The Cognitive Cost of Poor Item Design: How Item-Writing Flaws Shape Response Time and Difficulty Across Assessment Domains

Steven Moore
Information Sciences and Technology
George Mason University
StevenJamesMoore@gmail.com

ABSTRACT

Item-writing flaws (IWFs), violations of established guidelines for constructing multiple-choice questions, are known to compromise assessment quality. Their effects on item difficulty have proved inconsistent, with flawed items sometimes appearing easier, sometimes harder, and often no different from well-constructed items. We propose that this reflects a measurement problem, as different IWFs impose distinct cognitive demands that cancel or dilute when aggregated. Drawing on cognitive load theory, we classify 19 automatically detected IWFs into three categories of cue-giving, confusion-inducing, and structural complexity based on their expected cognitive function. Analyzing 1,230 items across three domains (medical licensure, K-12 mathematics, and university computer science), we find that each domain exhibits a different mix of IWFs rather than simply more or fewer flaws. Flaws categorized as confusion-inducing consistently increase response time by 14-29% across all three datasets, while cue-giving and structural flaws show no significant effects. This pattern holds across all three datasets, with no significant flaw-by-dataset interactions, suggesting the effect is robust to the differences in subject matter. Bootstrap mediation analyses reveal that confusion-inducing flaws increase processing time, which in turn associates with greater difficulty. This indirect effect is 2.5 times the magnitude of the aggregate indirect effect. These findings suggest that item quality improvement efforts should prioritize confusion-inducing flaws, which are the primary source of unnecessary cognitive burden on examinees.

Keywords

Assessment Quality, Response Time, Question Difficulty

1. INTRODUCTION

Multiple-choice questions (MCQs) remain the most widely used assessment format in education, certification, and professional licensure, where their quality directly affects what we can validly infer about the knowledge and skills of students [58]. With generative AI now accelerating item writing, quality is both more necessary and more possible [4]. Clear standards govern how to create high quality questions, but violations of those standards, item-writing flaws (IWFs), continue to appear even in commercial e-learning instruction [9, 36]. This compromises construct validity, fairness, and the interpretability of learning analytics derived from these assessments [12, 56]. IWFs impose unnecessary cognitive load as students interpret unclear wording rather than demonstrate knowledge, artificially inflating time-on-task through extraneous processing [44, 26, 53]. Not all flaws are expected to impose the same kind of cognitive load, some may even reduce it by cueing the answer, while others introduce confusion that slows examinees down.

When researchers have tested whether IWFs actually affect item performance, the results have been equivocal. Studies examining whether IWFs predict item difficulty or discrimination have reported weak, inconsistent, or null associations [39, 41, 46, 49]. This matters practically as if flaws do not measurably affect how items perform, it becomes hard to justify the cost of detecting and remediating them, and automated detection tools lose their rationale. The problem may not be that flaws are inconsequential but that the standard approach of counting them, treating all 19 criteria as interchangeable contributors to a single "flaw burden" score, masks the signal. Consider that some flaws cue the correct answer, potentially making items faster and easier, while others introduce confusion that slows examinees down and makes items harder. Aggregating these into one count pits opposing effects against each other, producing the attenuated relationships that prior work has observed. The same logic applies to response time, because if cue-giving flaws speed responses and confusion-inducing flaws slow them, a composite IWF count should show a diluted association with processing time, even though specific flaw types carry real cognitive cost. To better guide revision efforts, what is needed is a categorization that separates flaws by what they do to the examinee's response process, so that the distinct (and potentially opposing) mechanisms can be estimated rather than canceled.

We test this hypothesis by categorizing 19 automatically detected item-writing flaws into three cognitive-function categories of cue-giving, confusion-inducing, and structural complexity, and then testing whether this categorization recovers

flaw-outcome relationships that are obscured in aggregate analyses. We analyze 1,230 MCQs bridging the three assessment domains of medical licensure (USMLE, n=653), K-12 mathematics (XES3G5M, n=365), and university computer science (DBE-KT22, n= 212). Each dataset provides item-level mean response time and a difficulty index, allowing us to trace the pathway from flaw presence through processing time to operational difficulty. The cross-domain design tests whether any observed effects reflect a general cognitive mechanism or are specific to a particular testing context. We investigate three research questions:

RQ1: How do item-writing flaw profiles differ across assessment domains?

RQ2: Do the effects of item-writing flaws on response time differ by cognitive-function category and are these effects consistent across domains?

RQ3: Does response time mediate the relationship between item-writing flaws and item difficulty and is this mediation strengthened when flaws are categorized by cognitive-function rather than treated in aggregate?

This study contributes (1) a cognitive-function taxonomy that classifies 19 IWFs into cue-giving, confusion-inducing, and structural complexity categories, revealing that only confusion-inducing flaws impose measurable cognitive cost on examinees, (2) evidence that this confusion effect replicates across three heterogeneous assessment contexts, increasing response time by 14–29% across medical, mathematics, and computer science assessments with no significant domain interactions, and (3) an explanation for why prior aggregate analyses found weak flaw–difficulty relationships. When flaws are categorized, the indirect effect of confusion-inducing flaws through slower response times to higher difficulty is 2.5× the magnitude of the aggregate effect, showing that lumping flaw types together masks the strongest signal.

2. RELATED WORK

2.1 Item-Writing Flaws

Item-writing flaws (IWFs) are features of multiple-choice questions that introduce testwiseness cues, compromise validity, and create unfair advantages for test-takers who recognize these patterns regardless of their actual knowledge [20, 56]. The fundamental concern is that IWFs shift what is being measured from domain knowledge to test-taking skill, introducing construct-irrelevant variance that distorts inferences about examinee ability [12]. Established taxonomies typically use a 19-criterion rubric applicable across subject areas [55, 9]. These capture grammatical or length cues that signal the correct answer, option-set pathologies such as K-type structures, and clarity violations including unfocused or vague question stems [36]. Despite awareness of these standards, IWFs remain prevalent even in professionally developed assessments, with studies finding 36–65% of items flawed in medical education [13, 8, 5] and similar rates in nursing [55], e-learning [9], and STEM disciplines [46, 49]. This persistence has motivated automated detection approaches, including rule-based systems [16] and hybrid systems like the Scalable Automatic Question Usability Evaluation Toolkit (SAQUET) [35]. This system operationalizes the full rubric and has achieved strong agreement with expert raters across diverse domains.

The empirical case that IWFs actually affect item performance is surprisingly weak. Studies in medical education found no consistent relationship between specific flaw types and difficulty or discrimination [41], with similar null results in veterinary [46] and broader STEM assessments [49]. One exception found that flawed items were easier in nursing assessments, but only for specific IWFs rather than the overall count [56]. If IWFs do not measurably affect how items perform, it becomes difficult to justify the cost of detecting and remediating them. However, the problem lies not with the flaws themselves but with how they have been measured.

2.2 The Aggregation Problem

Prior work treats all 19 IWF criteria as interchangeable contributors to a single score, but the rubric encompasses flaws whose cognitive mechanisms theory predicts should operate in opposite directions [8]. Cognitive load theory (CLT) distinguishes intrinsic load, inherent to the assessed content, from extraneous load imposed by how material is presented [10, 44]. In a well-constructed MCQ, the dominant load should be intrinsic as examinees expend effort retrieving and applying domain knowledge. IWFs distort this balance, but the direction of distortion depends on what the flaw does to the response process. Downing [12] drew a parallel distinction at the score level, noting that flawed items produce either construct-irrelevant easiness, which inflates scores, or construct-irrelevant difficulty, which deflates them. For instance, a word repeating between the MCQ's text and answer provides a lexical shortcut that lets test-wise examinees bypass retrieval (construct-irrelevant easiness) [20, 56]. In contrast, an ambiguous stem or gratuitous information forces examinees to infer what is being asked before retrieval can begin (construct-irrelevant difficulty), adding extraneous load [44] on features unrelated to the assessed content. These two classes of flaws should push response time and difficulty in opposite directions.

