Overcoming the more measurements, the better intuitive heuristic in physics data evaluation: The role of inhibitory control

Wenting Hong,[1,2] Tong Li,[1,2] Li Wang,[3] Rui Dai,[4,§] Lei Bao,[5,6,7,‡] Keke Yu,[8,†] and Yang Xiao[1,2,*]

[1]*Guangdong Basic Research Center of Excellence for Structure and Fundamental Interactions of Matter, National Demonstration Center for Experimental Physics Education, School of Physics, South China Normal University, Guangzhou, Guangdong 510006, China*
[2]*Key Laboratory of Atomic and Subatomic Structure and Quantum Control (Ministry of Education), South China Normal University, Guangzhou, Guangdong 510006, China*
[3]*College of Physics and Electronic Engineering, Sichuan Normal University, Chengdu 610101, China*
[4]*School of Physics, Northeast Normal University, Changchun 130024, China*
[5]*School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
[6]*Research Center for Learning Science, Southeast University, Nanjing 210096, China*
[7]*Key Laboratory of Child Development and Learning Science (Ministry of Education), Southeast University, Nanjing 210096, China*
[8]*Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents (South China Normal University), Ministry of Education, and Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, Guangdong 510631, China*

(Received 28 April 2025; accepted 5 November 2025; published 19 December 2025)

In a data-driven world, the ability to critically assess empirical data is a core aspect of scientific literacy and a major goal of physics education. Yet, students often fall back on heuristics that conflict with statistical reasoning. A common example is the "the more measurements, the better the quality of the data" (MMB) heuristic, which naively equates quantity with quality, ignoring variance. According to conceptual change theory, such heuristics or misconceptions coexist with formal scientific knowledge and are automatically triggered during reasoning, resulting in conflicts and requiring active suppression. Inhibitory control, an executive function that overrides automatic responses, may be the central key to resolving these conflicts; however, its role in procedural domains, such as data literacy, is understudied. This study examines whether and how inhibitory control facilitates students to overcome the MMB heuristic in different representational formats (tabular vs pictorial). Using a modified Stroop paradigm, undergraduate physics students ($N = 30$) evaluated data quality presented in tabular (numerical tables) and pictorial (dot plots) formats across the congruent, incongruent, and neutral stimulus types. Results showed significant accuracy reductions in incongruent trials compared to both neutral and congruent trials across both formats, indicating significant Stroop effects. For the response times, we found a significant Stroop effect (longer response time in the incongruent than congruent trials) in the pictorial format, while a similar but non-significant trend in the tabular format. These findings first indicated the involvement of inhibitory control when the heuristic activation occurred in both formats. However, the pictorial format may evoke stronger or more persistent heuristic conflict, or the potential higher cognitive processing demands in the tabular format may mask the subtle RT-based Stroop effects. These findings extend the conceptual change research by identifying inhibitory control as crucial for resolving heuristic interference in procedural reasoning, and they stress the need to consider how representational formats modulate both heuristic activation and cognitive processing demands when designing instruction in physics.

DOI: 10.1103/lw7v-86jq

*Contact author: xiaoyang@m.scnu.edu.cn
†Contact author: kkyu@m.scnu.edu.cn
‡Contact author: lei.bao@outlook.com
§Contact author. dair986@nenu.edu.cn

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

I. INTRODUCTION

In an era increasingly driven by data, the ability to collect, interpret, and critically evaluate information is considered a cornerstone of scientific literacy [1]. The growing emphasis on data literacy reflects this shift. Data literacy has been broadly defined as the capacity to analyze and transform raw data into practical knowledge [2]. Within the science education community, the ability to organize and evaluate data is viewed as a particularly crucial subdomain of data literacy, especially in physics, which is a discipline deeply rooted in experimental measurement and inference.

Enhancing the capacity in data literacy is emphasized in curricular frameworks across all educational systems. For example, the Next Generation Science Standards (NGSS), the Common Core State Standards (CCSS), the International Baccalaureate (IB), and China's General Senior Secondary Physics Curriculum Standards [3–8] all include explicit learning objectives regarding the organization and interpretation of data. In contemporary physics education, the ability to critically evaluate the quality of empirical data is increasingly recognized as a fundamental scientific literacy skill. Beyond merely collecting or calculating data, students must learn to discern the reliability and appropriateness of evidence in ways that mirror authentic scientific inquiry.

As data analysis and modeling have become central components of science and physics curricula worldwide, increasing attention has been paid to how students evaluate the quality of experimental data. After conducting experiments, learners are expected to determine whether the data support valid conclusions, which often requires them to consider measurement uncertainty, consistency, and variability [9]. However, research consistently shows that students often struggle with such evaluations because they tend to rely on fast, effortless, and cognitively economical intuitive responses rather than systematic statistical reasoning [10–12].

These intuitive difficulties can be traced to two distinct but sometimes overlapping sources: heuristics and misconceptions. Heuristics are fast, automatic, and content-light reasoning strategies that help individuals make judgments under cognitive constraints [13,14]. While heuristics may be adaptive in everyday life [15], they are often triggered by salient but irrelevant features of a task, leading to systematic errors in scientific contexts [16]. In science and physics education, such heuristics typically operate unconsciously and are grounded in perceptual cues rather than formal scientific principles, making them particularly difficult for students to recognize and override. In contrast, misconceptions refer to stable, content-rich, yet scientifically inaccurate conceptual understandings that students construct through experience or prior instruction [17]. Although both can produce similar patterns of incorrect intuitive responses, their cognitive origins differ: heuristics are rooted in bottom-up, perceptual processing, while misconceptions reflect top-down, knowledge-based reasoning.

One prominent example in physics measurement reasoning is the belief that "the more measurements, the better" (MMB). This heuristic equates data quantity with data quality while neglecting statistical considerations such as spread or variability [18]. Students who adopt this strategy may focus on the number of repeated measurements while overlooking the dispersion of the data, thereby demonstrating a surface-level understanding of scientific measurement practices. Although the MMB belief is primarily heuristic in nature, it may gradually become stabilized and reified into a misconception through repeated reinforcement over time. However, the distinction between heuristics and misconceptions as such is not the main focus of the present study. Rather, we treat MMB as a prototypical case of heuristics that can mislead students' judgments under certain task conditions. Moreover, these types of heuristics are not limited to data evaluation; they are representative of a broader challenge in physics education, namely the persistence of heuristics- or misconception-based reasoning in conflict with normative scientific understanding [19,20].

Recent conceptual change models have emphasized that such heuristics or misconceptions are not passively replaced by scientific concepts but instead coexist with them [21]. Under task conditions that activate intuitive reasoning, these heuristics or misconceptions can reemerge and compete with formally learned scientific knowledge [22]. Dual-process theories posit that this conflict arises from two interacting cognitive systems: an intuitive, fast system and a reflective, effortful one [23,24]. Overcoming the nonscientific intuitive responses, which are observable outcomes that may arise from either heuristics or misconceptions [17], often requires inhibitory control, which is a component of the executive function responsible for suppressing prepotent but incorrect responses [25,26]. It plays a critical role in regulating cognitive interference during scientific reasoning by enabling learners to suppress heuristics or previously acquired misconceptions and engage in conceptually accurate interpretations [27]. It is also suggested that during conceptual conflict, the suppression of dominant intuitive representations allows students to redirect attention and cognitive resources toward correct scientific principles [16,28].

Despite extensive research on the involvement of inhibitory control in overcoming content-specific misconceptions, relatively few studies have examined how inhibitory control operates in procedural domains such as data evaluation and measurement reasoning, where intuitive heuristics are also pervasive [18,29–31]. Moreover, it remains unclear whether the activation of heuristics and the involvement of inhibitory control differ between different representation formats, such as numerical tables versus pictorial displays, as they may engage different cognitive processes and impose different cognitive loads. To address these gaps, the present study investigates whether and how

inhibitory control is involved in overcoming physics measurement-related heuristics or misconceptions during data quality assessment, taking the MMB as an example. We employed a modified Stroop paradigm [32,33] to examine students' cognitive conflict responses under different types of heuristic-scientific congruence. By quantifying response times (RTs) and accuracies (ACCs) across tabular and pictorial formats, this study aims to elucidate the inhibitory processes underlying conceptual interference and resolution in the context of data literacy in physics education.

II. LITERATURE REVIEW

A. Misconceptions and heuristics in evaluating data quality

Misconceptions in physics education have been extensively documented, particularly in core domains such as force, motion, and energy. For example, many students naively believe that an object at rest experiences no force or that heavier objects fall faster than lighter ones [34]. Similarly, the misconception that a constant force is needed to maintain constant velocity reflects a fundamental misunderstanding of Newtonian mechanics [35]. A more nuanced form of this reasoning, termed position-velocity indiscrimination (PVI), occurs when students incorrectly associate objects at the same position as having the same velocity, which is frequently triggered by visual representations of motion [19,20]. As Elby points out, such visual misinterpretations often stem from intuitive knowledge elements like the "what you see is what you get" (WYSIWYG), which leads learners to interpret graph features iconically (e.g., interpreting a flat velocity-time graph as stillness). These intuitive judgments, while context-sensitive and not necessarily "wrong" in isolation, can nonetheless coalesce into robust misconceptions when repeatedly activated inappropriately [36].

Beyond these content-specific misconceptions, students also exhibit procedure-oriented heuristics or misconceptions when evaluating empirical data [9]. Such evaluation often involves not only judging the adequacy and reliability of the data but also reasoning about measurement uncertainty, a key aspect where learners' difficulties become evident. Numerous studies have revealed a fundamental paradox: students tend to oscillate between two contrasting paradigms when interpreting empirical results (point paradigm vs set paradigm) [37,38]. The point paradigm reflects a belief that the measurement process can determine the true value of the measurand, viewing errors as potentially reducible to zero and treating single readings as possibly the true value. In contrast, the set paradigm assumes that measurements inherently contain irreducible uncertainties, requiring all available data to be analyzed as a distribution to approximate the measurand and quantify its uncertainty. Although instruction often promotes a shift toward setlike reasoning, students still struggle to appropriately account for uncertainty when drawing conclusions from measurement data [39].

A particularly robust example of such procedure-oriented difficulty is "the more measurements, the better" (MMB) heuristic [18]. Students tend to overgeneralize the productive insight that larger sample sizes can improve reliability, developing the belief that more measurements necessarily guarantee higher data quality, regardless of variability or systematic error. According to the intuitive rules theory [40], such a heuristic reflects a simplified experience-based rule of the form such as "more A, more B," which is efficient and frequently validated in everyday contexts but misleading in scientific ones. From the perspective of diSessa's phenomenological primitives (p-prims) (p. 52) [41], MMB aligns with the "more effort begets more results," a knowledge element derived from everyday experience that is overapplied in scientific contexts. Thus, the MMB reflects a cognitively appealing heuristic, rather than a mere factual error, and directly competes with formal statistical reasoning and normative principles emphasizing measurement uncertainty, experimental design quality, and statistical reliability.

