Differences in Individual Metacognitive Awareness Make Cognitive Offloading Tools a Double-Edged Sword: The case of Spell-Checking Tools

Jianning Wei
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan
j-wei@jaist.ac.jp

Kazushi Nishimoto
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
knishi@jaist.ac.jp

Abstract

Cognitive offloading research typically assumes uniform effects across users, yet individual differences in metacognitive awareness may influence these effects differently. We examined how users' everyday spell-checking practices interact with their metacognitive awareness. Using a behavioral measure of awareness (proportion of errors detected and corrected), we conducted a 2×2 experiment (awareness × tool usage, n=88). Participants completed a three-phase spelling task—baseline, error review, and validation—with performance assessed by the number of spelling errors. Results revealed a significant interaction effect ($F(1, 84) = 3.950, p = .050, \eta^2 = .045$): spell-checking tools minimally affected high-awareness users but resulted in significantly poorer performance among low-awareness users ($M = 1.70 \, vs. \, M = 7.74, d = 1.07$). These findings suggest that cognitive offloading tools may have differential effects depending on users' metacognitive capabilities, raising considerations for incorporating individual differences into assistive technology design and questioning one-size-fits-all design assumptions.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; **Interaction techniques**; • **Applied computing**;

Keywords

Cognitive Offloading, Metacognitive Awareness, Spell-checking Tools, Individual Differences, Human-computer Interaction

ACM Reference Format:

Jianning Wei and Kazushi Nishimoto. 2026. Differences in Individual Metacognitive Awareness Make Cognitive Offloading Tools a Double-Edged Sword: The case of Spell-Checking Tools. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3772318.3791355

1 Introduction

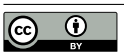
Cognitive offloading tools—from spell-checking tools to AI writing assistants—promise to enhance performance by delegating routine cognitive tasks to external systems [54]. However, research on these tools reveals contradictory findings about their impact on human capabilities. While some studies demonstrate performance benefits [41, 48, 56], others document skill degradation, such as the "Google effect" where search engine use reduces memory retention [60], and evidence that spell-checking tool dependence may impair spelling abilities [2, 22, 36].

These contradictory findings may stem from overlooking individual differences in how users interact with cognitive tools. Cognitive offloading theory distinguishes between strategic offloading—where users maintain core capabilities while selectively using tools—and non-strategic offloading driven by convenience or habit, which may lead to skill deterioration [28, 53]. This distinction suggests that identical tools may produce different or even opposite effects for users with varying cognitive characteristics, yet this hypothesis lacks systematic empirical validation.

Metacognitive theory provides a framework for understanding these individual differences. Nelson' [43] metacognitive regulation model distinguishes between monitoring (assessing one's cognitive state) and control (regulating cognitive processes based on monitoring outcomes). Users with strong metacognitive awareness can accurately assess their capability boundaries, appropriately judge when external assistance is needed, and maintain independent performance when tools are unavailable. Conversely, users with weak metacognitive abilities may lack accurate self-assessment and fail to regulate their tool dependence effectively, potentially leading to capability degradation over time.

Despite theoretical recognition of metacognition's role in cognitive offloading, empirical research faces critical challenges. Current metacognitive measures rely primarily on self-report instruments [40, 49, 58], which suffer from introspection limitations and social desirability bias, particularly the Dunning-Kruger effect [32] where less capable individuals overestimate their abilities. Moreover, static questionnaires cannot capture dynamic metacognitive decisions during task execution. In HCI research, most cognitive assistance studies focus on technical implementation or average effects while rarely considering individual cognitive differences as moderators [10, 17, 23, 34, 69]. This research gap limits both theoretical understanding of human-AI collaboration and the design of personalized cognitive assistance systems.

To address these challenges, we investigate how individual differences in metacognitive awareness moderate cognitive offloading effects, using spell-checking tools as our research context. Spell-checking provides an ideal testbed because: (1) spelling represents an objectively measurable cognitive skill, (2) spell-checking tools are widely used with high ecological validity, and (3) error detection



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791355

and correction involve the complete metacognitive cycle of monitoring and control. Our core innovation is a behavioral measurement method that captures metacognitive regulation through users' actual choices when presented with error feedback (whether to view corrections, whether to successfully revise), avoiding self-report biases while providing objective individual difference metrics.

This study makes three contributions to HCI research.

- **Empirical Contribution:** We provide the first systematic evidence that metacognitive individual differences moderate cognitive offloading effects, offering a unified framework for understanding previous contradictory findings.
- **Methodological Contribution:** We introduce behavioral metacognitive measurement that externalizes implicit cognitive processes into observable user interactions, providing a new assessment paradigm for HCI research.
- **Design Implications:** Our findings inform the design of adaptive cognitive assistance systems that can identify user metacognitive characteristics and adjust assistance strategies accordingly, maximizing benefits while minimizing potential harm.