Aggregating them into a single count pits these opposing effects against each other. An item with one cue-giving flaw and one confusion-inducing flaw registers the same count as an item with two flaws of the same type, despite imposing very different cognitive demands [44]. Across a dataset, this cancellation systematically attenuates observed flaw-outcome relationships, producing exactly the weak or null associations that prior work has reported [41, 46, 39]. What is needed is a categorization that separates flaws by their cognitive function, so that opposing mechanisms can be estimated rather than canceled. This study tests this hypothesis by categorizing the 19 IWFs into three cognitive-function categories and examining whether this categorization recovers flaw-outcome relationships that aggregate analyses obscure.

2.3 Response Time and Difficulty

Response time (RT) has been established as a valuable process measure that captures cognitive processing demands during assessment tasks [40, 26]. Often grounded in CLT [44], RT can help distinguish between productive struggle with

content that supports learning and extraneous cognitive load imposed by irrelevant features that burden working memory [10]. Unlike accuracy alone, which provides only a binary outcome, RT reveals how students interact with items and the degree of cognitive effort required to formulate a response [50, 47]. This temporal signature is particularly valuable for understanding processing burdens imposed by poor item design. Theoretically, questions that introduce unnecessary confusion or complexity should produce characteristically longer processing times [61].

Previous research consistently demonstrates associations between longer RT and higher item difficulty, though the relationship is complex and multifaceted [22]. Multiple factors are known to influence RT beyond content difficulty itself, including the cognitive complexity of required reasoning, linguistic features such as readability and syntactic complexity, and structural characteristics including stem length and number of options [42, 29]. While RT relates to item structure and operational difficulty, it is unclear if RT mediates the relationship between specific quality issues (IWFs) and difficulty outcomes. Existing RT research has focused primarily on prediction rather than diagnosis, modeling overall time-on-task or completion rates without connecting temporal patterns to specific quality deficits that could guide authors in creating higher quality questions [61, 54].

2.4 Automated Item Diagnostics

Traditional psychometric quality control relies on post-hoc approaches such as Item Response Theory (IRT) and Classical Test Theory (CTT) to evaluate item performance [15, 52]. These methods require substantial response data, often hundreds of examinees per item, to produce stable parameter estimates [48]. This means that flawed items have already been deployed and consumed learner time before problematic questions are identified. Operationally, many assessments rely on simpler metrics such as proportion-correct or transformed accuracy indices rather than full IRT parameterization [15]. These have been shown to correlate highly with IRT difficulty estimates and require less extensive calibration data [62]. More fundamentally, both IRT and CTT characterize *what* an item's difficulty is, not *why* it is difficult. Neither framework traces the process pathway from specific design features through cognitive processing to observed outcomes, which is the question motivating the present study. Recent advances in machine learning and artificial intelligence have introduced pre-deployment difficulty prediction methods that leverage linguistic features [7], readability metrics [33], and large language models [3, 14] to estimate item difficulty from text alone. These approaches move closer to identifying difficulty sources before deployment, but they still treat items as isolated text rather than as interactive cognitive tasks, focusing on surface features without modeling the processing demands that items impose [17]. Also, they typically identify which items are likely to be problematic without providing actionable guidance about what specifically should be revised, limiting their utility for iterative quality assurance.

Parallel developments in automated time-on-task estimation and IWF detection have operated in similar isolation. RT estimation methods typically rely on shallow features or aggregate completion patterns without connecting temporal signatures to specific quality issues that create processing burdens [51]. Automated IWF detectors, while effective at flagging rule violations, function as diagnostic checklists rather than risk assessors [16, 35]. They identify the presence of flaws but do not indicate whether those flaws impose meaningful processing costs or distort difficulty in practice. The fundamental gap across these lines of work is the absence of an integrated framework that links flaw patterns to process signatures to operational outcomes. This disconnect becomes especially critical as AI-accelerated item generation scales assessment development, yet still produces MCQs that contain flaws [23, 28], even in standardized exams reviewed by multiple experts [34]. Practitioners need quality signals that are both predictive enough to enable pre-deployment screening and diagnostic enough to guide specific, actionable improvements.

Table 1: Overview characteristics of the three datasets.

	USMLE	**DBE-KT22**	**XES3G5M**
Domain	Medicine	Comp. Sci.	Math
N items	653	212	365
N options	4–8	2–5	2–5
Stakes	High	Low	Low
Resp./item	∼300+	458–1,049	5–3,237
RT: M (SD) s	85.6 (29.2)	78.5 (40.8)	197.2 (116.1)
Difficulty	Continuous	Ordinal	Continuous

3. METHODS

3.1 Datasets

We draw on three MCQ datasets spanning different disciplines, examinee populations, and testing contexts to assess whether the relationships between item-writing flaws and item performance generalize beyond a single setting. Our cross-setting comparisons therefore test whether the flaw–RT mechanism replicates across heterogeneous assessment contexts rather than isolating a pure domain effect. Table 1 summarizes the key characteristics of each dataset.

3.1.1 USMLE

The first dataset comprises retired clinical MCQs released by the National Board of Medical Examiners for a shared task on predicting item difficulty and response time [63]. The items are drawn from Steps 1, 2 CK, and 3 of the United States Medical Licensing Examination (USMLE) [37] and were authored by experienced subject-matter experts following guidelines that emphasize a standard MCQ structure. They discourage construct-irrelevant cues such as extraneous material, misleading wording, or correct options that are considerably longer than distractors. We excluded 14 of the 667 MCQs in this dataset for data-quality and format issues, primarily invalid Unicode in the source file or having more than 8 answer options, which the automated IWF evaluator we utilized does not support [35]. This yielded an analytic sample of 653 medical licensure MCQs. Each item was administered to approximately 300 or more examinees on a live high-stakes exam. For every item we observe the mean response time (RT) in seconds, computed from on-screen presentation until the examinee advances past the item, aggregated across all test-takers who encountered the item. We also observe a difficulty index defined as a linearly

transformed proportion correct scaled so that higher values indicate more difficult items.

3.1.2 DBE-KT22

The second dataset is drawn from DBE-KT22, a knowledge-tracing dataset collected between 2018 and 2021 on the CodeBench platform at the Australian National University [1]. It contains MCQs from an undergraduate Relational Databases course attempted by students across disciplines including computer science, engineering, business, and arts. We use all 212 items in the dataset. Each item was answered by between 458 and 1,049 students in a self-paced, low-stakes practice environment. Response time for each item is the mean time in seconds taken by students to submit an answer. Unlike the USMLE's continuous difficulty scale, difficulty in DBE-KT22 is an ordinal rating assigned by the course instructors (1=easy, 2=medium, 3=hard), which we retain as an ordinal outcome in downstream analyses.

3.1.3 XES3G5M

The third dataset originates from XES3G5M, a large-scale knowledge-tracing benchmark of third-grade mathematics exercises collected from an online learning platform operated by TAL Education Group in China [30]. The original items are in Chinese, so we use the English translation produced by Ozyurt et al. [38], who translated the question text via the deep-translate Python library for a related study. Because these items were machine-translated rather than originally authored in English, some detected flaws may partly reflect translation artifacts. However, the flaw–RT relationships in XES3G5M are consistent in direction and magnitude with those observed in the two English-native datasets and the pooled model finds no significant flaw-by-domain interactions (Section 4.2). This suggests that any such artifacts do not differentially drive the results.

From the translated item bank we selected the subset of MCQs for which at least five student responses and their associated timestamps were available, yielding 365 items. We adopt this inclusive threshold to maximize the analytic sample and models for XES3G5M use WLS with $\sqrt{n}$ weights to down-weight less precisely estimated item-level means. Robustness checks excluding items with fewer than 30 and 50 responses confirm that results are not driven by low-n items (Section 4.2). Because XES3G5M does not ship with precomputed item statistics, we derived both variables from the student-level interaction logs. Difficulty is computed as 1 minus the proportion of correct first-attempt responses, so that higher values indicate harder items. Response time for each student-item pair is estimated as the elapsed time between the start timestamp of the given question and the start timestamp of the next question in that student's sequence. These per-student times are then averaged across all students who attempted the item.