Beyond MMB, additional procedure-oriented heuristics or misconceptions also indicate the dominance of intuitive reasoning in data evaluation [31,42–45]. For example, students may assume that high-precision instruments eliminate the need to consider uncertainty [46] or conflate decimal precision with measurement accuracy [47]. Others equate the mode (most frequent value) with the "true" measurement [37,48]. Even when students collect repeated data, they may ignore variability in favor of heuristic-based summary judgments [49–52].

Researchers attribute these persistent difficulties to students' failure to integrate statistical indicators into a coherent evaluative framework. Masnick and Morris argue that such fragmented reasoning reflects an inability to coordinate measures like the mean and standard deviation when making evaluative judgments [9]. Allie *et al.* further demonstrate that even students trained in setlike reasoning struggle to incorporate sample size (N) into reliability judgments, revealing a cognitive bottleneck in coordinating variability and N [48].

Building on these findings and aiming to better characterize students' reasoning about data quality, we propose an operational definition of data evaluation and measurement reasoning (DEMR) [9,18,37,38,48]. DEMR is defined as the capacity to make normative, context-sensitive judgments about data quality. It goes beyond mere statistical computation, emphasizing the reflective integration of multiple statistical indicators, and can be refined into three interrelated subcomponents: (i) central tendency appraisal [9,37], the ability to detect, compare, and interpret the mean values of datasets as indicators of central location; (ii) variability sensitivity [38,48], the ability to evaluate data spread

(e.g., variance, standard deviation, or range) and recognize that tighter clustering increases reliability; and (iii) coordinated reasoning about spread and sample size [18,48], the ability to incorporate the number of measurements (N) into judgments of data quality. Students should recognize that a larger sample size typically increases the reliability of statistical estimates, but only when interpreted alongside variability.

To systematically investigate how students engage with these types of heuristics or misconceptions, researchers have developed a number of diagnostic instruments. Tools such as the Physics Measurement Questionnaire (PMQ), the Concise Data Processing Assessment (CDPA), and the Physics Lab Inventory of Critical Thinking (PLIC) were designed to evaluate students' strategies for judging data quality, particularly when confronted with conflicting cues of precision, variability, or measurement count [48,53–55]. These tools consistently reveal the persistence of intuitive, rule-based strategies in students' thinking, particularly when surface features such as numerical precision are manipulated [9,18,47].

However, these tools primarily elicit students' procedural judgments, such as whether to take more measurements or improve equipment, under the assumption that such decisions reflect deeper conceptual understanding. Schang *et al.* argued that this approach does not fully capture students' underlying conceptions of data variability [18]. To address this limitation, they proposed a predictive format that directly asks students to anticipate what will happen to experimental data when sample size increases ("more data") or measurement quality improves ("better data"). This shift from procedural action to conceptual expectation provides a more direct lens into students' reasoning about variability and uncertainty.

These heuristics or misconceptions present a persistent challenge for instruction, as they have been proven to be resistant to explicit teaching interventions [43]. Recent conceptual change research suggests that learning involves negotiating between coexisting cognitive models, with heuristics or misconceptions competing but not passively replaced by scientific knowledge [21,56–58]. Heuristics or misconceptions that originate in everyday experience, i.e., the so-called "life-world domain," are particularly persistent and are not readily addressed through instruction [57,59]. Although much of this literature focuses on content-specific heuristics or misconceptions, similar cognitive mechanisms may appear to apply to procedure-oriented domains such as MMB. These heuristic-based misunderstandings are not just knowledge deficits; they function as intuitive default strategies that can be automatically triggered in scientific reasoning contexts.

Recent findings suggest that while scientific conceptions may prevail under certain conditions, they often fail to fully suppress entrenched misconceptions [60]. Therefore, rather than expecting heuristics or misconceptions to vanish, learners must engage in active cognitive regulation, notably by the inhibition of misleading intuitive responses, to apply normative scientific reasoning appropriately [21]. This points to the critical role of inhibitory control in conceptual change, particularly in situations that involve conflict between heuristics and scientifically grounded judgments. The next section explores this executive function in more detail, focusing on its role in suppressing heuristics or misconceptions and facilitating problem solving in physics education.

B. The role of inhibitory control in physics conceptual learning and problem solving

Inhibitory control, a core component of the executive function system, refers to the ability to suppress automatic, prepotent, or contextually inappropriate responses in favor of goal-directed behavior [26]. Within the broader framework of executive functions, alongside working memory and cognitive flexibility, inhibitory control allows individuals to maintain focus, resist distractions, and override intuitive but incorrect responses [61]. In science learning, particularly in physics, this capacity is crucial when students must reject appealing but incorrect reasoning patterns in favor of more abstract or counterintuitive scientific principles. As Heckler emphasized, many of these intuitive responses are not based on stable conceptions but arise from automatic, bottom-up perceptual processes that can mislead reasoning even in well-instructed individuals [16]. According to Houdé's [62] inhibitory control hypothesis, the acquisition of scientific concepts requires more than knowledge accumulation; it demands the active suppression of heuristics that conflict with normative reasoning. This view positions inhibition not as a peripheral support mechanism, but as a central agent in conceptual change and scientific reasoning.

This aligns with the dual-process theory [23,24], which suggests that human cognition operates through the interaction of two systems: system 1, which is fast, automatic, and intuitive; and system 2, which is slow, deliberate, and analytical. In physics learning, system 1 often activates misconceptions or heuristics, such as "heavier objects fall faster" or MMB, that can conflict with the formal knowledge managed by system 2. When such conflicts arise, learners must actively engage inhibitory control to suppress the default system 1 response and enable the application of system 2 reasoning [63]. Recent work in physics education by Kryjevskaia *et al.* [63], as well as Speirs *et al.* [64] and Lindsey *et al.* [65], has provided empirical support for this view, demonstrating how intuitive responses (system 1) must often be inhibited to allow for analytical reasoning in contexts involving air resistance, electric circuits, or graphical interpretation.

The role of inhibitory control in overcoming heuristics or misconceptions has been further exemplified in specific physics contexts [19,22,25]. One illustrative case is the PVI

misconception, where students incorrectly infer that objects at the same position must have the same velocity, showing how perceptual cues trigger misleading intuitive judgments [19,20]. Elby further emphasizes that such misinterpretations often arise from compelling visual attributes that disproportionately activate heuristics such as WYSIWYG, leading students to interpret graphical representations in a naively literal manner [36]. In such problems involving graphs or other motion representations, students often need to inhibit perceptually salient but incorrect interpretations to apply abstract physical laws correctly. These situations exemplify the role of inhibitory control, i.e., suppressing competing misconceptions and facilitating productive conceptual change [21,60].

The importance of inhibitory control in science learning has prompted researchers to adopt experimental paradigms from cognitive psychology to quantify its effects. Among these, the Stroop paradigm is one of the most widely used. In a classic Stroop task, participants are asked to name the ink color of a word that may itself denote a conflicting color (e.g., the word "blue" printed in red ink) [32,33,66]. Interference arises in incongruent trials, where the word meaning conflicts with the ink color, requiring participants to inhibit the dominant word-reading response [32]. Participants typically perform worse in incongruent than in congruent trials, showing longer RTs and/or lower ACCs. These two behavioral indicators, however, capture different aspects of conflict resolution: RT is generally more sensitive to the presence and degree of conflict, as many participants can still resolve the conflict correctly but require additional time to inhibit the dominant intuition response; ACC, by contrast, reflects the success or failure of this resolution, with lower ACC indicating that inhibitory control was insufficient and the automatic response dominated [33]. Thus, lower ACC implies stronger conflict or weaker inhibitory control, whereas longer RT indicates successful but effortful inhibition, reflecting the additional time required to override the prepotent heuristic response. Taken together, both RT and ACC are regarded as complementary behavioral indices of Stroop effects [67], as they jointly reflect the conflict triggered by an incongruent stimulus and the necessity of engaging inhibitory control to overcome it.

Modified Stroop paradigms have been adapted in educational research to investigate how learners suppress intuitive responses when solving mathematics and science problems [68–71]. For instance, numerical Stroop tasks have been employed to examine how students resolve conflicts between perceptual cues and semantic knowledge, such as judging the physical size of numerals like "3" and "5" when their numerical magnitude evokes conflicting intuitive responses [68]. Similarly, in science education, Babai and Amsterdamer demonstrated that tasks involving misconceptions, such as categorizing solids and liquids, elicited significantly longer reaction times and higher error rates in conflict conditions, implying a cognitive cost when learners must override naïve conceptions with formally learned scientific knowledge [70]. Potvin *et al.* further found that the reaction time increased proportionally with the strength of intuitive misconceptions, such as the belief that heavier objects sink more readily [71]. Collectively, these findings validate the utility of modified Stroop paradigms in capturing the competition between intuitive and analytic reasoning processes, shedding light on the role of inhibitory control in conceptual change within educational contexts.

Another relevant paradigm is the negative priming (NP) paradigm [72,73], in which a previously inhibited stimulus becomes the target in a subsequent task. Longer RTs in such cases suggest that prior inhibition remains active, affecting current cognitive performance. It offers a powerful tool for detecting how and when learners engage inhibitory control, particularly in conflict-rich problem-solving environments. Studies adopting NP paradigms across physics domains, such as kinematics, buoyancy, and projectile motion, have consistently shown that learners exhibit longer RTs when previously inhibited misconceptions are reactivated, indicating that conceptual change often requires the suppression of deeply rooted intuitive beliefs [19,74,75]. For example, Lin *et al.* employed an NP paradigm to investigate the PVI misconception in kinematics, where students incorrectly infer equal velocities from objects sharing the same position [19]. When students were primed with conflict-laden items (i.e., same position but different velocities), their performance on subsequent congruent probes significantly declined compared to neutral priming conditions, revealing a typical NP effect. This indicates that the suppression of intuitive but incorrect representations consumes cognitive resources and affects later task performance. Moreover, developmental studies have shown age-related variations in NP effects. Brault Foisy *et al.* demonstrated that children and adolescents exhibited stronger NP effects than adults when overcoming the misconception that "bigger objects sink more," suggesting that reliance on inhibitory mechanisms may decrease as scientific conceptions become more dominant through formal learning [27]. Taken together, these studies underscore the utility of the NP paradigm as a process-sensitive indicator of conceptual conflict resolution, highlighting its value in elucidating the developmental and situational dynamics of inhibitory control in science education.

Converging evidence from neuroimaging and educational neuroscience further supports these behavioral findings. Functional imaging studies have shown that solving problems involving misconception interference activates key prefrontal brain regions, including the anterior cingulate cortex, dorsolateral prefrontal cortex, and ventrolateral prefrontal cortex, which are widely associated with inhibitory control and conflict monitoring [76]. Moreover, expert

learners exhibit greater and more sustained activation in these regions compared to novices, suggesting that successful science learning is accompanied by the strengthening of inhibitory control [77,78]. Recent event-related potential (ERP) research has shed additional light on the temporal dynamics of these processes. For example, a study by Zhu *et al.* revealed that intuitive misconceptions are automatically activated at early processing stages to evaluate the plausibility of incoming information [79]. Only subsequently do learners retrieve the scientific knowledge, monitor the conflict, and engage inhibitory control to suppress the misconception. This finding supports a dynamic competition model: intuitive and scientific systems are activated sequentially, and the success of conceptual reasoning depends on the timely suppression of early-activated misconceptions. Together, this neurocognitive evidence reinforces the argument that inhibitory control is a necessary cognitive function for overcoming persistent misconceptions in science learning.