Through investigating how metacognitive awareness moderates the relationship between spell-checking tool usage and spelling performance, this research advances cognitive offloading theory in HCI contexts and establishes foundations for more intelligent, personalized human-AI collaboration systems.

2 Related Works

2.1 The Dual Effects of Cognitive Offloading Theory

Cognitive offloading theory provides an essential theoretical framework for understanding cognitive resource allocation in human-computer collaboration. Risko and Gilbert [54] define cognitive offloading as "the process by which individuals delegate cognitive tasks to external resources, thereby reducing the burden on internal cognitive processing". The core assumption of this theory is that appropriate task offloading can free up limited cognitive resources, enabling users to redirect their attention toward higher-level cognitive activities.

Supportive evidence demonstrates significant positive effects of cognitive offloading. In memory tasks, external memory aids (such as digital scheduling systems) effectively reduce cognitive load and enhance task performance [6, 26]. In computational domains, assistive tools enable users to shift attention from low-level numerical operations to high-level strategic thinking and problem-solving [55, 64]. The concept of "partnership with technology" proposed by Salomon et al. [56] further emphasizes that technological tools can serve as cognitive partners, effectively enhancing human intellectual performance.

However, contradictory evidence reveals potential risks of cognitive offloading. The "Google effect" discovered by Sparrow et al. [60] demonstrates that when information retrieval becomes overly convenient, individuals' deep encoding and long-term retention of information significantly decline. Similarly, widespread use of GPS navigation systems may weaken users' spatial cognitive abilities [29], while over-reliance on calculators can impair the maintenance of basic mathematical computation skills [39, 59].

These contradictory research findings expose important limitations of current cognitive offloading theory. Although Risko and Dunn [53] attempted to distinguish between strategic offloading and non-strategic offloading, this classificatory framework is primarily based on outcome descriptions and lacks in-depth mechanistic explanations. Existing theory cannot answer a crucial question: why do identical external tools produce dramatically different cognitive offloading strategies and effects across different users?

2.2 Cognitive Offloading Differentiation in Writing Assistance Tools

Writing assistance tools provide an ideal research context for observing the differentiation effects of cognitive offloading. From a theoretical perspective, these tools should free up users' cognitive resources by handling low-level cognitive tasks such as spell-checking and grammar correction, allowing users to focus on higher-level creative activities like content ideation and logical organization [16, 61]. However, the differentiation effects observed in HCI research precisely confirm the challenges faced by cognitive offloading theory.

Evidence for positive effects is substantial. Experimental research by Galletta et al. [25] found that spell-checking tools significantly improve document quality while reducing users' cognitive burden. Grammar checking tools are particularly beneficial for non-native speakers [3, 9]. Recent research on AI writing assistants further indicates that these intelligent tools not only enhance writing efficiency but also stimulate user creativity [15, 19, 34].

However, research on negative effects is equally noteworthy. Figueredo and Varnhagen [22] found that users accustomed to spell-checking showed significantly deteriorated writing performance when tools were unavailable. Rimbar [52] further showed that spell-checker dependence weakened students' independent error repair abilities. More recent research shows that while AI grammar tools can improve surface textual correctness, users' understanding of underlying grammatical rules does not improve correspondingly [50, 70].

This differentiation pattern strongly suggests that the effects of cognitive offloading may be significantly moderated by individual user characteristics. However, existing HCI research faces two critical limitations in understanding this phenomenon:

- **Methodological limitations:** Traditional HCI research, influenced by "average user" assumptions, pursues universal design principles and thus focuses more on group average effects rather than individual variation patterns. While this research orientation aligns with practical technology design needs, it may obscure critical information for understanding human-computer interaction mechanisms [21, 67].
- **Measurement limitations:** Existing research primarily employs static pre-post designs that cannot effectively capture the dynamic adjustment processes of users' cognitive strategies during tool use [14, 30]. This measurement approach limits our deep understanding of the mechanisms underlying cognitive offloading differentiation effects.

2.3 Metacognition: The Theoretical Key to Understanding Individual Differences

Metacognitive theory provides an important theoretical perspective for explaining individual differences in cognitive offloading effects [38, 41, 51]. Flavell [24] defined metacognition as "cognition about cognition", emphasizing individuals' awareness and control abilities over their own cognitive processes. Nelson' [43] classic dual-level model further operationalizes metacognitive operations into two core components: metacognitive monitoring (assessing the accuracy of current cognitive states) and metacognitive control (actively regulating cognitive behavior based on monitoring results).

The importance of metacognition has been validated across multiple cognitive domains. A large-scale meta-analysis by Veenman et al. [66] found that metacognitive skills' predictive power for learning performance even exceeds traditional intelligence measures. In problem-solving tasks, individuals with high metacognitive abilities demonstrate superior strategy selection and execution monitoring capabilities [4]. In information retrieval contexts, metacognitive awareness correlates with more effective search strategies and enhanced information quality assessment [8].