3.2 Automated Item-Writing Flaw Detection

We detect item-writing flaws (Table 2) with SAQUET (Scalable Automatic Question Usability Evaluation Toolkit) [1], an open-source and domain-agnostic system that operationalizes the standard 19-criterion IWF rubric to produce both per-flaw binary indicators and a total IWF count per item [35]. SAQUET was designed specifically as a tool to prescreen assessments before they are used with students, to capture testwiseness cues and clarity/alignment violations that introduce construct-irrelevant variance. This includes grammatical or length cues to the key, K-type option structures, negative or vague stems, non-parallel options, overlapping alternatives, and more. Prior work reports that such automated IWF application attains strong agreement with expert raters across diverse domains and that it outperforms readability-style proxies for question quality [35]. In this study we use SAQUET as the automated IWF engine and consume its outputs directly in our analyses. Following SAQUET's pipeline, we normalize Unicode and whitespace, lowercase, and tokenize stems and options before applying a layered set of detectors for the 19 IWF criteria.

We selected SAQUET because it is, to our knowledge, the only open-source tool that operationalizes the full 19-criterion IWF rubric in a domain-agnostic manner and has been used in previous research [49, 6]. Alternative MCQ evaluation systems either cover fewer criteria or are not publicly available [16]. Because SAQUET detects structural patterns rather than directly measuring cognitive disruption, its flags should be understood as indicators of flaw *presence* rather than flaw *impact*. This distinction bears on the interpretation of null results, particularly for cue-giving flaws (see Section 6). SAQUET returns a 19-dimensional binary vector per MCQ and derived aggregates (any-flaw and total IWF count). Full implementation details and previously reported accuracy (exact-match, Hamming loss, per-criterion F1) for this method are documented in prior work [35].

3.3 Flaw Category Taxonomy

The standard 19-criterion rubric [55] encompasses IWFs that should theoretically push examinee processing in opposite directions. We address this by mapping each of the 19 IWFs (described in Table 2) into one of three categories based on the type of cognitive demand the flaw is expected to impose. The mapping, summarized in Table 3, was fixed a priori before any outcome analyses were conducted based on the previous literature review. As outlined in Section 2.2, different IWFs are expected to distort the balance between intrinsic and extraneous cognitive load in opposing directions. We operationalize this distinction by mapping each flaw to one of three categories based on the direction of its expected effect as indicated in prior research.

We adopt three categories because the relevant theoretical distinction for our research questions is the direction of the predicted effect. Confusion-inducing flaws impose extraneous load with a clear directional prediction (↑ RT, ↑ difficulty), grounded in well-replicated effects of ambiguity, negation, and gratuitous information on processing demands [20, 44]. Structural-complexity flaws also alter processing, but the direction is not as predictable, as some (e.g. K-type with one obviously plausible combination) may simplify the task, others (e.g. "all of the above") expand it, and some operate primarily on guessing probabilities rather than RT. Collapsing these into a single extraneous load category would conflate flaws with directional predictions and flaws without, weakening the test the taxonomy is designed to perform. Cue-giving flaws form the third category because they carry the opposite directional prediction (↓ RT, ↓ diffi-

[1] `https://saquet.io/`

Table 2: The 19 Item-Writing Flaw (IWF) rubric criteria used to evaluate multiple-choice questions [55, 35].

Item-Writing Flaw	Description of Flaw
Ambiguous/Unclear	The question text or options contain unclear or ambiguous language
Implausible Distractors	One or more distractors are not plausible to be selected, reducing the effectiveness of the item
None of the Above	A "none of the above" option is included, which primarily tests the ability to detect incorrect answers
Longest Option Correct	The correct option is noticeably longer or more detailed than the distractors, cueing students
Gratuitous Information	The stem contains unnecessary information not required to answer the question
True/False Question	The options are structured as a series of true/false statements
Convergence Cues	Options contain overlapping combinations of components that hint at the correct answer
Logical Cues	The wording contains deductive shortcuts that allow the answer to be derived without domain knowledge
All of the Above	An "all of the above" option is included, allowing students to guess based on partial knowledge
Fill-in-the-Blank	Words are omitted from the middle of the stem for students to insert from the options
Absolute Terms	Absolute terms (e.g. always, never) appear in the options, which students recognize as typically false
Word Repeats	Words or phrases are repeated between the stem and the correct option but not in the distractors
Unfocused Stem	The stem does not present a clear, focused question that can be understood without reading the options
Complex or K-Type	The question requires selecting from combinations of responses rather than a single best answer
Grammatical Cues	Options are not grammatically consistent with the stem or parallel in style and form
Lost Sequence	Options are not arranged in a logical (e.g. chronological or numerical) order
Vague Terms	Vague terms (e.g. frequently, occasionally) are used in the options, whose meaning is subjective
More Than One Correct	The question does not have a single best answer, multiple options could be considered correct
Negative Wording	The stem uses negative wording, which can confuse students and obscure the intended learning outcome

culty), making a three-way structure necessary to represent the full range of theorized effects.

Category A (5 flaws): Cue-giving flaws reduce the effective cognitive work required to reach the correct answer. When a stem repeats a key phrase that appears only in the correct option, or when the correct option is conspicuously longer than the distractors, test-wise examinees can exploit these patterns to bypass the intended retrieval process [56, 20]. The five flaws in this category all share the property of providing a shortcut, such as word repeats creating lexical overlap between stem and key. The longest-answer-correct introduces a length asymmetry and grammatical cues allow elimination based on grammatical inconsistencies between an MCQ's text and options [59]. Convergence cues let overlapping options point toward the key and logical cues embed deductive pathways in the stem or option structure [18]. If these cues function as intended by the IWF literature, they should reduce both response time and difficulty [56].

Category B (6 flaws): Confusion-inducing flaws add processing demands that are extraneous to the assessed construct. An unfocused stem forces the examinee to infer what is being asked before retrieval can begin and ambiguous or unclear language requires re-reading and disambiguation [18]. Relatedly, gratuitous information consumes working memory without contributing to the answer [2]. Lost sequencing disrupts the natural scan order of options and negative wording introduces a reversal demand that is well documented to increase errors and slow responses [20]. Vague terms similarly force examinees to interpret subjective qualifiers rather than evaluate content, such as "usually" or "some". These six flaws share a common mechanism because they impose extraneous cognitive load in the sense of Puma et al. [44], consuming time-limited processing resources on construct-irrelevant features of the MCQ. We therefore predict that confusion-inducing flaws will increase response time and, through this added processing burden, associate with greater difficulty.

Category C (8 flaws): Structural complexity flaws alter the response format or option-set architecture in ways that change the task demands, but without a clear directional prediction. Complex K-type structures (e.g. "A and C", "B, C, and D") require combinatorial evaluation rather than single-best-answer selection, which may increase processing time but could also simplify the task if only one combination is plausible [60]. "None of the above" and "all of the above" options shift the response strategy from selecting the best answer to exhaustively verifying each option [11]. True/false formatting and fill-in-the-blank structures change the fundamental task from recognition to verification or recall [19]. Implausible distractors reduce the effective number of options, which could speed responses by shrinking the choice set or could paradoxically slow them if examinees second-guess why an obviously wrong option was included [4]. Absolute terms ("always", "never") occupy a genuinely ambiguous position as item-writing guidelines flag them because test-wise examinees learn to treat absolutes as likely incorrect [20]. This would place them in Category A, but the terms may also confuse less experienced examinees who take them at face value. We assign absolute terms to Category C to acknowledge this ambiguity. The eight flaws in this category share the property of restructuring the response task itself and we treat their effect direction as an empirical question rather than an a priori prediction.

The opposing predictions for Categories A and B are central to this study. If the taxonomy captures meaningful cognitive distinctions, an aggregate IWF count that collapses across categories should show a diluted effect, while category-level analyses should recover the underlying signals.