Despite mounting evidence for the role of inhibitory control in resolving content-specific heuristics or misconceptions, far less attention has been paid to procedural knowledge domains, such as data evaluation and measurement reasoning. Heuristics or misconceptions like MMB are deeply embedded in students' intuitive understanding of empirical work, and their persistence also suggests the involvement of inhibitory processes [18]. Unlike content-based heuristics or misconceptions, those related to procedural knowledge often entail statistical reasoning [44,48], which may impose greater demands on inhibitory control. However, procedural reasoning contexts, particularly those involving statistical judgments about data quality, remain understudied in the literature, despite their cognitive complexity. Addressing this gap is increasingly important as data literacy becomes a core objective in contemporary science education.

C. Inhibitory control and cognitive demands across different data representation formats

The format in which data is presented serves as more than a stylistic choice; it operates as a cognitive scaffold that fundamentally shapes how learners perceive, process, and reason with information. In data evaluation tasks, information is most commonly presented either in tabular form or as raw pictorial records (e.g., original experimental images) [51]. Different representational formats always engage distinct cognitive strategies and influence learners' reasoning in problem solving [80]. However, persistent heuristics or misconceptions may arise across these varied formats. For example, Nieminen *et al.* found that students often applied the same incorrect reasoning across isomorphic physics tasks, regardless of whether the problems were presented verbally, graphically, or vectorially [81]. This cross-representational consistency suggests that heuristics or misconceptions are robust and likely to be automatically activated, independent of the surface form.

However, the extent to which individuals inhibit these heuristics appears to be format-dependent, as suggested by several theoretical models. Dual coding theory distinguishes between the verbal-symbolic and visual-spatial channels, emphasizing that information can be processed through either or both channels simultaneously [82]. Building on this point, tabular representations are considered to primarily engage the verbal-symbolic channel because numerical information is processed sequentially and analytically [83,84]. In contrast, pictorial representations are thought to primarily activate the visual-spatial channel because patterns are recognized quickly through perceptual fluency [85–87]. These theoretical accounts converge with insights from dual-process theory, which implicitly posits that pictorial representations are more likely to activate fast, automatic system 1 reasoning, whereas symbolic representations recruit slow, reflective system 2 processes [23,24].

Building on this processing distinction, cognitive load theory suggests that such differences in processing modes may lead to differences in intrinsic cognitive load: tabular representations tend to impose greater intrinsic cognitive load because symbolic and sequential processing requires higher levels of abstraction, whereas pictorial representations impose comparatively lower intrinsic cognitive load due to their perceptual immediacy and holistic pattern recognition [88]. This higher cognitive load in tabular representations may also mask or "wash out" more subtle RT-based Stroop effects by prolonging baseline processing time, even when heuristic interference is present.

In parallel, perceptual load theory proposes that tasks with high perceptual complexity engage attentional capacity fully, thus reducing susceptibility to distraction—whereas low-load tasks leave attentional "spare capacity" vulnerable to interference [89]. For instance, pictorial representations may impose lower perceptual complexity, but their intuitive appeal makes learners more vulnerable to distraction. In contrast, tabular representations require more focused and sequential processing, potentially reducing learners' susceptibility to surface-level interference.

In sum, these points collectively imply that although heuristics or misconceptions are triggered across formats, the different formats would elicit varying cognitive loads, resulting in differences in the salience of heuristics and the involvement of inhibitory control when suppressing them. Specifically, pictorial representations may activate stronger heuristics or misconceptions and require a higher demand for inhibitory control. In contrast, tabular representations, while demanding more cognitive resources for symbolic and sequential processing, may elicit weaker heuristics activation and consequently impose lower inhibitory control demand.

Empirical evidence supports this differentiation from three aspects. First, numerous studies show that visually

salient but conceptually irrelevant cues interfere with student reasoning and require active suppression [19,58,75,76,90]. As mentioned previously, for example, Lin *et al.* found that perceptual features such as object position disrupted judgment in physics problem solving, underscoring the role of inhibitory control [19]. Second, the structure of the representation itself can modulate inhibition: Boyer *et al.* demonstrated that children exhibited a stronger whole number bias when proportions were presented as discrete units (e.g., blocks) rather than continuous bars, suggesting that certain visual formats intensify the need for suppression of misleading intuitions [91]. Third, there is growing evidence that cognitive and perceptual load interact dynamically with inhibition. Studies using NP paradigms [92,93] reveal that high-load environments impair inhibitory performance, as cognitive resources are diverted toward primary task demands. Recent findings by Yu *et al.* further confirm that inhibitory efficiency declines under high representational complexity, highlighting the resource-dependency of inhibition [94].

Taken together, these findings imply that while heuristics or misconceptions may be triggered across formats, the involvement of inhibitory control is format-sensitive. Representational forms would shape the salience of heuristics and the involvement of inhibitory control because they cause different cognitive loads. Therefore, it is essential to examine how specific heuristics like MMB interact with different representations, particularly tabular form versus pictorial form, and how these representations modulate inhibitory control involvement during scientific reasoning.

III. THE CURRENT STUDY

Despite growing interest in science education research on misconceptions and inhibitory control, most studies have focused on conceptual misunderstandings in content knowledge areas such as force, motion, or energy [19,34,40,71]. In contrast, procedural heuristics or misconceptions, particularly those involving data interpretation and measurement reasoning, have received limited attention. Among them, the MMB heuristic is particularly resilient because it conflates data quantity with data quality, causing learners to overlook variance, a key statistical indicator of reliability [9,18,47].

While prior studies have demonstrated the pervasiveness of MMB-like heuristics across educational levels and documented students' fragmented understanding of statistical reasoning [31,95], the cognitive mechanisms by which these heuristics or misconceptions are suppressed remain underexplored. Although inhibitory control, the ability to suppress intuitive but incorrect responses, has been widely studied in the context of physics conceptual change [19,22,25], its role in data evaluation tasks is still not well understood.

An additional gap concerns the role of data representation formats. Heuristics or misconceptions such as MMB are frequently elicited across both symbolic formats (e.g., numerical tables) and visual-spatial formats (e.g., pictorial representation or dot plots), yet it remains unclear whether suppressing these heuristics imposes equivalent inhibitory demands across representational forms. Theoretical perspectives from the dual coding theory, cognitive load theory, and perceptual load theory suggest that pictorial formats may have a lower cognitive load and amplify intuitive interference due to their visual salience, thus demanding more robust inhibitory engagement [23,24,82,88,89]. However, this prediction has not been empirically validated in the domain of data quality judgment, leaving open an important question regarding the interaction between representation formats and inhibitory control.

To address these gaps, the present study is guided by two research questions (RQs):

RQ1: Whether inhibitory control is involved in overcoming the MMB heuristic during data quality assessment tasks?

RQ2: If so, how do different data representation formats (tabular vs pictorial) affect the involvement of inhibitory control when suppressing the same heuristic?

To address these questions, we designed two behavioral experiments using a task adapted from the Physics Measurement Questionnaire (PMQ) [48] with a modified Stroop paradigm, based on our operational definition of DEMR. The task required participants to judge the relative quality of two datasets based on variance, presented in either tabular or pictorial form. Experiment 1 examined the MMB heuristic in tabular format (MMB-tabular) and experiment 2 explored it using pictorial (dot plot) format (MMB-pictorial). Based on the setting of the classic Stroop paradigm, the tasks in the two experiments consisted of three stimulus types: congruent (intuitive heuristic and scientific responses align), incongruent (intuitive heuristic and scientific responses conflict), and neutral (no intuitive heuristic cue). This enables precise comparisons of inhibitory demand based on the presence of intuitive-scientific conflict.

In the MMB-tabular task (experiment 1), participants compared two datasets that varied in the number of measurements and variance, while holding decimal precision and mean constant. Stimulus types were operationalized as neutral (same numbers of measurements, different variances), congruent (more measurements align with lower variance), and incongruent (more measurements align with higher variance). Accurate responses in incongruent trials required participants to override the MMB intuitive heuristic and instead base their judgment on statistical variance. Therefore, RTs are expected to be longest and ACC lowest in incongruent trials, reflecting the involvement of inhibitory control. Congruent trials should elicit the shortest RTs and highest ACCs, with neutral trials falling in between. In the MMB-pictorial task (experiment 2), the same logic applied, except that data were visualized as dot plots.

We explicitly operationalize the theoretical constructs of heuristic activation, cognitive conflict, and inhibitory control through two behavioral indicators: ACC and RT. Both ACC reductions and RT delays in incongruent trials are taken as evidence that the activation of heuristic response, the triggered cognitive conflict, and the involvement of inhibitory control [32,96,97]. ACC reductions indicate instances of failed inhibition, where the intuitive response intruded and dominated the judgment [33,96,98], whereas RT delays indicate successful but effortful inhibition, reflecting the additional processing time required for inhibitory control to override the prepotent heuristic response [96,97]. This mapping allows us to capture not only whether heuristic interference occurs but also how effectively and persistently inhibitory control is engaged.

Based on theoretical assumptions and prior findings, we formulated two hypotheses:

H1: Participants in the two experiments will exhibit longer RT and lower ACC under incongruent trials compared to congruent or neutral trials, reflecting the engagement of inhibitory control [69,70,99,100].

H2: Participants will perform different Stroop effects between the two experiments, indicating that the representation format will modulate the strength of the Stroop effect. Specifically, pictorial formats (i.e., dot plots) are expected to elicit stronger intuitive activation due to higher perceptual salience, thereby requiring greater inhibitory control and resulting in a larger Stroop effect, that is, larger differences in RT and ACC between incongruent and neutral/congruent trials [19,23]. In contrast, tabular formats impose greater cognitive load due to symbolic and sequential information processing demands, which may reduce the salience of intuitive interference and leave fewer resources available for active suppression. As a result, Stroop effects (i.e., the performance differences across trial types) are predicted to be smaller in tabular formats [88,89,94].

IV. METHODOLOGY

A. Participants

An *a priori* power analysis using G*Power 3.1.9.2 indicated that a minimum of 28 participants would be required to detect a medium within-subject effect size ($f = 0.25$) in a three-level repeated-measures design, with 80% power at a significance level of $\alpha = 0.05$. The final sample comprised 30 undergraduate physics majors (12 male; $M = 21.1$ years) from a university in southern China, thus exceeding the minimum requirement. All participants were native Chinese speakers and reported normal or corrected-to-normal vision. Because the modified Stroop tasks in the current study involved keyboard responses, handedness could systematically influence motor execution speed and thus RT measurements. To control this potential variability, all participants were right-handed. To further ensure that their performance reflected genuine reasoning rather than task misunderstanding, a structured practice phase was implemented before the formal trials (see Procedure). This phase ensured that participants clearly understood that judgments should be based on variance rather than surface-level cues like measurement count.