2.4 Limitations of Self-Report Measures and the Case for Behavioral Alternatives

Despite metacognition's theoretical importance, measurement remains a persistent challenge. Traditional measurement methods primarily rely on self-report questionnaire tools, such as the Metacognitive Awareness Inventory (MAI) [58], Motivated Strategies for Learning Questionnaire(MSLQ) [49] and Metacognitive Awareness of Reading Strategies Inventory (MARSI) [40]. However, these widely-adopted instruments carry significant and well-documented limitations that constrain their applicability to HCI research.

First, self-report measures fundamentally depend on participants' introspective abilities and accurate self-assessment [18, 44, 45]. Introspection itself is inherently biased, individuals often lack accurate insight into their own cognitive processes, particularly regarding automatic or implicit monitoring behaviors [44]. Second, self-report instruments are highly susceptible to social desirability effects, where participants provide idealized rather than accurate responses [7, 47, 57]. Third, and critically, these static measurement approaches cannot capture the dynamic, moment-to-moment changes in metacognitive processes as they unfold during interaction [5, 72]. Most fundamentally, self-report measures are vulnerable to the Dunning-Kruger effect [20, 32]—individuals with lower metacognitive ability tend to overestimate their own metacognitive levels, systematically distorting the validity of assessments, particularly among users who would most benefit from metacognitive support.

These limitations have particular implications for HCI research on cognitive offloading. When users interact with intelligent systems, their metacognitive monitoring and control occur in real-time, often involving implicit judgments about when to rely on external tools versus their own cognition. Static, retrospective self-reports cannot adequately capture these dynamic decision-making processes. Furthermore, the ecologically invalid laboratory context of self-report questionnaires may fail to activate the same metacognitive processes that occur during naturalistic technology use.

To address these constraints, researchers across cognitive science and learning analytics have increasingly developed behavioral measurement approaches that infer metacognitive processes directly from observable user actions [31, 42, 63, 65]. Eye-tracking has been employed to infer metacognitive monitoring during complex tasks [42]; log-based analyses have been used to model self-regulated learning strategies [13, 71]; and click-stream traces have been leveraged to track monitoring and strategy shifts in digital learning environments [35]. These behavioral approaches offer three key advantages over self-report: they are objective (not dependent on subjective assessment), they are dynamic (capturing moment-to-moment changes), and they are ecologically grounded (derived from actual interaction behavior).

2.5 True Awareness Rate (TAR): A Unified Behavioral Measure of Metacognition

Building on this foundation of behavioral measurement, we introduce the True Awareness Rate (TAR)—a unified behavioral metric specifically designed to overcome the limitations of self-report measures while capturing the dynamic nature of metacognitive processes in HCI contexts. Rather than relying on post-hoc introspection or general self-assessments, TAR quantifies users' metacognition through their actual error detection (monitoring) and error correction (control) behaviors during real-time interaction.

TAR is operationalized through two observable behavioral metrics derived directly from user interaction traces. Error detection rate captures whether users notice problematic outputs (e.g., spelling errors, grammatical mistakes, or AI-generated inaccuracies). Error correction rate captures whether users take corrective action once problems are identified. By combining these two metrics into a unified measure, TAR provides a direct behavioral assessment of users' awareness and control during interaction without requiring retrospective self-judgment. This approach builds on findings from HCI and learning analytics research showing that user behavior traces can reliably reveal underlying cognitive processes [13, 35, 71].

Within spell-checking tool contexts specifically, users' moment-to-moment decisions about whether to accept, reject, or ignore tool suggestions directly reflect their metacognitive state: their awareness of potential language errors and their willingness to engage in revision. These interaction patterns reveal how different users with varying awareness levels interact with the same tool, enabling investigation of individual differences in cognitive offloading strategies. Modern systems automatically record such interaction traces, making TAR practical to implement at scale across diverse user populations and device contexts without the methodological limitations inherent in self-report measures.

2.6 Research Gaps in HCI and Research Questions

Despite the rich accumulation of metacognitive theory in cognitive science, its application in HCI research remains in its infancy. This situation is closely related to HCI's insufficient attention to individual differences [10, 17, 23, 34, 69]. Recent CHI work has

Table 1: Overview of User Tasks and Collected Data in Each Step

	Step1	Step2	Step3	Step4
	Baseline Test (Test1)	Error Review	Verification Test (Test2)	Survey
User Behavior	Listen and fill in missing words (50 items)	View corrections for incorrect input (click red underline)	Same as Test1	Answer questionnaire
Data Collected	Input text, Play button clicks	Clicks and content on review phase	Input text, Play button clicks	Questionnaire responses

begun highlighting the metacognitive demands of emerging AI-mediated interactions [11, 33, 61]. For example, Lee et al. show that AI-powered metacognitive calibration support influences learning outcomes through changes in learners' monitoring and behavioral strategies, underscoring substantial individual differences in metacognitive responsiveness [33]. Similarly, Chen et al. demonstrate that metacognitive monitoring and calibration critically shape users' reasoning and confidence in visualization tasks, reinforcing the need for behavioral measures when modeling metacognitive variability in interactive systems [11]. However, several critical research gaps remain.

