinely internalized disclosure expectations rather than following them only to meet formal requirements. One implication of this specification deserves emphasis. Calibrated use is not low use. A student who uses GenAI heavily but accurately, discriminately, and transparently is well calibrated, whereas a student who uses it rarely but cannot say why is not. The construct is therefore separate from volume of use, and any instrument that conflates the two would fail to test Proposition 3. From the framework, five propositions are advanced to guide empirical research.

Proposition 1. Higher habitual offloading of interpretive statistical tasks to GenAI is negatively associated with performance on tool-free measures of statistical reasoning, when prior achievement is controlled.

Proposition 2. Statistics anxiety and low statistical self-efficacy positively predict wholesale offloading to GenAI, and GenAI reliance mediates the relationship between statistics anxiety and statistical reasoning development.

Proposition 3. Unstructured GenAI use widens the gap between perceived and demonstrated statistical competence, while calibrated use narrows it.

Proposition 4. Courses implementing AI-transparent assessment show a stronger association between course grades and independent measures of statistical reasoning than courses retaining conventional take-home procedural assessment.

Proposition 5. Reasoning-first pedagogy that treats the statistical failure modes of GenAI as explicit course content, rather than leaving them to incidental discovery, improves students' ability to detect errors in AI-generated analyses, and this auditing ability is positively associated with overall statistical thinking.

The five propositions are not all of one kind, and treating them as though they were would obscure both what each claims and what testing it would require. Propositions 1 and 2 concern the mechanism itself and are neutral as to pedagogical response. They specify what the framework holds to be occurring in the absence of intervention, and they establish the problem to which the pillars are the proposed answer. Propositions 3, 4, and 5 concern the pillars, with one proposition for each, so that no pillar is advanced without a corresponding empirical commitment. Table 1 sets out these assignments, together with the level at which each proposition operates and the design it implies.

Table 1. Mapping of propositions to ARSR pillars, levels of analysis, and implied research designs

Proposition	Claim tested	ARSR pillar addressed	Level of analysis	Implied design
P1	Habitual offloading of interpretive statistical tasks is negatively associated	None. This specifies the mechanism the	Student	Cross-sectional survey pairing an offloading measure with a tool-free

	with tool-free statistical reasoning.	pillars are designed to counter.		reasoning assessment, controlling for prior achievement.
P2	Statistics anxiety and low statistical self-efficacy predict wholesale offloading, which mediates their relation to reasoning development.	None. This motivates Pillar 3 by locating the affective origin of offloading.	Student	Mediation model estimated by PLS-SEM or equivalent. Longitudinal replication is preferred to support the causal ordering.
P3	Unstructured GenAI use widens the gap between perceived and demonstrated competence, whereas calibrated use narrows it.	Pillar 3: calibrated AI use	Student	Paired confidence and performance measures yielding a calibration index, alongside scenario-based offloading discrimination and disclosure fidelity. Pre-post or comparative-cohort design with calibration instruction as the contrast.
P4	Courses using AI-transparent assessment show stronger correspondence between grades and independent reasoning measures than courses retaining take-home procedural assessment.	Pillar 2: AI-transparent assessment	Course, with students nested within courses	Between-course comparison. Multilevel modelling is required, since the construct is a property of the course rather than of the student.
P5	Reasoning-first pedagogy that teaches GenAI failure modes as course content improves error detection, which is associated with statistical thinking.	Pillar 1: reasoning-first pedagogy	Student, within a course-level instructional condition	Quasi-experimental comparison of instructed and uninstructed cohorts using a purpose-built error-detection instrument.

Note. Propositions 1 and 2 specify the offloading mechanism the framework posits, while Propositions 3 to 5 test the three pillars advanced in response to it. Proposition 4 is the only course-level claim, and it cannot be tested by aggregating student-level correlations.

DISCUSSION AND IMPLICATIONS

Implications for Pedagogy

For lecturers, the framework implies a reallocation of instructional time rather than an increase in it. Procedural demonstration can be delegated to worked materials and, where appropriate, to GenAI itself, which frees contact hours for the discussion, critique, and error analysis through which reasoning develops. Lecturers also require professional development that integrates technological, pedagogical, and content knowledge for the GenAI context. Recent extensions of the TPACK framework to generative AI provide a suitable organising structure for this (Mishra et al., 2023). It also reflects the recommendation from UNITAR-based research that TPACK-grounded professional development, rather than tool training alone, is the mechanism for closing the AI implementation gap in Malaysian education (Saravanan & Kamrozzaman, 2025). In teacher education programmes, statistics courses carry a double burden, because pre-service teachers must both develop their own statistical reasoning and form professional norms about technology use that they will pass on to school classrooms.