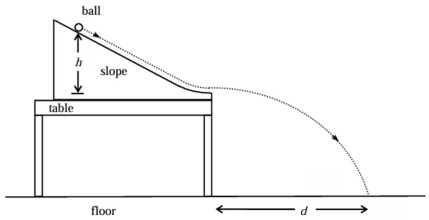
All participants had completed standard national physics and mathematics curricula before university. In grade 8, they learned fundamental statistical concepts (mean, median, mode, weighted mean, variance), while senior-high Mathematics (Compulsory Module 3) trained them to use samples to estimate population parameters, covering sample size, variance, standard deviation, range, and percentiles. At university, all had completed a 16-week General Physics Laboratory I course, which included a full module on Measurement Error and Data Processing. This module explicitly trained students to classify systematic vs random errors, compute standard deviations and uncertainties, and apply graphical and statistical data-analysis methods. Importantly, instructors repeatedly stressed that "larger sample sizes yield more reliable estimates," which may have unintentionally reinforced the MMB heuristic examined in this study.

In sum, participants were well prepared for statistical reasoning but also likely to hold the overgeneralized MMB heuristic, making them an appropriate group for investigating how inhibitory control regulates heuristic-driven interference in data evaluation tasks.

B. Materials

The task was adapted from a context involving horizontal projectile motion context in the Physics Measurement Questionnaire (PMQ) [48] (see Fig. 1), where participants were asked to evaluate which of two datasets demonstrated better quality (see Fig. 2).

Two experiments were developed using two kinds of data presentation formats: tabular (numeric) and pictorial (dot plot) formats. Specifically, experiment 1 tested the MMB heuristic using the tabular format of data, and experiment 2 used the pictorial format of data. Each experiment consisted of three stimulus types: neutral, in which intuitive cues were absent (e.g., same number of measurements), congruent trials, in which heuristics aligned with variance-based judgments (e.g., more measurements and lower variance), and incongruent trials, in which heuristics conflicted with variance-based judgments (e.g., more measurements and higher variance). Such settings are explicitly mapped onto the three DEMR subcomponents. Central tendency appraisal was controlled by holding the means of the two datasets equal on every trial, preventing choices based on average value differences. Variability sensitivity was targeted by manipulating the spread of the datasets, making variance the only valid diagnostic cue for data quality. Coordinated reasoning about spread and sample size was elicited by orthogonally manipulating measurement times (N) and variance across



As shown in the figure, students are conducting the following experiment in the laboratory:

1. Place an inclined plane on a horizontal table, ensuring the end of the plane is level with the table edge.

2. Release a small ball from rest at a fixed height *h* on the inclined plane, allowing it to roll down smoothly and leave the plane with a horizontal velocity.

3. Position a sheet of white paper overlaid with carbon paper (carbon side facing down) on the floor below. When the ball strikes the surface, it leaves an imprint marking the landing position.

4. After each trial, measure the horizontal distance *d* from the launch point (table edge) to the landing imprint.

After reading the instructions, press "Q" to continue.

FIG. 1. Projectile motion task context.

congruent and incongruent trials. This design compels participants to resolve conflicts between the MMB heuristic and normative statistical reasoning, thereby providing a process-sensitive measure of how well participants integrate the core DEMR subcomponents.

Each experiment consisted of 54 trials (18 per stimulus type). To balance potential bias caused by different sample size combinations, we systematically varied the number of measurements (i.e., 4, 5, or 6 data points) across trials. In the neutral trials, both datasets within a pair had the same number of data points (4, 5, or 6), and these three combinations were equally distributed. In the congruent and incongruent trials, three pairings were used and equally represented: (4 vs 5), (4 vs 6), and (5 vs 6). All datasets within a trial shared the same mean but differed in variance.

In both experiments, participants judged which dataset exhibited higher data quality, i.e., lower variance given equal means. In the MMB-tabular format, two columns of numerical data were labeled "F" (left) and "J" (right), matching the "F" and "J" keys for responses. In the MMB-pictorial format, dot plots were displayed above and below a central horizontal axis and labeled with upward ("↑") and downward ("↓") arrows, corresponding to the ↑ and ↓ keys. This consistent labeling minimized motor ambiguity and facilitated efficient responses. In the pictorial format, each dot's projection onto the horizontal axis represented a

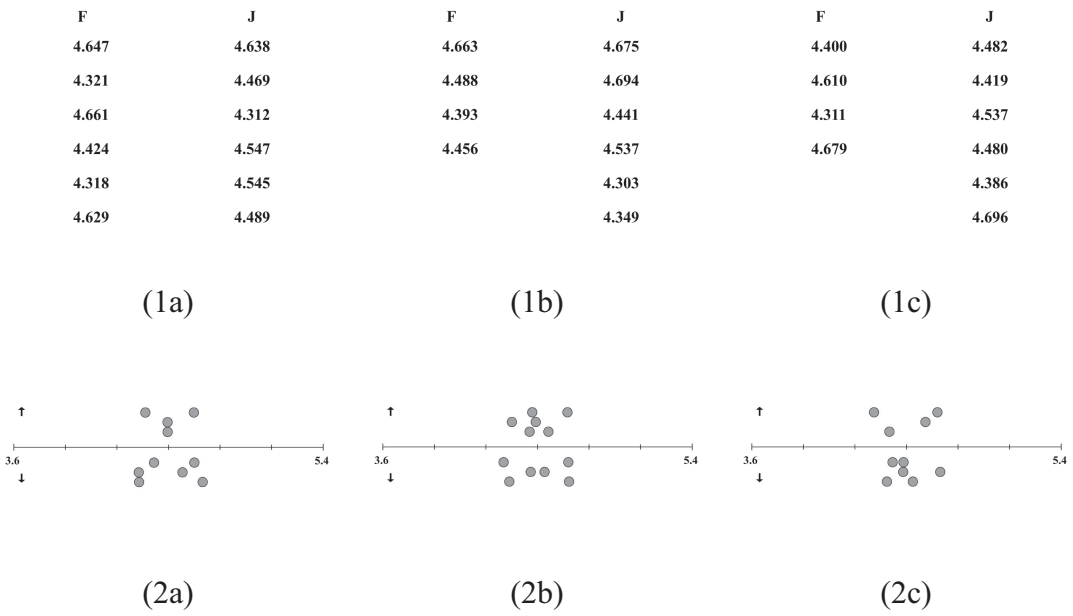


FIG. 2 Example stimuli from two experiments. Each row represents a particular experiment: (1) MMB with tabular format and (2) MMB with pictorial format. Columns (a)–(c) represent neutral, incongruent, and congruent trials, respectively. Participants judged which set—labeled "F" vs "J" or "↑" vs "↓"—had better data quality.

single data value, while the vertical spread, i.e., the distribution perpendicular to the axis, was visually arranged to simulate the natural randomness of projectile landing points and to minimize overlap between closely spaced data. This vertical visual arrangement was not intended to convey statistical dispersion (e.g., standard deviation) and was irrelevant to the judgment task. Step 4 of the task introduction (Fig. 1) explicitly informed participants that the measurement was the horizontal distance d from the launch point to the landing imprint, thereby implicitly indicating that vertical spread was not meaningful. Although the instruction did not explicitly ask participants to disregard vertical spread, postpractice interview data confirmed that they primarily focused on horizontal variability when making their judgments. Indeed, students' intuitive judgments may have been influenced by horizontal visual grouping, as indicated by the density of dots along the axis, which could serve as a potential heuristic cue.

To further support the appropriateness of the task as a measure of DEMR, we examined both its reliability and validity. Following methodological recommendations for reliability assessment of Stroop tasks [101], we estimated split-half reliability for our modified Stroop-based test. Reliability was evaluated separately for each condition (neutral, congruent, and incongruent) in both experiments. Trials were divided into first and second halves based on presentation order, and participants' mean RTs across halves were correlated, with the Spearman-Brown correction applied. The results indicated high internal consistency for condition-level RTs, with coefficients ranging from 0.88 to 0.96 (all $p < 0.001$), supporting the task's measurement stability.

To provide validity evidence, the task materials were reviewed by seven preservice physics teachers and two experienced physics educators. All reviewers agreed that the items appropriately assessed college-level understanding of measurement evaluation in physics experiments and correctly aligned with the targeted heuristic (MMB) and corresponding trial type (neutral, congruent, or incongruent). Furthermore, to provide response process validity evidence, postpractice interviews were conducted immediately after the practice session (see representative excerpts from students 1 and 2 in the Appendix). Participants confirmed that they interpreted the task as intended, i.e., initially relying on the MMB heuristic but shifting to variance-based reasoning, supporting that the task effectively elicited the intended construct.

C. Procedure

All participants completed the two experiments (tabular vs pictorial) in a counterbalanced order to control order effects. Each experiment was presented on a laptop computer with a 1920×1080 pixel resolution and a refresh rate of 60 Hz using E-prime 3.0 software (Psychology Software Tools, Inc., Pittsburgh, PA). To ensure that all the participants were fully acquainted with the task requirement, each experiment began with a dedicated practice session of 12 trials, comprising four trials for each of the three stimulus types (neutral, congruent, and incongruent), with each stimulus type presented twice in tabular format and twice in pictorial format. These practice trials were structurally identical to the formal trials but were excluded from subsequent analyses. After each practice trial, participants received immediate correctness feedback, and they could ask clarifying questions at any point during this phase. The formal session commenced only after participants completed the entire practice block and demonstrated a clear understanding of the task. This procedure minimized any confounding influence of task unfamiliarity, allowing the study to isolate data-evaluation reasoning processes.

Figure 3 illustrates the trial sequence, using the pictorial (dot plot) format from experiment 2 as a representative example. A fixation cross was presented at the center of the screen for 500 ms, followed by the stimulus display, which remained visible until a response was made. Participants were asked to judge which of the two datasets represented better data quality. Immediately after the response, a blank screen appeared for 800 ms before the onset of the next trial. No feedback was provided. The formal session in each experiment consisted of 54 trials: 18 neutral, 18 congruent, and 18 incongruent trials. They were presented in a randomized sequence to ensure that participants could not anticipate stimulus type or develop response heuristics.