First, little empirical evidence compares behavioral measures like TAR to traditional self-report instruments within HCI contexts, leaving unclear which approaches best capture users' actual metacognitive awareness during technology use. Second, the moderating role of metacognitive awareness on the effectiveness of cognitive offloading tools remains underexplored. Although prior work has examined various individual differences in HCI outcomes [1, 12, 37, 68], the specific mechanisms through which metacognitive awareness influences cognitive offloading effects have not been systematically investigated. Third, the design of cognitive offloading tools has rarely considered metacognitive variation among users. It remains an open question whether users with different levels of metacognitive ability benefit equally from the same tool features, or whether tools should be tailored to align with users' distinct metacognitive profiles.

To address these gaps, this study addresses three key questions:

RQ1: How do behavioral measures (specifically TAR) compare to self-report measures in assessing metacognitive awareness for HCI research?
RQ2: Does metacognitive awareness moderate the effects of cognitive offloading tools on user performance?
RQ3: What are the differential impacts of spell-checking tools on users with varying levels of metacognitive awareness?

3 Method

To address these research questions, we conducted a controlled experiment examining the role of metacognitive awareness in cognitive offloading effects. We employed a 2×2 between-subjects design examining the interaction between daily spell-checking tool usage (users vs. non-users) and metacognitive awareness level (high vs. low). The dependent variable was objective spelling performance (error count in Test2). This design allows us to test whether individual differences in metacognitive awareness moderate the effects of cognitive offloading tools.

3.1 Experimental Procedure and Materials

3.1.1 Experimental Procedure and Data Collection. The experiment employed a four-step sequential design to systematically capture multiple dimensions of cognitive processes and user behaviors (Table 1).

- **Step 1: Baseline Test (Test1)** required participants to listen to audio prompts and fill in missing words for 50 spelling items. We recorded participants' text input content while monitoring play button click frequency to ensure task completion validity and exclude interference from random responses due to insufficient audio playback.
- **Step 2: Error Review** provided opportunities for actively examining Test1 errors. The system marked incorrect answers with red underlines, allowing participants to autonomously choose which errors to examine by clicking to view correct spellings. This phase served as the core component for measuring metacognitive monitoring behavior, as we precisely recorded all user clicks on marked errors and the specific correction content participants accessed.
- **Step 3: Verification Test (Test2)** employed the same format as the baseline test, re-presenting the 50 spelling items to assess learning effectiveness. We continued collecting text input and play button click data, measuring performance changes through pre-post comparative analysis while similarly controlling for potential impacts of insufficient audio playback on result accuracy.
- **Step 4: Questionnaire Survey** collected participants' metacognitive self-assessment data through a questionnaire. All questionnaire responses were systematically recorded to support comprehensive analysis integrating subjective and objective data.

Figure 1 illustrates the error review interface used in Step 2. Correct responses remained unmarked, while misspelled words were underlined in red. When users clicked on an underlined word, the system displayed the correct spelling as immediate feedback.

This design achieved comprehensive integration of behavioral data (interaction patterns, text input), objective learning outcomes (pre-post comparisons), and subjective metacognitive processes (self-reports), establishing a solid data foundation for proposing novel behavioral measurement indicators and conducting in-depth analysis of their relationships with metacognitive awareness.

3.1.2 Materials.

- **Spelling Task:** Test1 and Test2 used a dictation format with 50 independent English sentences, each containing one target word replaced by a blank. Target words were common yet frequently misspelled words. Audio was generated using