3.4 Analysis

For each item, we computed three structural control variables: stem length (word count of the question text), number of options (count of non-null answer choices, ranging from 2–8), and mean option length (average word count across options). From the 19 IWFs, we derived binary category variables indicating the presence of at least one cue-giving, confusion-inducing, or structural complexity flaw, as well as corresponding count variables for dose–response analyses. Response time (the item-level mean across all respondents)

Table 3: Mapping of 19 item-writing flaws to cognitive-function categories based on the expected effect on the examinee response process.

Item-Writing Flaw	Rationale
A: Cue-Giving (predicted: ↓ RT, ↓ difficulty)	
Word Repeats	Shared wording between stem and key lets examinees match rather than retrieve
Longest Ans. Correct	Length asymmetry singles out the key without content evaluation
Grammatical Cues	Agreement mismatches allow elimination of options on form alone
Convergence Cues	Overlapping option content narrows the viable set toward the key
Logical Cues	Deductive structure in the stem or options reveals the answer without domain knowledge
B: Confusion-Inducing (predicted: ↑ RT, ↑ difficulty)	
Ambiguous/Unclear	Unclear language forces examinees to infer intent before engaging content
Unfocused Stem	Absence of a clear question expands the space of plausible interpretations
Gratuitous Info	Irrelevant details load working memory without aiding the response
Lost Sequence	Unordered options disrupt systematic comparison across alternatives
Negative Stem	Negation imposes an additional reversal step during option evaluation
Vague Terms	Subjective qualifiers (e.g., "frequently") require interpretation beyond content
C: Structural Complexity (predicted: direction uncertain)	
Complex/K-Type	Combinatorial format may increase evaluation load or simplify when only one combination is plausible
None of the Above	Shifts task from selecting the best option to exhaustively verifying all options
All of the Above	Shifts task from selecting the best option to exhaustively verifying all options
Fill-in-the-Blank	Recall-oriented format changes the cognitive task from recognition to generation
True/False	Binary format alters baseline difficulty and guessing probability
Absolute Terms	May cue test-wise examinees ("always" = likely false) or confuse naïve ones
Implausible Distractors	Reduces effective choice set, but may also prompt second-guessing
More Than One Correct	Multiple defensible keys change elimination strategy unpredictably

was winsorized at the 1st and 99th percentiles within each dataset and log-transformed to approximate normality [32]. For within-dataset models, log(RT) serves as the outcome, enabling interpretation of coefficients as approximate percent changes in response time. For pooled cross-domain models, both RT and difficulty were z-scored within each dataset prior to concatenation, so that one unit represents one within-dataset standard deviation and cross-domain comparisons reflect relative rather than absolute effects. Difficulty is treated as continuous for USMLE and XES3G5M and as ordinal (1 = easy, 2 = medium, 3 = hard) for DBE-KT22. Because USMLE items were each answered by approximately 300+ examinees (roughly constant), those models use ordinary least squares with HC3-robust standard errors [31]. XES3G5M and DBE-KT22 have variable respondent counts per item (5–3,237 and 458–1,049, respectively), so models for these datasets use weighted least squares with $\sqrt{n}$ weights and HC3-robust standard errors to down-weight less precisely estimated item-level means [24].

To address RQ1, we compared the prevalence of each of the 19 IWFs across datasets using chi-square tests (Fisher's exact where expected cell counts fell below 5), with Benjamini-Hochberg false discovery rate correction ($q = 0.05$) applied across the 19 tests and Cramér's V quantified effect sizes at both the individual-flaw and category levels [21, 57]. For RQ2, we fit within-dataset regressions of log(RT) on the three category indicators and structural controls, followed by a pooled model with domain fixed effects and flaw-category-by-domain interactions to test cross-domain consistency. A non-significant joint Wald test for the interaction terms would support a domain-general mechanism [25]. For RQ3, we used bootstrap mediation [43] (5,000 iterations, bias-corrected and accelerated 95% confidence intervals) to estimate indirect effects of each flaw category on difficulty through response time. The aggregate single-indicator model serves as the baseline against which categorized indirect effects are compared. Mediation was conducted separately for USMLE and XES3G5M (continuous difficulty), pooled with domain fixed effects, and corroborated in DBE-KT22 using ordinal logistic regression in which coefficient attenuation upon adding RT indicates consistency with the mediation pathway. All analyses were conducted in Python using `statsmodels` (OLS/WLS, ordinal logistic, Wald tests) and `scipy` (chi-square, Spearman correlations, bootstrap implementations for mediation inference).

4. RESULTS

The final samples comprised 653 items (USMLE, medical), 365 items (XES3G5M, math), and 212 items (DBE-KT22, computer science), for a pooled 1,230 MCQs. All 19 IWFs were evaluated and detailed in the taxonomy in Table 2. Note that the more-than-one-correct IWF had zero instances across all three datasets and was excluded from all models.

4.1 IWF Profiles Across Domains

The three domains exhibited different flaw profiles in both their composition and prevalence (Table 4). Across all three datasets USMLE was the least flawed, with 64.8% of the MCQs being free of IWFs. DBE-KT22 sat at the opposite extreme, with 93.4% of the MCQs containing at least one flaw and XES3G5M occupied an intermediate position with 58.6% flawed. The datasets also differed in which flaw

Table 4: Flaw prevalence, category distribution, and mutual exclusion group sizes by domain.

	USMLE (Medical)	XES3G5M (Math)	DBE-KT22 (CS)
N items	653	365	212
% with any flaw	35.2%	58.6%	93.4%
Mean IWF count (SD)	0.47 (0.75)	0.99 (1.11)	2.02 (1.22)
% with cue-giving flaw	29.9%	21.1%	39.6%
% with confusion flaw	4.6%	39.2%	41.5%
% with structural flaw	6.1%	17.5%	65.1%
Mutual exclusion groups			
No cue/confusion	440 (67.4%)	178 (48.8%)	70 (33.0%)
Cue only	183 (28.0%)	44 (12.1%)	54 (25.5%)
Confusion only	18 (2.8%)	110 (30.1%)	58 (27.4%)
Both	12 (1.8%)	33 (9.0%)	30 (14.2%)

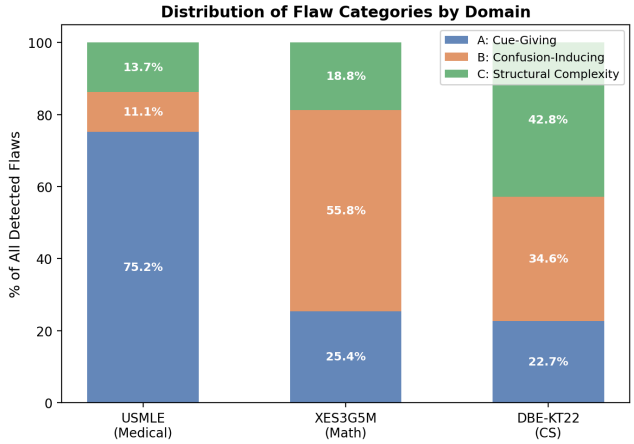

Figure 1: Distribution of detected flaws by cognitive-function category within each domain.