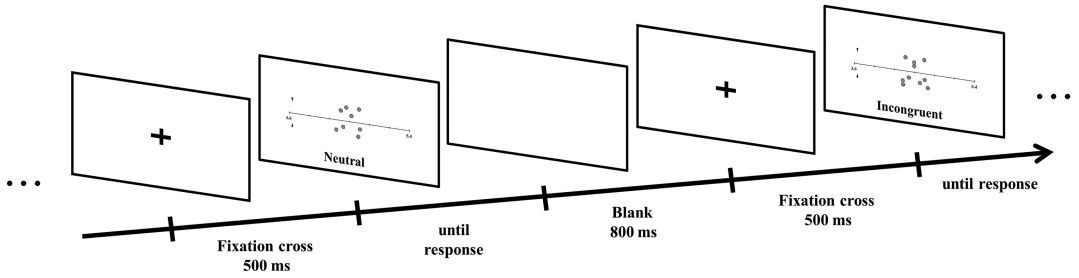


FIG. 3. Example trial sequence, shown here using the pictorial (dot plot) format from experiment 2. Each trial included a 500 ms fixation cross, a stimulus display (until response), and an 800 ms blank screen. Format-specific differences included layout and key mapping, but the sequence remained identical across experiments (see main text).

The main procedural difference across experiments lay in the data format and spatial arrangement of the stimuli. In experiment 1 (tabular format), two columns of numerical data were presented side-by-side, labeled "F" and "J." Participants responded using the left and right arrow keys to indicate whether the left or right group of data was of higher quality. In experiment 2 (pictorial format), two groups of dots representing projectile landing points were positioned symmetrically above and below a horizontal number axis. Here, participants responded with the up ("↑") and down ("↓") arrow keys to indicate which group (above or below the axis) showed better data quality.

D. Data analysis

All analyses were conducted using R (version 4.4.2), with the LME4 and lmerTest packages for mixed-effects modeling [102–104]. Two dependent variables were analyzed: ACC, defined as the proportion of correct responses and RT for correct trials only. Following the procedure outlined by Viarouge *et al.* [105], trial-level exclusions were performed within each participant for each experimental condition, defined by the combination of stimulus type (neutral, congruent, and incongruent) and presentation format (tabular or pictorial). Specifically, trials were excluded if they had incorrect responses (3.66% of all trials) or if their RTs deviated more than ±3 standard deviations from the participant's mean RT within each condition (0.88% of all trials). The average proportion of excluded trials per participant ranged from 0.79% to 9.52%. Since Shapiro-Wilk tests revealed that raw RT data were not normally distributed, a natural logarithm (ln) transformation was applied to RTs [106,107]. The transformed data met normality assumptions and were used in all subsequent analyses.

To appropriately analyze our within-subject, repeated-measures data, we used mixed-effects models instead of traditional ANOVAs. Traditional ANOVAs assume independence of observations [108], which is violated in our design, where multiple trials were nested within each participant. Moreover, by-participants or by-items ANOVAs cannot simultaneously account for errors arising from both participants and items and ignore individual differences in baseline tendencies (e.g., being generally faster or more accurate) or sensitivity to experimental manipulations. Mixed-effects models address these issues by incorporating both fixed and random effects, thereby controlling participant and item variability simultaneously. Random intercepts capture individual differences in baseline performance, and random slopes allow the effects of predictors to vary across participants, yielding more accurate and generalizable inferences, particularly for hierarchical data structures such as those in this study [109–112].

In both models, stimulus type (neutral, congruent, incongruent) and representation format (tabular vs pictorial) were entered as fixed effects. Stimulus type was dummy-coded with neutral as the reference level, and representation format was dummy-coded with tabular as the reference level. This model structure allowed us to test the H1 by assessing the main effect of stimulus type (Stroop effect) and the H2 by examining the interaction between representation format and stimulus type (representation modulation for Stroop effect).

We tested several candidate random-effects structures to account for individual variability and retained only those that significantly improved model fit while ensuring convergence. The final GLMM for ACC included a random intercept and slope for representation format by subject,

$$\mathrm{ACC} \sim 1 + \mathrm{format} + \mathrm{type} + \mathrm{format}:\mathrm{type} + (1 + \mathrm{format}|\mathrm{subject}) \quad (1)$$

Meanwhile, the final LMM for ln-transformed RTs included a random intercept and slope for both representation format and stimulus type by subject,

$$\ln(\mathrm{RT}) \sim 1 + \mathrm{format} + \mathrm{type} + \mathrm{format}:\mathrm{type} + (1 + \mathrm{format} + \mathrm{type}|\mathrm{subject}) \quad (2)$$

When the interaction effect is significant, estimated marginal means (EMMs) were calculated using the EMMEANS package to compare levels of stimulus type and their interaction with representation format [113]. Where applicable, Bonferroni-adjusted *post hoc* comparisons were applied.

V. RESULTS

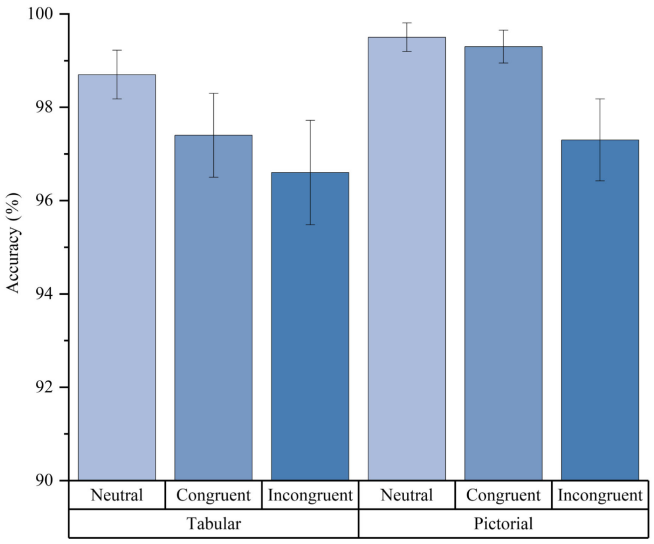
Table I and Fig. 4 present the mean ACCs and ln-transformed RTs [ln(RT)] across stimulus types (neutral, congruent, incongruent) and representation formats (tabular vs pictorial).

The GLMM for the ACC data revealed a significant main effect of stimulus type $[\chi^2(2) = 17.550, p < 0.001]$. Specifically, the ACC was significantly lower for incongruent trials (95.28%) than for neutral trials (98.43%), $\beta = -1.605$, $\mathrm{SE} = 0.556$, $Z = -2.887$, $p < 0.01$, and congruent trials (97.22%), $\beta = -1.379$, $\mathrm{SE} = 0.509$, $Z = -2.709$, $p < 0.01$. The difference between congruent and neutral trials was not significant, $\beta = -0.226$, $\mathrm{SE} = 0.676$, $Z = -0.335$, $p = 0.738$. Moreover, the main effect of representation format was not significant $[\chi^2(1) = 2.040$, $p = 0.153]$, nor was the stimulus type × representation format interaction $[\chi^2(2) = 3.830, p = 0.147]$.

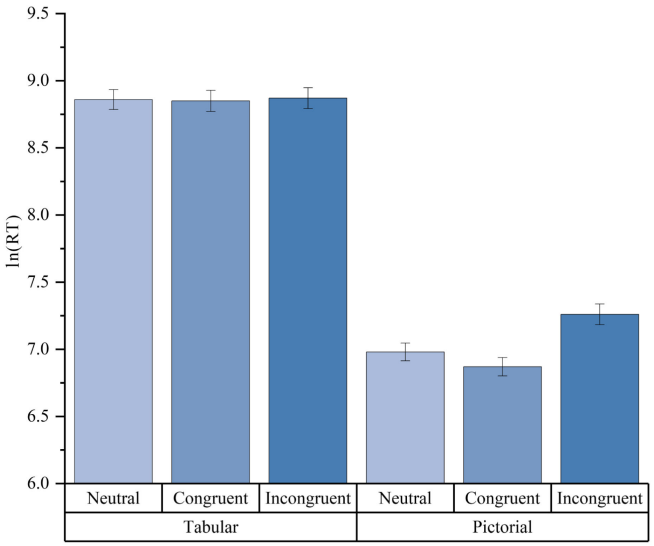
The LMM for the ln-transformed RTs revealed significant main effects of stimulus type $[F(2, 29) = 23.051$, $p < 0.001$, incongruent > neutral > congruent] and representation format, $[F(1, 29) = 639.206$, $p < .001$, tabular > pictorial], and a significant interaction between stimulus type and representation format $[F(2, 2994) = 46.458, p < 0.001]$. In the pictorial representation format,

TABLE I. Mean ACCs and ln(RT) by stimulus type and representation format (standard errors in parentheses).

Representation format	Stimulus type	ACC (%)	ln(RT)
Tabular	Neutral	97.59 (0.660)	8.86 (0.027)
	Congruent	95.37 (0.905)	8.85 (0.028)
	Incongruent	94.07 (1.017)	8.87 (0.026)
Pictorial	Neutral	99.26 (0.369)	6.98 (0.022)
	Congruent	99.07 (0.413)	6.87 (0.024)
	Incongruent	96.48 (0.794)	7.26 (0.029)
Overall	Neutral	98.43 (0.379)	7.92 (0.034)
	Congruent	97.22 (0.500)	7.85 (0.036)
	Incongruent	95.28 (0.646)	8.06 (0.032)



(a)



(b)

FIG. 4. Interaction between stimulus type and representation format on (a) ACC and (b) ln(RT).

Bonferroni-adjusted *post hoc* contrasts revealed that incongruent trials (7.26) were significantly slower than congruent (6.87), $\beta = 0.385$, SE $= 0.036$, $t(56) = 10.845$, $p < 0.001$ and neutral trials (6.98), $\beta = 0.277$, SE $= 0.034$, $t(64) = 8.140$, $p < 0.001$. Meanwhile, the neutral trials had significant longer RTs than congruent trials, $\beta = 0.108$, SE $= 0.029$, $t(97) = 3.744$, $p = 0.046$. In the tabular representation, although the same directional trend was observed (incongruent > neutral > congruent), none of the differences were statistically significant.

VI. DISCUSSION

A. The role of inhibitory control in overcoming the MMB heuristic in data quality assessment

This study provides direct evidence that inhibitory control plays a critical role in overriding the MMB heuristic during data evaluation tasks in physics. Consistent with H1, both ACC and RT data revealed significant main effects of stimulus type, indicating the presence of Stroop effects in the MMB context. Specifically, participants demonstrated significantly lower ACC and longer RT in the incongruent trials than in the congruent or neutral ones, suggesting that intuitive but incorrect heuristic responses were activated and required inhibition.

The Stroop effects observed in the current study support the dual-process account of thinking [23,24], wherein system 1 activates heuristics that must be inhibited, and hence system 2 is activated to enable normative reasoning. This is particularly salient in data judgment tasks where surface features, such as the number of measurements, can trigger heuristic-based judgments, as evidenced by reduced accuracy across both formats and extended response times in the pictorial format. These findings suggest cognitive conflict and the engagement of inhibitory control to suppress misleading intuitions [70,71].

These results align with prior research demonstrating that heuristics or misconceptions compete with scientific knowledge and persist even after instruction [21,60]. Importantly, these findings extend prior research into the domain of procedural reasoning, which has been comparatively

underexplored. Whereas most prior studies have focused on content knowledge (e.g., force or motion), our results show that heuristics in procedural contexts, such as the data quality evaluation used in this study, are equally prone to misjudgment and require inhibitory control for overcoming the heuristics interference. This reinforces the view that heuristic interference is a domain-general phenomenon in science learning and must be addressed even in tasks grounded in data literacy.