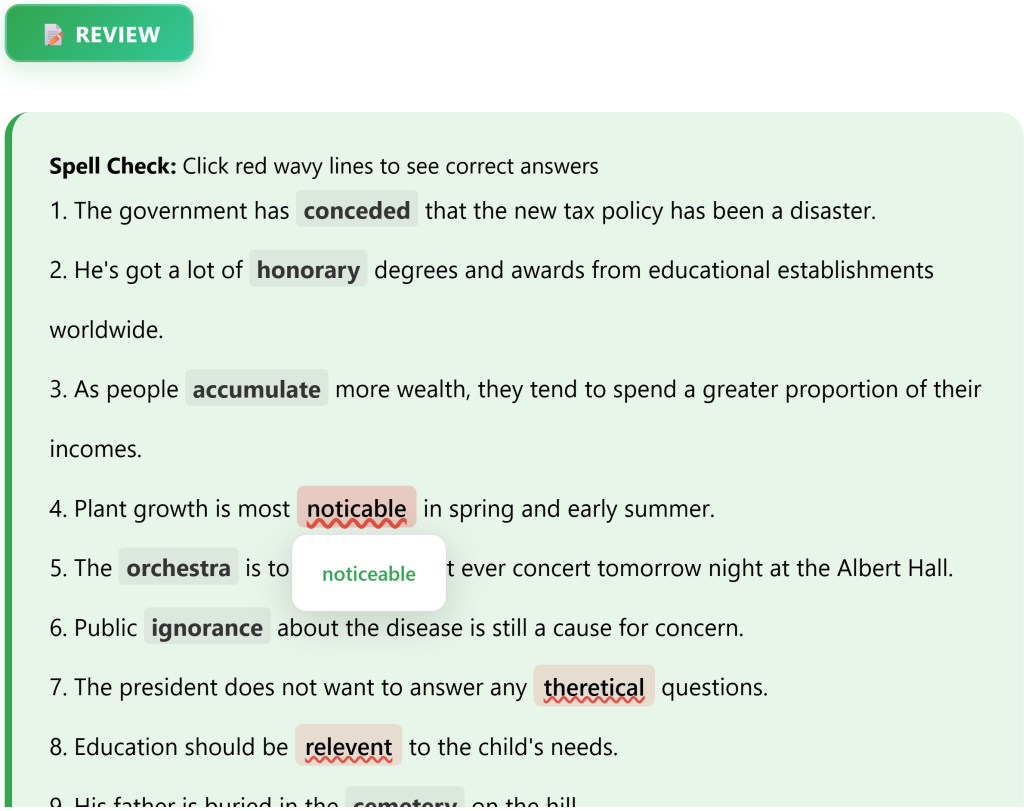


Figure 1: Error Review Interface

TTS at a natural pace, with unlimited replay available via a Play button. Both tests used the same sentences in the same order. Spell-checking tools were prohibited.

- **Questionnaire:** Subjective measures included task difficulty (5-point scales:1=very difficult, 5=very easy), confidence (5-point scales:1=not confident at all, 5=very confident), and self-reported frequency of spelling errors in daily writing (5-point scale: 1=almost never, 5=always). Participants' spell-checking tool usage was assessed using a binary yes/no question ("Do you use spell-checking tools in your daily work or studies?"), which was used to classify participants into spell-check users and non-users.

The system and data are open-sourced on Github[1].

3.2 Participants

One hundred English native speakers were recruited via Prolific Academic (www.prolific.co) with pre-screening criteria: UK/US nationality, native English speakers, age 18-50, high school education or above, and 95-100% platform approval rating.

Data Quality Control: Two exclusion criteria ensured valid data: (1) participants with zero baseline errors (n = 6) were excluded as our proposed behavioral measurement method requires errors as denominator, and (2) participants with insufficient audio engagement (n = 6) were excluded based on minimal play button interactions indicating possible repeat participation or task disengagement.

Final Sample: 88 participants (63.6% female, Mean age = 32.41, SD = 7.31). Geographic distribution: 83.0% US, 17.0% UK. Education: 90.9% bachelor's degree or higher. Daily spell-checking tool usage: 48.9%, providing balanced groups for analysis.

3.3 Behavioral Measurement of Metacognitive Awareness

3.3.1 True Awareness Rate (TAR): A Behavioral Approach. Given the inherent limitations of traditional metacognitive self-report measures, including subjective bias [18, 44, 45], social desirability effects [7, 47, 57], and the influence of the Dunning-Kruger effect [20, 32], this study proposes a task behavior-based metacognitive measurement method: True Awareness Rate (TAR). This method externalizes implicit metacognitive processes into observable interactive behaviors, measuring individuals' metacognitive monitoring-regulation capabilities through objective behavioral sequences, thereby avoiding the subjective judgment biases inherent in traditional self-report methods.

3.3.2 Theoretical Decomposition: MAR and CE. While TAR provides a comprehensive measure of metacognitive functioning, we decomposed it into two theoretically-grounded components to address methodological concerns and provide deeper insights:

[1]https://github.com/weismiling/audio-listening-experiment

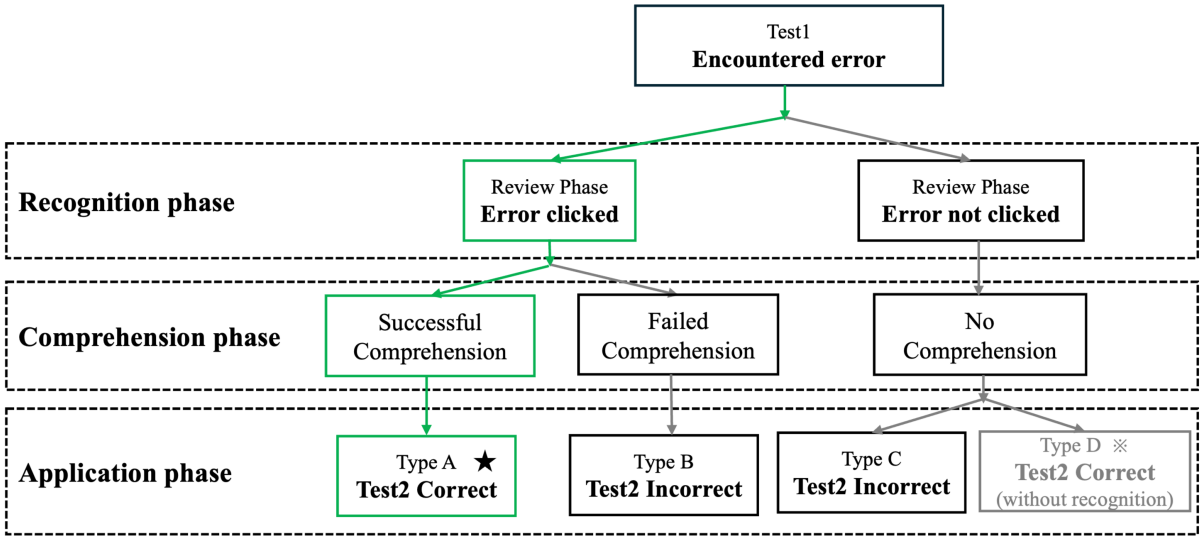


Figure 2: Three-Phase Model of True Awareness Rate (TAR) Measurement

Monitoring Awareness Rate (MAR): Represents the metacognitive monitoring process—the ability to recognize errors requiring attention. MAR captures the recognition phase of metacognition:

$$\text{MAR} = \frac{\text{Errors viewed during Review}}{\text{Total Test1 errors}} \tag{1}$$

Correction Efficiency (CE): Represents metacognitive control—the ability to process error information and apply corrections. CE captures the comprehension and application phases:

$$\text{CE} = \frac{\text{Errors both viewed during Review and corrected in Test2}}{\text{Errors viewed during Review}} \tag{2}$$

This decomposition aligns with Nelson' [43] framework while addressing potential circular dependencies in statistical analyses. The overall TAR metric is the product of these two processes:

$$\begin{aligned}\text{TAR} &= \text{MAR} \times \text{CE} \\ &= \frac{\text{Errors both viewed during Review and corrected in Test2}}{\text{Total Test1 errors}}\end{aligned} \tag{3}$$

3.3.3 Metacognitive Processing Pathways. As shown in Figure 2, we identify four distinct cognitive pathways from Test1 errors to Test2 outcomes, which can be understood through the lens of MAR (Monitoring Awareness Rate) and CE (Correction Efficiency):

Pathway 1: True Awareness Pathway (Type A)
This is the only complete pathway counted in TAR. After encountering an error in Test1, users demonstrate good metacognitive monitoring ability by actively clicking to view error information (contributing to high MAR). They then successfully understand and process the error information, correcting their knowledge and applying this understanding accurately in Test2 (contributing to high CE). This pathway embodies the complete metacognitive process from monitoring to regulation, representing the multiplicative success of MAR × CE.

Pathway 2: Comprehension Failure Pathway (Type B)
Users possess metacognitive monitoring ability and actively click to view error information in the recognition phase (high MAR), but problems occur in the comprehension and correction phase—they fail to effectively understand the error or update their incorrect memory to correct memory. Therefore, despite the viewing behavior, due to comprehension failure, they still answer incorrectly in Test2 (low CE). This pathway reveals the monitoring-control dissociation: high MAR × low CE = low TAR. It indicates that monitoring behavior alone is insufficient without effective cognitive processing.

Pathway 3: Monitoring Deficiency Pathway (Type C)
Users lack metacognitive monitoring ability and fail to recognize the need to check errors in the recognition phase, not clicking to view error information (low MAR). Due to the absence of crucial error information input, they cannot engage in meaningful error processing and continue making mistakes in Test2. This pathway reflects the fundamental impact of insufficient metacognitive monitoring ability on learning. Even if these users had high potential CE, their low MAR prevents meaningful correction opportunities.

Pathway 4: Incidental Correction Pathway (Type D)
This is a special case where users similarly lack monitoring ability (contributing to low MAR), do not click to view errors, and do not engage in error processing, yet answer correctly in Test2. This situation includes two subtypes: one is genuine "incidental correction," and the other involves users who made careless input errors in Test1 despite knowing the correct answer. Regardless of the subtype, when users face error checking in the Review interface, they lack conscious cognitive monitoring processes and are therefore excluded from TAR calculation. This design ensures that TAR measures conscious, understanding-based learning processes rather than chance factors or prior knowledge.

Quantitative Distribution: Through MAR-CE decomposition, these pathways can be quantified:

- Pathway 1 (High MAR × High CE): Complete metacognition
- Pathway 2 (High MAR × Low CE): Monitoring without control
- Pathway 3 (Low MAR × Any CE): Monitoring deficiency
- Pathway 4: Excluded from conscious metacognition measurement

This framework reveals that True Awareness requires both components: MAR captures whether learners identify problems requiring attention, while CE captures whether they can successfully remediate identified problems. The multiplicative relationship (TAR = MAR × CE) ensures that both monitoring and control are necessary for metacognitive success.