Implications for Assessment Design

For assessment designers, the practical agenda is to move summative weight towards formats that GenAI cannot complete on the student's behalf. These include invigilated conceptual and interpretive items, oral defence, personalised datasets, and critique of flawed analyses. Take-home analytical work need not be abandoned, but its role shifts from certification to formation, with process documentation and disclosed AI use assessed rather than prohibited. This redesign agenda is consistent with the conclusion that structural design, not detection, is the sustainable response to GenAI capability, and that validity should be the organising concern (Chiu, 2024; Cotton et al., 2024; Dawson et al., 2024). It should be acknowledged that this agenda currently runs ahead of disciplinary guidance rather than deriving from it, since the assessment section of the GAISE revision was still in preparation at the time of writing (American Statistical Association, 2025). The second pillar is therefore offered as one contribution to that gap, and should be re-examined against the ASA's guidance once it is issued.

Implications for Policy and SDG 4

For Malaysian institutions and quality assurance bodies, the framework argues against both prohibitionist and laissez-faire positions. The Malaysian Qualifications Agency's Advisory Note No. 2/2023 already articulates the principle that GenAI should be used responsibly and with integrity while learning outcomes are safeguarded (Malaysian Qualifications Agency, 2023). The task now is operationalisation. Policy should follow the multi-dimensional approach recommended in the emerging literature, which addresses governance, pedagogy, and operational support together (Chan, 2023). In practical terms, programme standards and course learning outcome statements for quantitative courses should be revised to name AI-resilient competencies explicitly, including unaided statistical reasoning and the auditing of machine-generated analysis. This would allow accreditation and curriculum review to create incentives aligned with the pedagogical agenda described above. Because Malaysian evidence indicates that student adoption is driven by perceived learning value rather than institutional facilitation (Foroughi et al., 2024), policy that works through assessment design and through the deliberate development of AI literacy appears more likely to succeed than policy that attempts to restrict access. Comparative Southeast Asian evidence that AI literacy and AI self-efficacy vary markedly across student groups (Asio, 2024, in a Philippine sample) further suggests that such capability cannot be assumed to be evenly distributed across a cohort, and should therefore be built into programme design rather than left to student initiative. Framed against SDG 4, the implications are clear. Target 4.4 commits education systems to equipping graduates with relevant skills, and Target 4.7 to the knowledge needed for sustainable development. Statistics courses that certify competence students do not possess would fall short of both, whereas AI-resilient design may convert GenAI from a threat to quality education into an instrument of it.

The equity concern raised at the outset can now be stated more precisely. If wholesale offloading is driven in part by statistics anxiety and weak prior preparation, as Proposition 2 holds, then the erosion of statistical reasoning is unlikely to fall evenly across a cohort. It may concentrate among students who entered least well prepared, who are disproportionately those for whom the credential carries the greatest weight, and who are least equipped to detect that the substitution has occurred. An unregulated GenAI environment is therefore not neutral with respect to educational equity. A prohibitionist environment is not neutral either, since it penalises the same students for using a support their better-prepared peers do not need. The framework's response is that equity here is served by design rather than by rule. Reasoning-first pedagogy protects the mastery experiences that low-efficacy students most require, AI-transparent assessment removes the advantage that undisclosed tool use confers, and calibrated use is taught rather than presumed. Target 4.5 of SDG 4 commits education systems to eliminating disparities in access to quality education. Treating AI literacy and calibrated use as course-delivered entitlements, rather than as background competencies students are assumed to arrive with, is the operational form that commitment takes in a statistics curriculum.

LIMITATIONS

The central limitation of this work is the one every conceptual paper carries. Its claims, however carefully grounded in theory, remain claims until they are tested. The framework is built on inference from established theory and from early GenAI scholarship, which is a body of literature that is itself young, rapidly moving, and dominated by perception studies rather than learning-outcome evidence. The propositions should therefore be read as a research agenda rather than as settled conclusions. In addition, the analysis is framed at the level of the typical large-enrolment introductory course. Boundary conditions, such as small honours cohorts or fully project-based curricula, may alter the dynamics described.