B. Representation format modulates demands in inhibiting the MMB heuristic

Another contribution of this study lies in demonstrating that the requirement for inhibitory control is not consistent across representation formats. The results revealed that the ACC was significantly reduced in the incongruent trials across the pictorial and tabular formats, confirming the heuristic interference in these two formats. However, a significant Stroop effect in RTs was observed only in the pictorial format, whereas no such effect emerged in the tabular format. This pattern partially supports H2's prediction that the pictorial format imposes greater inhibitory demands. This suggests that the visual data representations, while intuitively engaging, may also intensify cognitive conflict by increasing the salience of misleading perceptual cues [19,91]. As such, learners may require greater cognitive effort to suppress visually triggered heuristics like MMB. Importantly, although heuristic-based interference was evident in both formats, the extent to which each format activates intuitive responses remains an open question for future research.

Our findings on the different Stroop effects in RTs between the pictorial and tabular formats can be interpreted through an integrated framework combining dual coding theory [82], cognitive load theory [88], and perceptual load theory [89]. Although heuristic interference was present across both formats as reflected in ACC reductions in incongruent trials, the Stroop effect in RTs was specific to the pictorial format, suggesting differential cognitive dynamics underlying conflict detection and resolution. The tabular format primarily engages the verbal-symbolic system and imposes a higher intrinsic cognitive load due to its analytical demands. According to perceptual load theory, this high load is assumed to saturate attentional resources, which may in turn reduce (but not necessarily eliminate) susceptibility to heuristic interference by narrowing perceptual focus [80]. However, as evidenced by the lack of an RT Stroop effect, this reduced susceptibility might also be explained by a masking mechanism. That is, the higher cognitive processing demands and longer overall processing time in the tabular format may wash out subtle conflict effects in RT (i.e., a "washed-out" Stroop effect), even though heuristic interference, as shown by ACC, was clearly present. Taken together, these complementary explanations may jointly account for the absence of a significant RT Stroop effect in tabular format despite clear heuristic activation.

In contrast, the pictorial format (e.g., dot plots) primarily recruits the visual-spatial channel, enabling rapid pattern recognition and reducing symbolic abstraction. This lowers cognitive load and facilitates faster initial data access. Postpractice interviews provided qualitative support for this interpretation (see representative excerpts from students 3 and 4 in the Appendix). These participants reported that the pictorial format allowed for immediate visual comparison of data spread, whereas the tabular format required mentally translating numbers into visual patterns or engaging in more analytic processing, which was more time-consuming and effortful. However, the perceptual salience of visual features, such as the number of dots, may inadvertently trigger heuristic responses (e.g., MMB), increasing the risk of heuristic-based responses. The lower perceptual and cognitive load in this format may leave excess attentional resources available to process these misleading cues. This stronger activation of heuristic responses may also contribute to the pronounced RT Stroop effect observed in pictorial representations. Thus, while pictorial representations promote fluency and accessibility, they may simultaneously demand greater inhibitory control to override erroneous intuitions. This highlights a critical trade-off: ease of access through visual design may come at the cost of increased vulnerability to heuristic interference.

In addition, the two formats may differ in their task difficulty, which could also help explain the observed patterns. Dot plots may reduce task difficulty by affording perceptual immediacy and visual fluency in detecting variance, whereas tables may increase difficulty by requiring symbolic abstraction and sequential, analytic processing. This task difficulty contrast could affect the Stroop effects beyond inhibitory control dynamics: easier pictorial items may leave sufficient attentional resources to amplify conflict effects in RTs, whereas more difficult tabular items may prolong overall RTs and mask subtle interference. Thus, the absence of an RT Stroop effect in the tabular condition does not necessarily indicate a lack of heuristic activation, but rather, may reflect difficulty-related masking.

C. Implications for conceptual change and physics instruction

The present findings reinforce conceptual change models that emphasize the coexistence of scientific and intuitive misconceptions, rather than simple replacement [58,76,90,114]. The need for active inhibition of heuristic interference during correct responding, evidenced by significantly longer RT in incongruent trials in the pictorial format and significantly lower ACC in incongruent trials across both formats, supports theories viewing learning as a dynamic regulation of competing knowledge systems [57,58].

Importantly, this study extends the scope of inhibitory control research beyond content-based misconceptions,

such as those in electricity [76,79] or mechanics [25], to procedural misconceptions related to scientific reasoning, particularly the evaluation of data quality (i.e., MMB). To our knowledge, this is among the first empirical investigations to examine how inhibitory control is involved in conflict resolution in data assessment tasks. This highlights that procedural domains are, in fact, deeply shaped by intuitive interference and executive function regulation.

These insights carry direct implications for instructional design in physics education. Pictorial representations, as evidenced by the pronounced RT Stroop effect, may trigger stronger or more persistent heuristic interference and conflict and thus greater inhibitory demands. Paradoxically, this may be pedagogically advantageous: such formats can surface students' misconceptions, making them explicit and therefore more amenable to conceptual change. Instructional strategies could intentionally harness these intuitive pitfalls as "productive errors," creating moments of cognitive conflict that drive metacognitive reflection and, ultimately, more durable conceptual change [58,115,116].

While the pictorial format reduces processing effort, it may simultaneously increase the salience and accessibility of heuristic features. Conversely, the tabular format imposes higher representational cognitive load, as evidenced by the generally long RTs observed in all tabular format trials, and this increased demand, as suggested by cognitive load theory [88] and perceptual load theory [89], may help reduce but not entirely prevent learners' susceptibility to intuitive distractions by narrowing their perceptual focus. Therefore, a strategic balance between representational cognitive load and inhibitory accessibility is vital. Educators must design tasks that preserve enough cognitive resources for conflict monitoring and resolution, especially when introducing high-intuition formats.

Moreover, instruction should be sequenced to first activate intuitive conflict and then guide learners toward reflective resolution. Our findings suggest that cognitive load levels modulate conflict detection: lower cognitive load (pictorial format) may leave more attentional resources available to process misleading cues, whereas higher load (tabular format) may either reduce such processing or mask RT-based conflict effects. Thus, inhibitory success depends on how cognitive resources are allocated among decoding, attention control, and conflict suppression. Pedagogically, this suggests the need for instructional scaffolds that first activate intuitive conflict and then guide learners through reflective resolution, for instance, via representational switching, or inhibitory prompts [117].

Finally, efforts to strengthen learners' executive functions, particularly inhibitory control, could enhance their ability to resist intuitive interference. Recent studies show that inhibitory control can be effectively strengthened through domain-specific, classroom-embedded interventions, offering a promising pathway for integrating cognitive training into physics education [117,118].

D. Limitations and future directions

Several limitations should be acknowledged. First, the sample was limited to 30 physics undergraduates from a single university, which may limit generalizability. Replication with larger and more diverse populations, including preservice teachers or general science learners, is warranted.

Second, regarding validity, the present study provides only initial construct validity evidence. Content validity was supported through expert review, and response process validity was indicated by post-practice interviews confirming that participants understood and engaged with the task as intended. Nevertheless, the validity argument remains preliminary. Future research should extend this work by employing convergent validation approaches, such as correlating Stroop-task performance with external measures of statistical reasoning or testing the stability of the task across different populations.

Third, the tabular format yielded weaker or insignificant Stroop effect in RTs. Its higher cognitive load may obscure the subtle inhibitory mechanism, suggesting the need for task designs that more clearly dissociate processing effort from inhibition cost. Meanwhile, although the tabular format was intended to primarily elicit symbolic processing, we acknowledge that it also contains visual features such as column height, which may act as perceptual cues for some participants. The longer RTs observed across tabular trials suggest that participants generally engaged in analytical reasoning beyond these cues. However, future research should explicitly manipulate and control such visual features to clarify their influence on heuristic activation.

Fourth, in addition to format-related factors, another limitation concerns the potential confounding between representation format and item difficulty. While our analyses compared pictorial (dot plots) and tabular formats, it is possible that task difficulty differences also contributed to the observed Stroop patterns. Future studies should consider the potential confounding factor of task difficulty across different formats when interpreting the findings.

Fifth, while the Stroop effects were used as behavioral proxies for inhibitory control, future work should integrate the neuroscience techniques, such as electroencephalography (EEG), event-related potentials (ERP) [79], and functional magnetic resonance imaging (fMRI) [25], to provide converging evidence on the neural and temporal dynamics of how inhibitory control functions in heuristics.

Finally, this study only focused on the MMB heuristic. Future research should examine other data-related heuristics or misconceptions and their interaction with different instructional strategies or assessment modalities.

VII. CONCLUSION

This study examined how inhibitory control contributes to overcoming the MMB heuristic, a common belief in physics data evaluation that assumes more measurements imply higher data quality. Using a modified Stroop paradigm across two representation formats, we found that the incongruent trials, which required participants to override intuitive but incorrect responses, elicited significantly reduced ACC in both formats, and significantly longer RT in the pictorial format, with a similar trend in the tabular format. These results suggest that inhibitory control is engaged when individuals confront conflicts between MMB heuristic and scientific reasoning.

In addition, the findings revealed that the representation format modulates the involvement of inhibitory control when suppressing the MMB heuristic. Pictorial formats, such as dot plots, may evoke stronger or more persistent heuristic conflict, as evidenced by the pronounced RT Stroop effect, whereas tabular formats, despite also eliciting heuristic interference as shown by accuracy reductions, imposed greater symbolic processing demands that may reduce or mask the Stroop effects in RT. This interaction highlights the need to consider representational features when designing instructional materials intended to target intuitive heuristic or misconception-driven reasoning.

Taken together, the results support the view of conceptual change as involving competition between coexisting knowledge systems and extending this framework to the domain of procedural reasoning in data analysis, a relatively underexplored area in physics education research. Future work may explore how instructional interventions that scaffold inhibitory engagement can support students' ability to reason scientifically with data.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China under Grant 62407015, by construction funding from the Guangdong Basic Research Center of Excellence for Structure and Fundamental Interactions of Matter, and by the 2025 Guangzhou Education Science Planning Project Provincial Teaching Achievement Cultivation Project titled "Exploration and Practice of a Physics Micro Project Teaching Model that Engages Students in Creative Learning Activities" under Grant 2024111484.

Any opinion expressed in this work is that of the authors and does not necessarily represent those of the funding agencies.

DATA AVAILABILITY

The data are not publicly available. The data are available from the authors upon reasonable.

APPENDIX: REPRESENTATIVE POSTPRACTICE INTERVIEW EXCERPTS

Interviews were conducted immediately after participants completed the practice phase. All interviews were conducted individually.

Student 1:

Interviewer:	What do you think should be the basis for judging data quality in the experiment?
Response:	The degree of dispersion. I would notice the difference in data quantity, and although more data is generally better in real experiments, if the goal is just to judge which data is better, we should focus on the degree of dispersion.
Interviewer:	Does the amount of data affect your judgment?
Response:	At first, I picked the group with more data, but then I realized I was wrong and switched to the more concentrated one.