3.4 Data Analysis

Primary Variables:

- Dependent variable: Test2 spelling error count
- Independent variables:
  - Metacognitive awareness level: High (n = 44) vs. low (n = 44), based on TAR median split (Mdn = 0.52)
  - Tool usage: Users (n = 43) vs. non-users (n = 45)

Analysis Strategy: We examined TAR's predictive validity using Pearson correlations and linear regression analysis. A 2×2 between-subjects ANOVA (metacognitive awareness × tool usage) tested for interaction effects, followed by simple effects analysis. Effect sizes were calculated using Cohen's conventions, with demographic variables examined for potential confounding effects.

Statistical Software: All analyses were conducted using Microsoft Excel's Data Analysis ToolPak and R (version 4.2.2). Excel was used for descriptive statistics, ANOVA, and initial regression modeling, while R was used to validate regression coefficients, compute standardized β values, and confirm model statistics (R^2, ΔR^2). Both tools produced consistent results.

3.5 Research Hypotheses

Based on the research questions and our theoretical framework, we test the following specific hypotheses:

- **H1 (addressing RQ1):** Behavioral assessment of metacognitive awareness will demonstrate superior predictive validity compared to self-report measures in predicting spelling performance.
- **H2 (addressing RQ2):** A significant interaction will emerge between metacognitive awareness and spell-checking tool usage on learning outcomes.
- **H3 (addressing RQ3):** Tool usage will show differential effects—minimal impact for high-awareness users but performance degradation for low-awareness users.

4 Results

4.1 Validation of Behavioral Metacognitive Measures: TAR, MAR, and CE

Sample Characteristics: Table 2 presents participant characteristics and descriptive statistics by metacognitive awareness groups. The 88 participants averaged 7.70 spelling errors across 50 spelling items in Test1 (SD = 6.16) with a mean TAR of 0.52 (SD = 0.43), indicating appropriate task difficulty that provided adequate challenge while maintaining potential for learning gains and substantial individual differences.

4.1.1 Behavioral vs. Self-Report Measures (Testing H1). We tested H1 using hierarchical regression analysis predicting Test2 errors. Notably, we used the decomposed components (MAR and CE) rather than composite TAR to avoid circular dependency concerns, as TAR inherently includes correction counts that also appear in our outcome measure.

We constructed three nested models predicting Test2 errors:
Model 1 (Baseline): Test1 spelling errors significantly predicted Test2 performance ($R^2 = .662$, $p < .001$).
Model 2 (Self-Reported Awareness): Adding self-reported metacognitive awareness (Notice Mistakes subscale) yielded negligible improvement ($\Delta R^2 = .002$, $p = .443$), indicating that self-reported awareness does not enhance prediction beyond baseline.
Model 3 (Behavioral Indicators): Adding behavioral measures (MAR, CE) significantly enhanced prediction ($\Delta R^2 = .066$, $p < .001$; final $R^2 = .730$), indicating that behavioral indicators of metacognitive awareness meaningfully predict spelling performance.

CE emerged as a strong negative predictor ($\beta = -0.30$, $p < .001$), while MAR showed marginal effects ($p = .099$). This decomposition not only addresses methodological concerns but also reveals that correction efficiency contributes more than error detection to performance. The self-report measure remained non-significant throughout.

These findings directly support H1: behavioral assessments of metacognitive awareness provide superior predictive validity for spelling performance compared to self-report measures. Specifically, correction efficiency (a behavioral indicator) significantly improved model prediction ($\Delta R^2 = .066$, $p < .001$), while self-reported awareness contributed negligibly ($\Delta R^2 = .002$, $p = .443$).

In human-computer interaction research, which inherently deals with complex behavioral phenomena where numerous uncontrolled factors influence outcomes, effect sizes require careful interpretation. Our $\Delta R^2 = .066$ for behavioral indicators may appear modest compared to laboratory standards, but represents meaningful practical significance in the HCI context. Recent meta-analysis of CHI publications established that effect size thresholds in HCI are lower than Cohen's traditional guidelines, with small effects starting at $r \approx .14$ ($R^2 \approx .02$) and medium effects at $r \approx .27$ ($R^2 \approx .07$) [46]. Our incremental $R^2 = .066$ (corresponding to $r \approx .26$) falls at the threshold of medium effects in HCI research. More importantly, the contrast with self-report measures is stark: whereas behavioral indicators (MAR and CE) significantly improved prediction ($\Delta R^2 = .066$), self-report measures contributed negligibly ($\Delta R^2 = .002$, $p = .443$). This demonstrates that behavioral assessment provides superior predictive validity for actual performance outcomes compared to subjective self-appraisal. The finding that notice self-reports explained essentially zero additional variance ($\Delta R^2 = .002$) suggests that participants cannot accurately introspect about their metacognitive processes, or that self-report captures different constructs than those driving actual performance.