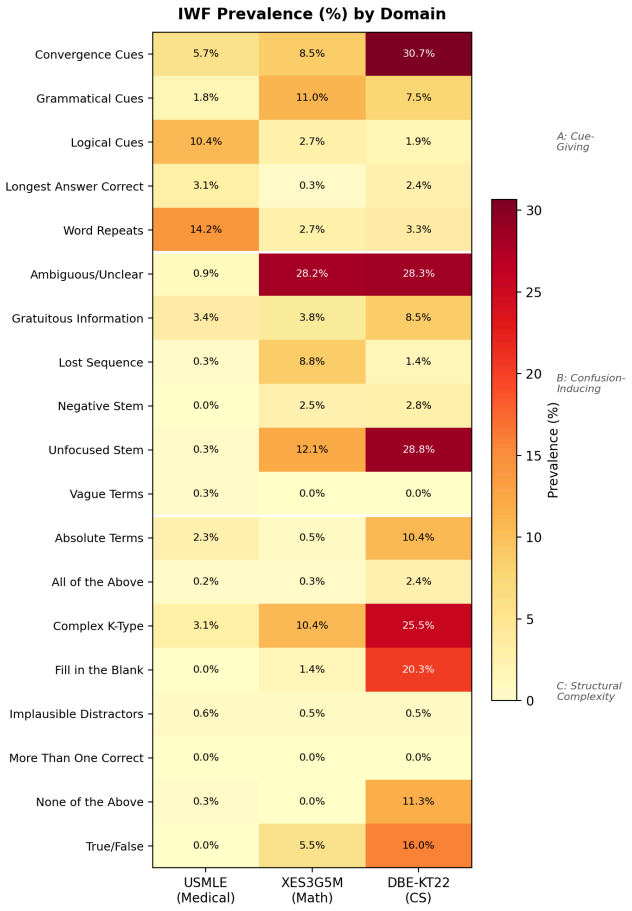

Figure 2: Prevalence of individual item-writing flaws by dataset.

categories dominated. Cue-giving flaws were most common in USMLE, driven by word repeats between stem and key (14.2%) and logical cues (10.4%). Confusion-inducing flaws were rare in USMLE but prevalent in both XES3G5M (39.2%) and DBE-KT22 (41.5%), driven primarily by ambiguous or unclear information (28% in both datasets) and unfocused stems (12.1% in XES3G5M, 28.8% in DBE-KT22). Structural complexity flaws were concentrated in DBE-KT22 (65.1%), where complex K-type items (25.5%), fill-in-the-blank formats (20.3%), and true/false items (16.0%) were common. As shown in Figure 1, when expressed as a share of all detected flaws the category composition of each domain is uniquely different. A majority of USMLE flaws were cue-giving, XES3G5M flaws were confusion-inducing, and DBE-KT22 was more evenly split between confusion and structural complexity.

Chi-square tests comparing each of the 19 IWFs across the three domains, with Benjamini-Hochberg correction for multiple comparisons ($q = 0.05$), found that 16 differed significantly in prevalence (Cramér's $V = 0.085$-0.396), with per-flaw prevalence shown in Figure 2. The three non-significant IWFs, vague terms, implausible distractors, and more-than-one-correct, all had near-zero prevalence in every dataset. At the category level, all three flaw categories differed significantly across domains (all $p < .0001$). The largest effect was for structural complexity ($V = 0.536$), followed by confusion-inducing flaws ($V = 0.433$) and cue-giving flaws ($V = 0.137$). These differences are not merely quantitative, as each domain has a distinct flaw *signature*. USMLE is dominated by cue-giving flaws, XES3G5M by confusion-inducing flaws, and DBE-KT22 by structural complexity. This heterogeneity suggests that treating IWFs as a single aggregate construct risks obscuring domain-specific patterns that may have consequences for examinee performance.

4.2 Item-Writing Flaws and Response Time

Before examining category-level effects, we established the aggregate association between IWF presence and response time. Spearman correlations between total IWF count and mean response time were significant and positive in all three datasets (Table 5). The OLS slopes translate this association into a practical metric of cognitive cost, where each additional detected IWF was associated with roughly 5-16 additional seconds of response time depending on the dataset. That this association holds across all three domains, despite differences in item format, examinee population, and absolute response time scales, establishes a consistent baseline that flawed items take longer to answer. This motivates the subsequent category-level analysis to determine whether this

Table 5: Aggregate association between flaw count and response time. Spearman ρ and OLS slope with 95% CI, all with $p < .001$.

	DBE-KT22	USMLE	XES3G5M
ρ	0.306	0.139	0.214
sec/flaw	+10.45	+4.86	+16.20
95% CI	[6.14, 14.76]	[1.89, 7.82]	[5.54, 26.86]

aggregate signal reflects a uniform contribution from all flaw types or is driven by a specific subset of them.

Within-dataset regression models predicting log(RT) from flaw-category indicators and structural controls revealed a clear asymmetry among flaw categories (Table 6). All models control for z-scored stem length, number of options, and z-scored mean option length, with USMLE estimated via OLS and XES3G5M and DBE-KT22 via WLS ($\sqrt{n}$ weights). Additionally, all use HC3-robust standard errors and coefficients are reported as percent change in RT ($\%\Delta = 100 \times (\exp(\beta) - 1)$). Confusion-inducing flaws were the only category to significantly increase RT, doing so in all three domains (+14-29%, all $p < .05$). A dose-response sensitivity analysis using IWF counts rather than binary indicators confirmed these patterns with the confusion count significant in all three domains ($p = .003$, $.010$, $< .001$). Cue-giving flaws showed no significant effect on RT in any dataset (all $p > .15$), with coefficients that were positive rather than negative, contrary to the a priori prediction that cue-bearing items would be answered faster. Structural complexity flaws were also non-significant everywhere (all $p > .52$).

A mutual exclusion grouping reinforced these findings as items with only confusion flaws took 16.5-27.0% longer than items free of IWFs. MCQs containing both cue-giving and confusion flaws showed nearly identical RT to the confusion-only group ($\Delta < 2\%$, both $p > .88$ in USMLE and DBE-KT22), indicating that the confusion component drives the effect. These results held across multiple robustness checks that included median RT (confusion significant in XES3G5M and DBE-KT22), unweighted models (all conclusions unchanged), exclusion of low-n items at thresholds of $n \geq 30$ and $n \geq 50$ (confusion $p = .009$ in both), and variance inflation factors below conventional thresholds (all VIFs < 1.5).

The pooled model with IWF-category-by-domain interactions ($R^2 = 0.146$, $N = 1{,}230$) provided a direct test of cross-domain consistency for IWFs increasing RT. The main effect of confusion-inducing flaws was significant ($\beta = +0.516$, $SE = 0.169$, $p = .002$), while the joint Wald test for all six interaction terms was non-significant ($F(6, 1215) = 0.956$, $p = .454$). Per-category interaction tests were also non-significant (cue × domain: $p = .853$, confusion × domain: $p = .167$, and structural × domain: $p = .275$). This pattern of a significant main effect with non-significant interactions supports the interpretation that confusion-inducing flaws impose a processing cost that replicates across these three heterogeneous assessment settings. We found 83-100% of estimable confusion-inducing flaws showed positive RT coefficients across datasets, while most cue-giving flaws did not show the predicted negative direction (20-60% alignment).

Table 6: Flaw category effects on log(RT) from within-dataset regressions, reported as percent change in RT.

Dataset	Cue-giving %Δ (p)	Confusion %Δ (p)	Structural %Δ (p)
USMLE ($N = 653$)	+3.7% (.151)	+14.0% (.016*)	−2.3% (.590)
XES3G5M ($N = 365$)	+15.0% (.233)	+29.4% (.005**)	+7.1% (.524)
DBE-KT22 ($N = 212$)	+8.1% (.295)	+18.9% (.040*)	−3.3% (.676)

4.3 Response Time Mediation of the Flaw-Difficulty Relationship

The aggregate flaw-RT association did not extend to a direct flaw-difficulty link. Spearman correlations between total IWF count and difficulty were non-significant in USMLE and XES3G5M, with only DBE-KT22 showing a reliable association ($\rho = 0.331$, $p < .001$). Meanwhile, the RT-difficulty link varied sharply across datasets, as it was strong in USMLE ($\rho = 0.525$, $p < .001$), moderate in DBE-KT22 ($\rho = 0.245$, $p < .001$), and near-zero in XES3G5M ($\rho = 0.060$, $p = .250$). Total effects of IWFs on difficulty, the direct association without accounting for RT, were non-significant in both datasets, for both the aggregate indicator and all three category-specific indicators (all $p > .18$). With flaws reliably increasing RT, but rarely associating directly with difficulty, mediation analysis is needed to test whether the flaw signal transmits *through* response time even when the direct path is negligible.