Student 2:

Interviewer:	What do you focus on when judging data quality?
Response:	I mainly look at the degree of dispersion. If two groups have similar dispersion, then I'll consider quantity. If a smaller group is more concentrated, I would choose it.
Interviewer:	Did the number of data points affect your initial judgment?
Response:	Yes, during the practice I intuitively thought more data meant better quality. But gradually I started paying attention to the fluctuations and learned how to judge better.

Student 3:

Interviewer:	What differences did you notice when doing the tasks with the two representations?
Response:	The pictorial representation felt much easier and faster because I could immediately see which group was more dispersed

Student 4:

Interviewer:	How did the formats affect your performance?
Response:	The table task was much more time-consuming because I processed the numbers directly without turning them into an image in my mind. I felt like I had to go through each number analytically, which slowed me down. The pictorial representation, in contrast, allowed me to make decisions much more quickly by just looking at the overall spread of the dots.

[1] T. Koltay, Data literacy: In search of a name and identity, J. Doc. **71**, 401 (2015).

[2] J. B., Data literacy for educators: Making it count in teacher preparation and practice, Harv. Educ. Rev. **86**, 607 (2016).

[3] J. Calzada Prado and M. Á. Marzal, Incorporating data literacy into information literacy programs: Core competencies and contents, Libri Oncol. **63**, 123 (2013).

[4] NGSS Lead States, *Next Generation Science Standards: For States, By States* (National Academies Press, Washington, DC, 2013).

[5] The National Governors Association Center for Best Practices and The Council of Chief State School Officers, *Common Core State Standards for Mathematics* (NGA Center & CCSSO, Washington, DC, 2010).

[6] Ministry of Education of the People's Republic of China, *General Senior Secondary Physics Curriculum Standards* (People's Education Press, Beijing, 2020).

[7] Ministry of Education Higher Education University Physics Curriculum Teaching Steering Committee, *Basic Requirements for University Physics Courses*, 2023 ed. (Higher Education Press, Beijing, 2023).

[8] C. Maybee and L. Zilinski, Data informed learning: A next phase data literacy framework for higher education, Proc. Assoc. Inf. Sci. Technol. **52**, 1 (2015).

[9] A. M. Masnick and B. J. Morris, Investigating the development of data evaluation: The role of data characteristics, Child Dev. **79**, 1032 (2008).

[10] D. J. Merritt, Bias, the brain, and student evaluations of teaching, SSRN J. (2007), 10.2139/ssrn.963196.

[11] A. Kirk-Johnson, B. M. Galla, and S. H. Fraundorf, Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice, Cognit. Psychol. **115**, 101237 (2019).

[12] M. M. Zucchelli, N. Matteucci Armandi Avogli Trotti, A. Pavan, L. Piccardi, and R. Nori, The dual process model: The effect of cognitive load on the ascription of intentionality, Front. Psychol. **16**, 1451590 (2025).

[13] D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York, 2011).

[14] A. Tversky and D. Kahneman, Judgment under uncertainty: Heuristics and biases, Science **185**, 1124 (1974).

[15] P. M. Todd and G. Gigerenzer, Précis of simple heuristics that make us smart, Behav. Brain Sci. **23**, 727 (2000).

[16] A. F. Heckler, The role of automatic, bottom-up processes: In the ubiquitous patterns of incorrect answers to science questions, Psychol. Learn. Motiv. **55**, 227 (2011).

[17] D. Hammer, More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research, Am. J. Phys. **64**, 1316 (1996).

[18] A. Schang, M. Dew, E. M. Stump, N. G. Holmes, and G. Passante, New perspectives on student reasoning about measurement uncertainty: More or better data, Phys. Rev. Phys. Educ. Res. **19**, 020105 (2023).

[19] J. Lin, Y. Xing, Y. Hu, J. Zhang, L. Bao, K. Luo, K. Yu, and Y. Xiao, Inhibitory control involvement in overcoming the position-velocity indiscrimination misconception among college physics majors, Phys. Rev. Phys. Educ. Res. **19**, 010112 (2023).

[20] D. E. Trowbridge and L. C. McDermott, Investigation of student understanding of the concept of velocity in one dimension, Am. J. Phys. **48**, 1020 (1980).

[21] C. Dawson, Towards a conceptual profile: Rethinking conceptual mediation in the light of recent cognitive and neuroscientific findings, Res. Sci. Educ. **44**, 389 (2014).

[22] L. Mason and S. Zaccoletti, Inhibition and conceptual learning in science: A review of studies, Educ. Psychol. Rev. **33**, 181 (2021).

[23] J. St. B. T. Evans, Dual-processing accounts of reasoning, judgment, and social cognition, Annu. Rev. Psychol. **59**, 255 (2008).

[24] D. Kahneman and S. Frederick, Representativeness revisited: Attribute substitution in intuitive judgment, in *Heuristics and Biases*, edited by T. Gilovich, D. Griffin, and D. Kahneman, 1st ed. (Cambridge University Press, Cambridge, England, 2002), pp. 49–81.

[25] L.-M. Brault Foisy, P. Potvin, M. Riopel, and S. Masson, Is inhibition involved in overcoming a common physics misconception in mechanics?, Trends Neurosci. Educ. **4**, 26 (2015).

[26] A. Diamond, Executive functions, Annu. Rev. Psychol. **64**, 135 (2013).

[27] G. Allaire-Duquette, L.-M. Brault Foisy, P. Potvin, M. Riopel, M. Larose, and S. Masson, An fMRI study of scientists with a Ph.D. in physics confronted with naive ideas in science, npj Sci. Learn. **6**, 11 (2021).

[28] J. De Vries, M. Byrne, and E. Kehoe, Cognitive dissonance induction in everyday life: An fMRI study, Soc. Neurosci. **10**, 268 (2015).

[29] D. Ben-Zvi, Reasoning about variability in comparing distributions, SERJ **3**, 42 (2004).

[30] D. Canada, Elementary pre-service teachers' conceptions of variation in a probability context, SERJ **5**, 36 (2006).

[31] J. Garfield and D. Ben-Zvi, How students learn statistics revisited: A current review of research on teaching and learning statistics, Int. Stat. Rev. **75**, 372 (2007).

[32] J. R. Stroop, Studies of interference in serial verbal reactions, J. Exp. Psychol. **18**, 643 (1935).

[33] C. M. MacLeod, Half a century of research on the Stroop effect: An integrative review, Psychol. Bull. **109**, 163 (1991).

[34] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[35] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics Education Research, Am. J. Phys. **67**, 755 (1999).

[36] A. Elby, What students' learning of representations tells us about constructivism, J. Math. Behav. **19**, 481 (2000).

[37] F. Lubben and R. Millar, Children's ideas about the reliability of experimental data, Int. J. Sci. Educ. **18**, 955 (1996).

[38] N. A. Obrecht, G. B. Chapman, and R. Gelman, Intuitive t tests: Lay use of statistical information, Psychon. Bull. Rev. **14**, 1147 (2007).

[39] T. Wan and J. M. Mickelsen, Investigating student ability to draw conclusions from measurement data, presented at PER Conf. 2012, Virtual Conference, 10.1119/perc.2021.pr.Wan.
[40] R. Stavy and D. Tirosh, *How Students (Mis-)understand Science and Mathematics: Intuitive Rules* (Teachers College Press, New York, 2000).
[41] A. A. diSessa, Knowledge in pieces, in *Constructivism in the Computer Age*, edited by G. Forman and P. B. Pufall (Lawrence Erlbaum Publishers, Hillsdale, NJ, 1988), pp. 49–70.
[42] P. Van Kampen and O. Gkioka, Undergraduate students' reasoning about the quality of experimental measurements of covarying secondary data, Eur. J. Phys. **42**, 045704 (2021).
[43] R. L. Kung and C. Linder, University students' ideas about data processing and data comparison in a physics laboratory course, NorDiNa **2**, 40 (2012).
[44] A. Buffler, S. Allie, and F. Lubben, The development of first-year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).
[45] M. Séré, R. Journeaux, and C. Larcher, Learning the statistical analysis of measurement errors, Int. J. Sci. Educ. **15**, 427 (1993).
[46] S. Pillay, A. Buffler, F. Lubben, and S. Allie, Effectiveness of a GUM-compliant course for teaching measurement in the introductory physics laboratory, Eur. J. Phys. **29**, 647 (2008).
[47] K. Kok, B. Priemer, W. Musold, and A. Masnick, Students' conclusions from measurement data: The more decimal places, the better?, Phys. Rev. Phys. Educ. Res. **15**, 010103 (2019).
[48] S. Allie, A. Buffler, B. Campbell, and F. Lubben, First-year physics students' perceptions of the quality of experimental measurements, Int. J. Sci. Educ. **20**, 447 (1998).
[49] T. Wan, Investigating student reasoning about measurement uncertainty and ability to draw conclusions from measurement data in inquiry-based university physics labs, Int. J. Sci. Educ. **45**, 223 (2023).
[50] C. F. J. Pols, P. J. J. M. Dekkers, and M. J. De Vries, What do they know? Investigating students' ability to analyze experimental data in secondary physics education, Int. J. Sci. Educ. **43**, 274 (2021).
[51] A. Susac, A. Bubic, P. Martinjak, M. Planinic, and M. Palmovic, Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study, Phys. Rev. Phys. Educ. Res. **13**, 020125 (2017).
[52] B. Priemer, S. Pfeiler, and T. Ludwig, Firsthand or secondhand data in school labs: It does not make a difference, Phys. Rev. Phys. Educ. Res. **16**, 013102 (2020).
[53] J. Day and D. Bonn, Development of the concise data processing assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).
[54] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking (PLIC), Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).
[55] F. Lubben, B. Campbell, A. Buffler, and S. Allie, Point and set reasoning in practical science measurement by entering university freshmen, Sci. Educ. **85**, 311 (2001).
[56] R. Duit and D. F. Treagust, Conceptual change: A powerful framework for improving science teaching and learning, Int. J. Sci. Educ. **25**, 671 (2003).
[57] S. Ohlsson, Resubsumption: A possible mechanism for conceptual change and belief revision, Educ. Psychol. **44**, 20 (2009).
[58] P. Potvin, Proposition for improving the classical models of conceptual change based on neuroeducational evidence: Conceptual prevalence, Neuroeducation **2**, 16 (2013).
[59] J. Solomon, Learning about energy: How pupils think in two domains, Eur. J. Sci. Educ. **5**, 49 (1983).
[60] P. Potvin and G. Cyr, Toward a durable prevalence of scientific conceptions: Tracking the effects of two interfering misconceptions about buoyancy from preschoolers to science teachers, J. Res. Sci. Teach. **54**, 1121 (2017).
[61] S. Lemire-Rodger, J. Lam, J. D. Viviano, W. D. Stevens, R. N. Spreng, and G. R. Turner, Inhibit, switch, and update: A within-subject fMRI investigation of executive control, Neuropsychologia **132**, 107134 (2019).
[62] O. Houdé, Inhibition and cognitive development: Object, number, categorization, and reasoning, Cognit. Dev. **15**, 63 (2000).
[63] M. Kryjevskaia, P. R. L. Heron, and A. F. Heckler, Intuitive or rational? Students and experts need to be both, Phys. Today **74**, No. 8, 28 (2021).
[64] J. C. Speirs, M. R. Stetzer, B. A. Lindsey, and M. Kryjevskaia, Exploring and supporting student reasoning in physics by leveraging dual-process theories of reasoning and decision making, Phys. Rev. Phys. Educ. Res. **17**, 020137 (2021).
[65] B. A. Lindsey, A. Boudreaux, D. J. Rosen, M. R. Stetzer, and M. Kryjevskaia, Reinforcing mindware or supporting cognitive reflection: Testing two strategies for addressing a persistent learning challenge in the context of air resistance, Phys. Rev. Phys. Educ. Res. **20**, 020116 (2024).
[66] C. M. MacLeod, The Stroop effect, in *Encyclopedia of Color Science and Technology*, edited by R. Luo (Springer, Berlin, Heidelberg, 2015), pp. 1–6.
[67] J. M. Bugg, L. L. Jacoby, and J. P. Toth, Multiple levels of control in the Stroop task, Mem. Cognit. **36**, 1484 (2008).
[68] S. Bugden, G. R. Price, D. A. McLean, and D. Ansari, The role of the left intraparietal sulcus in the relationship between symbolic number processing and children's arithmetic competence, Dev. Cognit. Neurosci. **2**, 448 (2012).
[69] R. Babai, R. Sekal, and R. Stavy, Persistence of the intuitive conception of living things in adolescence, J. Sci. Educ. Technol. **19**, 20 (2010).
[70] R. Babai and A. Amsterdamer, The persistence of solid and liquid naive conceptions: A reaction time study, J. Sci. Educ. Technol. **17**, 553 (2008).
[71] P. Potvin, S. Masson, S. Lafortune, and G. Cyr, Persistence of the intuitive conception that heavier objects sink