One evidential gap should be named specifically, since the argument depends on it. The claim that current GenAI systems reliably solve conventional statistics assessment items rests on benchmarking conducted in a related quantitative discipline rather than in statistics itself. A discipline-specific benchmarking study would be valuable, in which current systems are evaluated against a representative sample of introductory statistics assessment items stratified by item type, from routine computation through output interpretation to assumption evaluation and critique. Such a study would establish which parts of the existing assessment repertoire are in fact compromised and which retain their discriminating power. It is logically prior to the five propositions rather than one of them, because it would calibrate the urgency of the case advanced here. If it were to show that interpretive and assumption-checking items resist automation more than is assumed, the framework would remain sound, but the redesign agenda could be correspondingly narrower. A related caveat concerns the standard against which the framework should ultimately be judged, which is itself provisional. The GAISE revision released for consultation from 2025 was incomplete when this paper was written, and its recommendations may change before endorsement (American Statistical Association, 2025). The framework is therefore developed against the endorsed 2016 report, and readers should treat its alignment with disciplinary guidance as subject to revision.

CONCLUSION

Generative artificial intelligence has not created the weaknesses of undergraduate statistics education, but it has made them harder to sustain. Courses that certified procedural performance loosely coupled to reasoning could survive as long as the procedures still had to pass through the student's mind. They are unlikely to survive a technology that performs procedure, interpretation, and prose alike. This paper has conceptualised the resulting problem as one of cognitive offloading, and has identified its mechanisms through the statistical thinking framework, cognitive load theory, self-efficacy theory, and evaluative judgement. It has proposed the AI-Resilient Statistical Reasoning framework, built on reasoning-first pedagogy, AI-transparent assessment, and calibrated AI use, as a constructive response grounded in the Malaysian higher education context and aligned with SDG 4 (Quality Education). Future research should test the five propositions using designs that pair validated measures of statistical reasoning with behavioural measures of GenAI use. Cross-sectional studies employing partial least squares structural equation modelling (Hair et al., 2022) could examine the mediating role of GenAI reliance between statistics anxiety and reasoning outcomes among Malaysian undergraduate cohorts. Such studies could use instruments previously administered with Malaysian samples, such as the Statistics Anxiety Rating Scale employed by Maat et al. (2022), subject to fresh validation in the target population. Quasi-experimental course-level studies could evaluate the assessment redesigns proposed under the second pillar. The choice facing statistics educators is not whether students will use GenAI, but whether graduates will command statistical reasoning that the tool can extend, or hold credentials that no longer reflect it.

ACKNOWLEDGEMENTS

The author would like to thank UNITAR International University for supporting this research.

REFERENCES

1. American Statistical Association. (2025). Guidelines for assessment and instruction in statistics education (GAISE) college report 2025 [Draft in development]. Retrieved July 21, 2026, from

https://amstat.quarto.pub/college-gaise

2. Asio, J. M. R. (2024). AI literacy, self-efficacy, and self-competence among college students: Variances and interrelationships among variables. Malaysian Online Journal of Educational Sciences, 12(3), 44–60. https://doi.org/10.22452/aldad.vol12no3.4
3. Bandura, A. (1997). Self-efficacy: The exercise of control. W. H. Freeman.
4. Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. British Journal of Educational Technology, 54(5), 1160–1173. https://doi.org/10.1111/bjet.13337
5. Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20, Article 38. https://doi.org/10.1186/s41239-023-00408-3
6. Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20, Article 43. https://doi.org/10.1186/s41239-023-00411-8
7. Chiu, T. K. F. (2024). Future research recommendations for transforming higher education with generative AI. Computers and Education: Artificial Intelligence, 6, Article 100197. https://doi.org/10.1016/j.caeai.2023.100197
8. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148
9. Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20, Article 22. https://doi.org/10.1186/s41239-023-00392-8
10. Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. Computers & Education, 210, Article 104967. https://doi.org/10.1016/j.compedu.2023.104967
11. Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. Assessment & Evaluation in Higher Education, 49(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662
12. Dodeen, H., & Alharballeh, S. (2024). Predicting statistic anxiety by attitude toward statistics, statistics self-efficacy, achievement in statistics and academic procrastination among students of social sciences colleges. Journal of Applied Research in Higher Education, 17(4). https://doi.org/10.1108/JARHE-01-2024-0021
13. delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. Statistics Education Research Journal, 6(2), 28–58.
14. Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology, 56(2), 489–530. https://doi.org/10.1111/bjet.13544
15. Friedrich, A., Schreiter, S., Vogel, M., Becker-Genschow, S., Brünken, R., Kuhn, J., Lehmann, J., & Malone, S. (2024). What shapes statistical and data literacy research in K–12 STEM education? A systematic review of metrics and instructional strategies. International Journal of STEM Education, 11, Article 58. https://doi.org/10.1186/s40594-024-00517-z
16. Foroughi, B., Senali, M. G., Iranmanesh, M., Khanfar, A., Ghobakhloo, M., Annamalai, N., & Naghmeh-Abbaspour, B. (2024). Determinants of intention to use ChatGPT for educational purposes: Findings from PLS-SEM and fsQCA. International Journal of Human–Computer Interaction, 40(17), 4501–4520. https://doi.org/10.1080/10447318.2023.2226495
17. GAISE College Report ASA Revision Committee. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report 2016. American Statistical Association. https://www.amstat.org/education/gaise
18. Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. International Statistical Review, 70(1), 1–25. https://doi.org/10.2307/1403713
19. Garfield, J., & Ben-Zvi, D. (2008). Developing students' statistical reasoning: Connecting research and teaching practice. Springer. https://doi.org/10.1007/978-1-4020-8383-9

20. Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), Article 6. https://doi.org/10.3390/soc15010006
21. Gok, S., & Goldstone, R. L. (2024). How do students reason about statistical sampling with computer simulations? An integrative review from a grounded cognition perspective. Cognitive Research: Principles and Implications, 9, Article 33. https://doi.org/10.1186/s41235-024-00561-x
22. Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). A primer on partial least squares structural equation modeling (PLS-SEM) (3rd ed.). Sage.
23. Jaakkola, E. (2020). Designing conceptual articles: Four approaches. AMS Review, 10(1–2), 18–26. https://doi.org/10.1007/s13162-020-00161-0
24. Jie, A. L. X., & Kamrozzaman, N. A. (2024). The challenges higher education students face in using artificial intelligence (AI) against their learning experiences. Open Journal of Social Sciences, 12(10), 362–377. https://doi.org/10.4236/jss.2024.1210025
25. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274
26. Koga, S. (2022). Characteristics of statistical literacy skills from the perspective of critical thinking. Teaching Statistics, 44(2), 59–67. https://doi.org/10.1111/test.12302
27. Lodge, J. M., Thompson, K., & Corrin, L. (2023). Mapping out a research agenda for generative artificial intelligence in tertiary education. Australasian Journal of Educational Technology, 39(1), 1–8. https://doi.org/10.14742/ajet.8695
28. Maat, S. M., Mohd, N., Che Ahmad, C. N., & Puteh, M. (2022). Statistical anxiety and achievement among university students. TEM Journal, 11(1), 420–426. https://doi.org/10.18421/TEM111-53
29. Malaysian Qualifications Agency. (2023). Advisory note no. 2/2023: The use of generative artificial intelligence technology in higher education. MQA.
30. Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the age of ChatGPT and generative AI. Journal of Digital Learning in Teacher Education, 39(4), 235–251. https://doi.org/10.1080/21532974.2023.2247480
31. Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. European Journal of Engineering Education, 48(4), 559–614. https://doi.org/10.1080/03043797.2023.2213169
32. Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments—a comprehensive review of the literature. Teaching in Higher Education, 8(2), 195–209. https://doi.org/10.1080/1356251032000052447
33. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002
34. Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? Journal of Applied Learning and Teaching, 6(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9
35. Saravanan, T., & Kamrozzaman, N. A. (2025). Artificial intelligence in primary education: Teacher perceptions, instructional decisions, and integration challenges in urban Malaysia. International Journal of Research and Innovation in Social Science, 9(9), 9393–9406. https://doi.org/10.47772/IJRISS.2025.909000772
36. Sabbag, A., Zieffler, A., & Ng, C. (2025). Can we distinguish statistical literacy and statistical reasoning? Statistics Education Research Journal, 24(1). https://doi.org/10.52041/serj.v24i1.587
37. Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. Journal of Applied Learning and Teaching, 6(1), 31–40. https://doi.org/10.37074/jalt.2023.6.1.17
38. Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), Psychology of learning and motivation (Vol. 55, pp. 37–76). Academic Press. https://doi.org/10.1016/B978-0-12-387691-

1.00002-8

39. Trassi, A. P., Leonard, S. J., Rodrigues, L. D., Rodas, J. A., & Santos, F. H. (2022). Mediating factors of statistics anxiety in university students: A systematic review and meta-analysis. Annals of the New York Academy of Sciences, 1512(1), 76–97. https://doi.org/10.1111/nyas.14746
40. Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. International Statistical Review, 67(3), 223–265. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x
41. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55(1), 90–112. https://doi.org/10.1111/bjet.13370
42. Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. Smart Learning Environments, 11, Article 28. https://doi.org/10.1186/s40561-024-00316-7