Formal mediation was conducted on USMLE and XES3G5M, which have continuous difficulty measures (Figure 3). Using 5,000 bootstrap iterations with bias-corrected confidence intervals, the aggregate indirect effect was modest in USMLE (+0.093) and non-significant in XES3G5M (+0.019). Categorizing by cognitive function revealed that the USMLE aggregate was driven almost entirely by the confusion pathway, as the confusion-inducing indirect effect (+0.231) was the only category to reach significance and was 2.5 times the aggregate magnitude. Cue-giving (+0.064) and structural (−0.040) indirect effects were non-significant. In XES3G5M, all category-specific indirect effects were non-significant, consistent with the weak RT-difficulty link in that dataset. The USMLE results also exhibited a suppression pattern as the positive indirect effect through RT coexisted with a near-zero total effect (total effect = −0.024) because the direct effect was negative (direct effect = −0.243 for confusion). This specifically indicates that the indirect and direct paths operate in opposing directions. The pooled model ($N = 1{,}018$, USMLE + XES3G5M with domain fixed effects) strengthened these findings. Both confusion (+0.117) and cue-giving (+0.051) indirect effects reached significance, with confusion roughly 2.3 times the cue-giving effect. Domain interaction terms were non-significant on both the flaw-to-RT and flaw-to-difficulty equations (both $p > .41$), supporting domain-general mediation.

In DBE-KT22 ($N = 212$), where difficulty is an instructor-assigned ordinal label, formal mediation is not applicable. Instead we tested whether adding RT to an ordinal logistic regression attenuated flaw-difficulty associations. Without

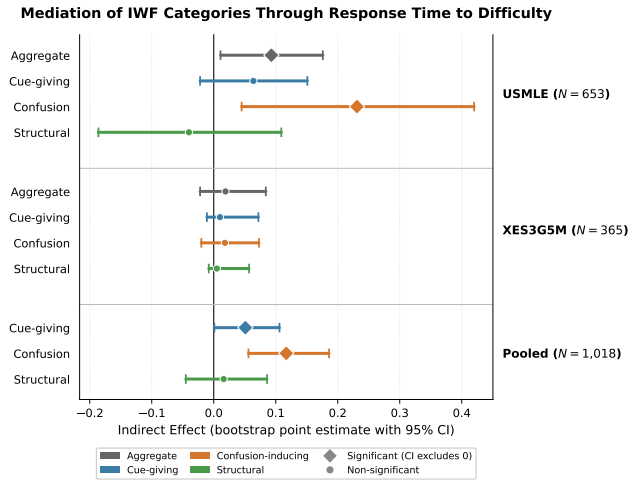

Figure 3: Bootstrap indirect effects (point estimates with 95% bias-corrected CIs) for the flaw category → RT → difficulty mediation pathway. All models control for stem length, option length, and number of options.

RT, both confusion ($\beta = +1.418$, $p < .001$) and structural ($\beta = +1.536$, $p < .001$) flaws predicted higher difficulty ratings. Adding log(RT) improved model fit (LR $\chi^2 = 12.85$, $p = .0003$) and attenuated the confusion coefficient by 10.3%, consistent with partial mediation through RT. Structural flaw coefficients showed no attenuation, suggesting their link to instructor-judged difficulty operates through other channels.

5. DISCUSSION

This study set out to test whether the inconsistent relationships between item-writing flaws with response time and difficulty reported in prior work [41, 46, 49, 39] reflect a measurement problem rather than a null effect. Across 1,230 items from three assessment domains we found that IWF profiles differ qualitatively, with each exhibiting a distinct signature rather than simply more or fewer flaws. When these flaws are categorized by their expected cognitive function, confusion-inducing flaws emerge as the sole category that imposes measurable processing cost, increasing response time by 14-29% across all three domains with no significant domain interactions. This processing cost transmits to difficulty where the RT-difficulty link is strong. In the medical MCQ dataset from the USMLE, the confusion-mediated indirect effect is 2.5 times the aggregate and the pooled model confirms confusion as the dominant pathway. The aggregate flaw-difficulty relationship is weak not because flaws do not matter, but because lumping cue-giving, confusion-inducing, and structural flaws into a single count dilutes the one category that carries the signal.

The domain specific IWF profiles are interpretable in light of each assessment's development context. USMLE's low flaw rate and cue-giving dominance is consistent with a professionally developed high-stakes exam where item writers follow established guidelines, but occasionally allow subtler cue-giving flaws (word repeats, logical cues) to persist [5]. DBE-KT22's saturation with structural complexity and confusion inducing flaws reflects the reality of instructor-authored items in a university CS course, where formal item-writing training is likely minimal [18]. XES3G5M's distinctive confusion profile, 39% of items flagged, driven by ambiguous stems and unfocused questions, may reflect the particular challenge of writing unambiguous items for younger learners in mathematics [27]. These domain specific signatures matter because they determine which category level effects will dominate in any given assessment context. A medical MCQ with one cue-giving flaw and a computer science MCQ with one confusion flaw both register identically under an aggregate count. However, our results show they carry fundamentally different cognitive consequences. Treating flaws as a monolithic construct, as prior work has largely done, obscures this heterogeneity and motivates the category-level analyses we conducted.

Our central finding is that confusion-inducing IWFs (ambiguous or unclear information, unfocused stems, gratuitous details, lost sequencing, and negative wording) are the active ingredient behind the aggregate flaw-RT association. Adding 5 to 16 additional seconds per IWF is carried entirely by the confusion-inducing category. Cue-giving flaws did not reduce RT in any dataset, contrary to the prediction from cognitive load theory and the IWF literature that they should speed responses by signaling the correct answer [56, 20]. Several explanations may account for this null result. SAQUET may detect structural patterns such as word overlap or length asymmetry that do not function as usable cues for examinees under time pressure. Cue-giving flaws may also co-occur with item properties that increase complexity, such as longer stems or more technical language, offsetting any facilitation. Alternatively, the cueing effect may operate on accuracy rather than speed, where examinees who would otherwise guess may benefit from the cue but not necessarily arrive at the answer faster [45]. This may be particularly if recognizing and exploiting the cue requires its own processing effort. Regardless of the cue-giving null, the confusion effect itself is robust across domains. Non-significant interaction terms (Wald $p = .454$) indicate that confusion-inducing flaws impose a comparable relative processing cost across these three datasets. However, since each domain is represented by a single dataset, we cannot disentangle domain from these co-varying factors. This consistency of the effect across such heterogeneous contexts is consistent with a construct-irrelevant processing burden, extraneous processing load in the language of cognitive load theory [44], that is independent of the domain knowledge being assessed.

The mediation results complete the chain from IWF presence through processing time to operational difficulty, while also revealing an informative boundary condition. In USMLE, where RT and difficulty are strongly linked ($\rho = .525$), the confusion-inducing indirect effect was 2.5 times the aggregate, revealing a mediation pathway that aggregation dilutes. In XES3G5M, confusion flaws still increased RT, but RT did not associate with difficulty ($\rho = .060$) and mediation was accordingly null. This divergence is informative rather than problematic as it suggests that what "difficulty" captures varies across assessment contexts. In USMLE, difficulty tracks closely with processing time where harder items take longer. In XES3G5M, difficulty may instead reflect conceptual barriers, such as misunderstanding the underly-

ing mathematics, that do not manifest as additional processing time. A student who lacks the requisite knowledge may answer quickly and incorrectly, producing high difficulty without elevated RT. This dataset dependence helps explain why prior work treating flaws in aggregate across single datasets has found weak or null flaw–difficulty relationships [41, 46]. Those studies conflated flaw types with different (or null) indirect effects and did not account for variation in how strongly aspects like RT track to difficulty. DBE-KT22 provides corroborative evidence from a third operationalization of difficulty. The confusion coefficients attenuated 10.3% when RT was added to an ordinal logistic model of instructor-assigned difficulty ratings, while structural flaw coefficients showed no attenuation. The differential attenuation pattern reinforces the category distinction where confusion flaws operate partly through response time and structural flaws reach difficulty through other channels. This may be through visible format features that instructors associate with complexity when assigning ratings. Together, these results indicate that categorization by cognitive function can help provide insight into the mechanism through which IWFs affect examinee performance.