more: A reaction time study with different levels of interference, Int. J. Sci. Math. Educ. **13**, 21 (2015).
[72] W. T. Neill, Inhibitory and facilitatory processes in selective attention, J. Exp. Psychol. **3**, 444 (1977).
[73] S. P. Tipper, The negative priming effect: Inhibitory priming by ignored objects, Q. J. Exp. Psychol. A **37**, 571 (1985).
[74] L.-M. Brault Foisy, E. Ahr, J. Blanchette Sarrasin, P. Potvin, O. Houdé, S. Masson, and G. Borst, Inhibitory control and the understanding of buoyancy from childhood to adulthood, J. Exp. Child Psychol. **208**, 105155 (2021).
[75] Y. Wen, J. Lin, Y. Ming, J. Zhang, X. Wu, L. Bao, K. Yu, and Y. Xiao, Role of inhibition in overcoming interferences of misconception under similar feature saliency: An eye-tracking study of the projectile motion problem, Phys. Rev. Phys. Educ. Res. **20**, 020121 (2024).
[76] S. Masson, P. Potvin, M. Riopel, and L. B. Foisy, Differences in brain activation between novices and experts in science during a task involving a common misconception in electricity, Mind Brain Educ. **8**, 44 (2014).
[77] A. A. diSessa and B. L. Sherin, What changes in conceptual change?, Int. J. Sci. Educ. **20**, 1155 (1998).
[78] P. Potvin *et al.*, Models of conceptual change in science learning: Establishing an exhaustive inventory based on support given by articles published in major journals, Stud. Sci. Educ. **56**, 157 (2020).
[79] Y. Zhu, L. Zhang, Y. Leng, R. Pang, and X. Wang, Event-related potential evidence for persistence of an intuitive misconception about electricity, Mind Brain Educ. **13**, 80 (2019).
[80] A. Lichtenberger, C. Wagner, S. I. Hofer, E. Stern, and A. Vaterlaus, Validation and structural analysis of the kinematics concept test, Phys. Rev. Phys. Educ. Res. **13**, 010115 (2017).
[81] P. Nieminen, A. Savinainen, and J. Viiri, Force Concept Inventory-based multiple-choice test for investigating students' representational consistency, Phys. Rev. ST Phys. Educ. Res. **6**, 020109 (2010).
[82] A. Paivio, *Imagery and Verbal Processes*, 1st ed. (Psychology Press, New York, 1979).
[83] J. H. Larkin and H. A. Simon, Why a diagram is (sometimes) worth ten thousand words, Cogn. Sci. **11**, 65 (1987).
[84] P. Shah and J. Hoeffner, Review of graph comprehension research: Implications for instruction, Educ. Psychol. Rev. **14**, 47 (2002).
[85] A. Paivio, Empirical case for dual coding, in *Imagery, Memory and Cognition*, edited by edited by R. D. Pearson and J. T. E. Richardson (Psychology Press, London, 1980).
[86] M. Hegarty, The cognitive science of visual-spatial displays: Implications for design, Top. Cognit. Sci. **3**, 446 (2011).
[87] J. A. Fugelsang and V. A. Thompson, A dual-process model of belief and evidence interactions in causal reasoning, Mem. Cogn. **31**, 800 (2003).
[88] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. **12**, 257 (1988).
[89] N. Lavie and Y. Tsal, Perceptual load as a major determinant of the locus of selection in visual attention, Percept. Psychophys. **56**, 183 (1994).
[90] P. Potvin, G. Malenfant-Robichaud, C. Cormier, and S. Masson, Coexistence of misconceptions and scientific conceptions in chemistry professors: A mental chronometry and fMRI study, Front. Educ. **5**, 542458 (2020).
[91] T. W. Boyer, S. C. Levine, and J. Huttenlocher, Development of proportional reasoning: Where young children go wrong, Dev. Psychol. **44**, 1478 (2008).
[92] N. Lavie and E. Fox, The role of perceptual load in negative priming, J. Exp. Psychol. **26**, 1038 (2000).
[93] J. J. Ortells, C. Noguera, D. Álvarez, E. Carmona, and G. Houghton, Individual differences in working memory capacity modulates semantic negative priming from single prime words, Front. Psychol. **7** (2016).
[94] X. Yu, H. Zhang, and Y. Qi, The role of inhibitory control in the development of analogical mapping: Cognitive load matters, Psychol. Dev. Educ. **39**, 1 (2023), https://devpsy.bnu.edu.cn/CN/Y2023/V39/I1/1.
[95] P. Van Kampen and O. Gkioka, Undergraduate students' reasoning about the quality of experimental measurements of covarying secondary data, Eur. J. Phys. **42**, 045704 (2021).
[96] M. M. Botvinick, C. S. Carter, T. S. Braver, D. M. Barch, and J. D. Cohen, Conflict monitoring and cognitive control, Psychol. Rev. **108**, 624 (2001).
[97] S. P. Tipper, T. A. Bourque, S. H. Anderson, and J. C. Brehaut, Mechanisms of attention: A developmental study, J. Exp. Child Psychol. **48**, 353 (1989).
[98] S. Kertzman, I. Reznik, T. Hornik-Lurie, A. Weizman, M. Kotler, and D. Amital, Stroop performance in major depression: Selective attention impairment or psychomotor slowness?, J. Affective disord. **122**, 167 (2010).
[99] R. Stavy and R. Babai, Overcoming intuitive interference in mathematics: Insights from behavioral, brain imaging and intervention studies, ZDM Math. Educ. **42**, 621 (2010).
[100] B. A. Parris, N. Hasshim, M. Wadsley, M. Augustinova, and L. Ferrand, The loci of Stroop effects: A critical review of methods and evidence for levels of processing contributing to color-word Stroop effects and the implications for the loci of attentional selection, Psychol. Res. **86**, 1029 (2022).
[101] L. Ziaka, D. Skoteinou, and A. Protopapas, Task format modulates the relationship between reading ability and Stroop interference, J. Exp. Psychol. **48**, 275 (2022).
[102] The R Core Team, *R: A Language and Environment for Statistical Computing, version 3.3.0* (R Foundation for Statistical Computing, Vienna, Austria, 2025).
[103] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, lmerTest package: Tests in linear mixed effects models, J. Stat. Softw. **82** (2017).
[104] D. Pilson and K. L. Decker, Compensation for herbivory in wild sunflower: Response to simulated damage by the head-clipping weevil, Ecology **83**, 3097 (2002).
[105] A. Viarouge, H. Lee, and G. Borst, Attention to number requires magnitude-specific inhibition, Cognition **230**, 105285 (2023).
[106] M. K. Morphew *et al.*, Metallothionein as a clonable tag for protein localization by electron microscopy of cells, J. Microsc. **260**, 20 (2015).

[107] D. Curran-Everett, Explorations in statistics: The log transformation, Adv. Physiol. Educ. **42**, 343 (2018).
[108] R. C. Littell, G. A. Milliken, W. W. Stroup, and R. D. Wolfinger, *SAS for Mixed Models*, 2nd ed. (SAS Institute, Cary, NC, 2006).
[109] R. H. Baayen, D. J. Davidson, and D. M. Bates, Mixed-effects modeling with crossed random effects for subjects and items, J. Mem. Lang. **59**, 390 (2008).
[110] D. Bates, M. Mächler, B. Bolker, and S. Walker, Fitting linear mixed-effects models using LME4, J. Stat. Softw. **67**, 1 (2015).
[111] R. Gueorguieva and J. H. Krystal, Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry, Arch. Gen. Psychiatry **61**, 310 (2004).
[112] E. Theobald, Students are rarely independent: When, why, and how to use random effects in discipline-based education research, CBE Life Sci. Educ. **17**, rm2 (2018).
[113] R. V. Lenth, Estimated Marginal Means aka Least Squares Means, version version 1.8.4, R package (2023), available at https://CRAN.R-project.org/package=emmeans.
[114] Y. Zhu, L. Zhang, Y. Leng, R. Pang, and X. Wang, Event-related potential evidence for persistence of an intuitive misconception about electricity, Mind Brain Educ. **13**, 80 (2019).
[115] M. Limón, On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal, Learn. Instr. **11**, 357 (2001).
[116] B. Y. White and J. R. Frederiksen, Inquiry, modeling, and metacognition: Making science accessible to all students, Cognit. Instr. **16**, 3 (1998).
[117] H. R. Wilkinson *et al.*, Domain-specific inhibitory control training to improve children's learning of counterintuitive concepts in mathematics and science, J. Cogn. Enhancement **4**, 296 (2020).
[118] M. Letang, P. Citron, J. Garbarg-Chenon, O. Houdé, and G. Borst, Bridging the gap between the lab and the classroom: An online citizen scientific research project with teachers aiming at improving inhibitory control of school-age children, Mind Brain Educ. **15**, 122 (2021).