There are interpretive caveats related to resposne time, as longer resposne time can reflect productive engagement, extraneous load, or sources unrelated to either. This makes it challenging to directly observe cognitive load at the item level. In this work, we treat the additional 14–29% response time under confusion-inducing flaws as cognitive cost because the flaws are by construction construct-irrelevant, but this is an inference from design rather than direct measurement. Additionally, the association of response time translating to performance harm depends on assessment context. In timed high-stakes tests, examinee time is bounded and per-item delays compound into real consequences for completion. In low-stakes practice the throughput cost is less acute, and the primary harm of confusion-inducing flaws may be to the construct validity rather than to performance directly. The XES3G5M boundary case, where confusion flaws increased RT but RT did not associate with difficulty, is consistent with this distinction.

The practical implication is that item quality improvement efforts should prioritize confusion-inducing flaws that create ambiguity, unfocused stems, and gratuitous information because these are the flaws that impose measurable cognitive cost on examinees. Cue-giving flaws remain a concern for validity if they inflate scores, and structural flaws may matter for format standardization, but neither category affects response time in our data. Because the confusion effect is domain general, this recommendation does not require adaptation for specific domains. The guidance to eliminate ambiguity and sharpen stem focus applies equally to medical licensure exams, K-12 math assessments, and university CS courses. For automated detection tools like SAQUET [35] or related ones [16], these findings suggest that not all flags deserve equal weight. A confusion-inducing IWF should trigger revision priority over a cue-giving flag when the goal is reducing construct-irrelevant cognitive load and saving examinee time. More broadly, the results argue for moving beyond binary "flawed or not" classifications toward taxonomies that distinguish flaws by their cognitive consequences. This approach may prove fruitful for both item quality assurance and for understanding how assessment design shapes examinee behavior. The 5-16 additional seconds per flaw documented in our analysis is a real cost, but attributing it equally to all IWFs misidentifies the source and therefore misdirects the remedy.

6. LIMITATIONS AND FUTURE WORK

Several design constraints temper the conclusions drawn here. All analyses operate at the item level using mean response times aggregated across respondents, which precludes examining within-person effects or distinguishing processing patterns, such as fast-correct versus slow-incorrect responses. The flaw indicators themselves are structural pattern matches produced by an automated detector, not ground-truth judgments of cognitive disruption. This distinction may partly account for the cue-giving null, if SAQUET flags formal patterns (e.g. word overlap) that do not function as usable cues under real testing conditions. DBE-KT22's small zero-flaw baseline ($n = 14$) increases estimator variance for that dataset, though the regression framework does not require a large joint-zero group in the same way ANOVA would. USMLE lacks median response time data, so the median-RT robustness check could not be applied to the dataset with the strongest mediation finding. While the three datasets used spanned unique domains, they also varied in examinee population, authoring practices, and more, which also attribute to the observed differences. Finally, although we use the language of mediation throughout, the design is observational and cross-sectional at the item level. These indirect effects should be interpreted as associational pathways rather than confirmed causal chains.

Student-level analysis with individual response times would permit testing whether confusion-inducing flaws impose uniform processing costs or disproportionately burden lower-ability examinees, a question the item-level design cannot address. Examining whether cue-giving flaws affect accuracy rather than speed would test the alternative interpretation that cueing operates on correctness without reducing response time. The taxonomy could also be applied to AI-generated items to determine whether large language models produce different flaw signatures than human item writers, a question of growing practical relevance as automated item generation scales. More broadly, the RT-based screening approach demonstrated here could be formalized into a practical quality assurance tool with decision thresholds that weight confusion flags above other flaw types, moving from the associational evidence presented in this study toward an actionable pre-deployment diagnostic.

7. CONCLUSION

Across 1,230 multiple-choice questions from three assessment domains, we show that weak or inconsistent relationships between item-writing flaws and item outcomes are often a measurement consequence of aggregation, not evidence that flaws are inconsequential. By introducing a cognitive-function taxonomy that separates cue-giving, confusion-inducing, and structural-complexity flaws, we reveal that different flaw types operate through distinct mechanisms, so collapsing them into a single flaw count can mask real effects. In particular, confusion-inducing flaws are the primary driver of processing burden, increasing mean response time by 14-29% across domains, an effect that held across

all three domains without significant moderation. When difficulty covaries with response time, the added time burden from confusion-inducing flaws is consistent with an indirect pathway to higher operational difficulty via response time. In the USMLE dataset, this indirect effect is 2.5× larger than the indirect effect based on the aggregate flaw count. For assessment practice, the implication is that not all IWFs deserve equal attention. Automated evaluation tools become more useful when they prioritize confusion-related issues, such as gratuitous information, that reliably add time due to wording rather than the target skill. As AI-generated assessment content scales, the taxonomy introduced here provides a cognitively grounded framework for quality assurance that moves beyond simply flagging flaws, to prioritizing the ones that impose measurable time costs on examinees and distort what assessments measure.

8. REFERENCES

[1] G. Abdelrahman, S. Abdelfattah, Q. Wang, and Y. Lin. Dbe-kt22: A knowledge tracing dataset based on online student evaluation. *arXiv preprint arXiv:2208.12651*, 2022.

[2] J. Abedi. Language issues in item development. In *Handbook of test development*, pages 391–412. Routledge, 2011.

[3] S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914, 2024.

[4] S. Asthana, T. Arif, and K. C. Thompson. Field experiences and reflections on using llms to generate comprehensive lecture metadata. In *NeurIPS'23 Workshop on Generative AI for Education (GAIED)*, 2023.

[5] M. H. Balaha, M. T. El-Ibiary, A. A. El-Dorf, S. L. El-Shewaikh, and H. M. Balaha. Construction and writing flaws of the multiple-choice questions in the published test banks of obstetrics and gynecology: adoption, caution, or mitigation? *Avicenna Journal of Medicine*, 12(03):138–147, 2022.

[6] N. Balepur, B. Rajasekaran, H. J. Oh, M. Xie, A. Desai, V. Gupta, S. J. Moore, E. Choi, R. Rudinger, and J. L. Boyd-Graber. Benchmarker: An education-inspired toolkit for highlighting flaws in multiple-choice benchmarks. In *Association for Computational Linguistics*, 2026.

[7] L. Benedetto, A. Cappelli, R. Turrin, and P. Cremonesi. R2de: a nlp approach to estimating irt parameters of newly generated questions. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 412–421, 2020.

[8] E. Costello, J. Holland, and C. Kirwan. The future of online testing and assessment: question quality in moocs. *International Journal of Educational Technology in Higher Education*, 15(1):1–14, 2018.

[9] E. Costello, J. C. Holland, and C. Kirwan. Evaluation of mcqs from moocs for common item writing flaws. *BMC research notes*, 11(1):849, 2018.

[10] K. E. DeLeeuw and R. E. Mayer. A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of educational psychology*, 100(1):223, 2008.

[11] D. DiBattista, J.-A. Sinnige-Egger, and G. Fortuna. The "none of the above" option in multiple-choice testing: An experimental study. *The Journal of Experimental Education*, 82(2):168–183, 2014.

[12] S. M. Downing. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, 77(10):S103–S104, 2002.

[13] S. M. Downing. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education*, 10(2):133–143, 2005.

[14] A. Dutulescu, S. Ruseti, M. Dascalu, and D. Mcnamara. How hard can this question be? an exploratory analysis of features assessing question difficulty using llms. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 802–808, 2024.

[15] X. Fan. Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, 58(3):357–381, 1998.

[16] T. Firoozi, L. Daniels, V. Daniels, and M. Gierl. An augmented intelligence system for automated quality control and feedback generation of multiple choice test items. In *International Conference on Artificial Intelligence in Education*, pages 86–93. Springer, 2025.

[17] F. Goldhammer, J. Naumann, A. Stelter, K. Tóth, H. Rölke, and E. Klieme. The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3):608, 2014.

[18] V. Gupta, E. R. Williams, and R. Wadhwa. Multiple-choice tests: A–z in best writing practices. *Psychiatric Clinics*, 44(2):249–261, 2021.

[19] A. Gurung, K. Vanacore, A. A. McReynolds, K. S. Ostrow, E. Worden, A. C. Sales, and N. T. Heffernan. Multiple choice vs. fill-in problems: The trade-off between scalability and learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 507–517, 2024.

[20] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333, 2002.

[21] W. Haynes. Benjamini–hochberg method. In *Encyclopedia of systems biology*, pages 78–78. Springer, 2013.

[22] M. Jiang and J. E. McLaughlin. Item discrimination, difficulty, and response time for 4-option mcqs versus 3-option mcqs. *American Journal of Pharmaceutical Education*, page 101408, 2025.

[23] B. Johnson, J. Dittel, and R. V. Campenhout. Investigating student ratings with features of automatically generated questions: A large-scale analysis using data from natural learning contexts. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 194–202, 2024.

[24] H. A. Kiers. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 62(2):251–266, 1997.

[25] D. A. Kodde and F. C. Palm. Wald criteria for jointly testing equality and inequality restrictions. *Econometrica: journal of the Econometric Society*, pages 1243–1248, 1986.

[26] P. C. Kyllonen and J. Zu. Use of response time for measuring cognitive ability. *Journal of Intelligence*, 4(4):14, 2016.

[27] J. Lee, D. Smith, S. Woodhead, and A. Lan. Math multiple choice question generation via human-large language model collaboration. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 941–946, 2024.

[28] U. Lee, Y. Kim, S. Lee, J. Park, J. Mun, E. Lee, H. Kim, C. Lim, and Y. J. Yoo. Can we use gpt-4 as a mathematics evaluator in education?: Exploring the efficacy and limitation of llm-based automatic assessment system for open-ended mathematics question. *International Journal of Artificial Intelligence in Education*, 35(3):1560–1596, 2025.

[29] Y.-H. Lee and Y. Jia. Using response time to investigate students' test-taking behaviors in a naep computer-based study. *Large-scale Assessments in Education*, 2(1):8, 2014.

[30] Z. Liu, Q. Liu, T. Guo, J. Chen, S. Huang, X. Zhao, J. Tang, W. Luo, and J. Weng. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems*, 36:32958–32970, 2023.

[31] J. S. Long and L. H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[32] P. Mair and R. Wilcox. Robust statistical methods in r using the wrs2 package. *Behavior research methods*, 52(2):464–488, 2020.

[33] W. Marinho, E. W. Clua, L. Martí, and K. Marinho. Predicting item response theory parameters using question statements texts. In *LAK23: 13th International learning analytics and knowledge conference*, pages 1–10, 2023.

[34] A. Masrur, A. Tayyab, Z. A. Khan, S. Jaffar, M. Mansoor, S. Mansoor, A. Rauf, and S. S. Shah. The impact of item-writing flaws on student test scores, pass rate and psychometric properties of the test in end of clerkship ophthalmology assessment. *Multicultural Education*, 8(4), 2022.

[35] S. Moore, E. Costello, H. A. Nguyen, and J. Stamper. An automatic question usability evaluation toolkit. In *International Conference on Artificial Intelligence in Education*, pages 31–46. Springer, 2024.

[36] R. Nedeau-Cayo, D. Laughlin, L. Rus, and J. Hall. Assessment of item-writing flaws in multiple-choice questions. *Journal for nurses in professional development*, 29(2):52–57, 2013.

[37] J. Norcini, I. Grabovsky, M. A. Barone, M. B. Anderson, R. S. Pandian, and A. J. Mechaber. The associations between united states medical licensing examination performance and outcomes of patient care. *Academic Medicine*, 99(3):325–330, 2024.

[38] Y. Ozyurt, S. Feuerriegel, and M. Sachan. Automated knowledge concept annotation and question representation learning for knowledge tracing. *arXiv preprint arXiv:2410.01727*, 2024.

[39] J. Pais, A. Silva, B. Guimarães, A. Povo, E. Coelho, F. Silva-Pereira, I. Lourinho, M. A. Ferreira, and M. Severo. Do item-writing flaws reduce examinations psychometric quality? *BMC Research Notes*, 9(1):399, 2016.

[40] Z. Papamitsiou, A. A. Economides, I. O. Pappas, and M. N. Giannakos. Explaining learning performance using response-time, self-regulation and satisfaction from content: An fsqca approach. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 181–190, 2018.

[41] H. Pham, J. Besanko, and P. Devitt. Examining the impact of specific types of item-writing flaws on student performance and psychometric properties of the multiple choice question. *MedEdPublish*, 7:225, 2018.

[42] C. Pires, A. Cavaco, and M. Vigário. Towards the definition of linguistic metrics for evaluating text readability. *Journal of Quantitative Linguistics*, 24(4):319–349, 2017.

[43] K. A. Pituch, L. M. Stapleton, and J. Y. Kang. A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. *Multivariate Behavioral Research*, 41(3):367–400, 2006.

[44] S. Puma, N. Matton, P.-V. Paubel, and A. Tricot. Cognitive load theory and time considerations: Using the time-based resource sharing model. *Educational Psychology Review*, 30(3):1199–1214, 2018.

[45] H. Reuss, A. Kiesel, and W. Kunde. Adjustments of response speed and accuracy to unconscious cues. *Cognition*, 134:57–62, 2015.

[46] B. R. Rush, D. C. Rankin, and B. J. White. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC medical education*, 16(1):250, 2016.

[47] I. Rushkin, I. Chuang, and D. Tingley. Modelling and using response times in online courses. *Journal of Learning Analytics*, 6(3):76–89, 2019.

[48] A. Sahin and D. Anil. The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice*, 17(1n):321–335, 2017.

[49] R. Schmucker and S. Moore. The impact of item-writing flaws on difficulty and discrimination in item response theory. *arXiv preprint arXiv:2503.10533*, 2025.

[50] S. D. Schneid, C. Armour, Y. S. Park, R. Yudkowsky, and G. Bordage. Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. *Medical Education*, 48(10):1020–1027, 2014.

[51] M. A. Shinnick and M. A. Woo. Validation of time to task performance assessment method in simulation: A comparative design study. *Nurse education today*, 64:108–114, 2018.

[52] A. Strugatski and G. Alexandron. Applying irt to distinguish between human and generative ai responses to multiple-choice assessments. In *Proceedings of the*

15th International Learning Analytics and Knowledge Conference, pages 817–823, 2025.

[53] J. Sweller, P. Ayres, and S. Kalyuga. Altering element interactivity and intrinsic cognitive load. In *Cognitive load theory*, pages 203–218. Springer, 2011.

[54] B. Tan and O. Bulut. The value of individual screen response time in predicting student test performance: Evidence from timss 2019 problem solving and inquiry tasks. *Journal of Intelligence*, 13(7):82, 2025.

[55] M. Tarrant, A. Knierim, S. K. Hayes, and J. Ware. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, 26(8):662–671, 2006.

[56] M. Tarrant and J. Ware. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical education*, 42(2):198–206, 2008.

[57] A. Telford, C. C. Taylor, H. M. Wood, and A. Gusnanto. Properties and approximate p-value calculation of the cramer test. *Journal of Statistical Computation and Simulation*, 90(11):1965–1981, 2020.

[58] D. R. Thomas, C. Borchers, S. Kakarla, J. Lin, S. Bhushan, B. Guo, E. Gatz, and K. R. Koedinger. Does multiple choice have a future in the age of generative ai? a posttest-only rct. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 494–504, 2025.

[59] M. H. Towns. Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, 91(9):1426–1431, 2014.

[60] A. Tripp and N. Tollefson. Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options?, 1985.

[61] S. Wang, S. Zhang, J. Douglas, and S. Culpepper. Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives*, 16(1):45–58, 2018.

[62] A. E. Wyse and S. Hao. An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement*, 36(7):602–624, 2012.

[63] V. Yaneva, K. North, P. Baldwin, L. A. Ha, S. Rezayi, Y. Zhou, S. R. Choudhury, P. Harik, and B. Clauser. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, 2024.