4.1.2 Validation of TAR-Based Grouping.

- **Control Variable Balance** Groups showed no significant differences across background variables: age ($t(86) = 0.90$, $p =$

Table 2: Participant Characteristics and Task Performance by Awareness Level

Variable	Overall N=88	High Awareness n=44	Low Awareness n=44	t/x^2	p
Test Performance					
Test1 Spelling Errors	7.70(6.16)	6.32(4.92)	9.09(6.98)	-2.15	.034*
Test2 Spelling Errors	3.41(5.44)	1.34(2.30)	5.48(6.76)	-3.84	.000**
Review Phase Behavior					
Total Clicks	5.69(6.96)	8.14(6.18)	3.25(6.90)	3.50	.001**
Click Rate	0.58(0.46)	0.98(0.07)	0.19(0.30)	16.99	.000**
True Awareness Rate	0.52(0.43)	0.92(0.13)	0.12(0.18)	23.25	.000**
Subjective Measures (1-5)					
Task Difficulty (Very Difficult-Very Easy)	3.49(0.88)	3.68(0.83)	3.30(0.90)	2.09	.040*
Answer Confidence (Not confident at all-Very confident)	4.31(0.78)	4.52(0.66)	4.09(0.83)	2.69	.008**
Self-reported Daily Spelling Errors (Almost Never-Always)	3.47(0.77)	3.32(0.80)	3.61(0.72)	-1.82	.073
Individual Characteristics					
Age	32.41(7.31)	33.11(7.28)	31.70(7.36)	0.90	.369
Gender (Female), n (%)	56(63.6%)	29(65.9%)	27(61.4%)	0.20	.658
Education (Bachelor+), n (%)	80(90.9%)	40(90.9%)	40(90.9%)	0.00	1.000
Daily Spell-Checking Tool Use, n (%)	43(48.9%)	20(45.5%)	23(52.3%)	0.41	.522

Note: Values are M (SD) for continuous variables and n (%) for categorical variables. Statistical comparisons between High vs Low Awareness groups. Click Rate= (Total Spell Check Clicks/ Total Test1 errors), True Awareness Rate = (Clicks corrected in Test2 / Total Test1 errors) Significance: *p < .05, **p < .01, ***p < .001*

.369), gender ($\chi^2(1) = 0.20$, $p = .658$), education ($\chi^2(1) < 0.001$, $p = 1.000$), or tool usage habits ($\chi^2(1) = 0.41$, $p = .522$). Tool usage rates were well-balanced between low-awareness (45.5%) and high-awareness (52.3%) groups, ensuring unconfounded analysis of interaction effects.

- **Group Discrimination Validity** TAR-based median split grouping ($Mdn = 0.52$) showed excellent discriminant validity. High-awareness ($n = 44$) and low-awareness ($n = 44$) groups differed dramatically on TAR ($high: M = 0.92$, $SD = 0.13$ vs. $low: M = 0.12$, $SD = 0.18$, $t(86) = 23.25$, $p < .001$, $d = 4.96$). Objective performance differences validated this grouping: Test1 errors ($high: M = 6.32$, $SD = 4.92$ vs. $low: M = 9.09$, $SD = 6.98$, $t(86) = -2.15$, $p < .05$, $d = 0.46$) and Test2 errors showed even stronger discrimination ($high: M = 1.34$, $SD = 2.30$ vs. $low: M = 5.48$, $SD = 6.76$, $t(86) = -3.84$, $p < .001$, $d = 0.82$).
- **Behavioral Consistency Validation** High-awareness users showed significantly higher click rates during Review phase ($high: M = 0.98$, $SD = 0.07$ vs. $low: M = 0.19$, $SD = 0.30$, $t(86) = 16.99$, $p < .001$, $d = 3.62$) and more total clicks ($high: M = 8.14$, $SD = 6.18$ vs. $low: M = 3.25$, $SD = 6.90$, $t(86) = 3.50$, $p < .001$, $d = 0.75$). In subjective perceptions, high-awareness users reported higher answer confidence ($high: M = 4.52$, $SD = 0.66$ vs. $low: M = 4.09$, $SD = 0.83$, $t(86) = 2.69$, $p < .01$, $d = 0.57$) and perceived the task as relatively easier ($high: M = 3.68$, $SD = 0.83$ vs. $low: M = 3.30$, $SD = 0.90$, $t(86) = 2.09$, $p < .05$, $d = 0.45$).

4.1.3 Interaction Effect Analysis (Testing H2) and Simple Effects Analysis (Testing H3). Participants were categorized into four groups by metacognitive awareness (high/low) and daily tool usage (users/non-users): high-awareness users ($n = 20$), high-awareness non-users ($n = 24$), low-awareness users ($n = 23$), and low-awareness non-users ($n = 21$).

A two-way ANOVA examining Test 2 spelling errors revealed a significant main effect of metacognitive awareness, $F(1, 84) = 16.318$, $p < .001$, $\eta^2 = .163$, and a significant main effect of daily tool usage, $F(1, 84) = 6.943$, $p = .010$, $\eta^2 = .076$. The interaction between metacognitive awareness and tool usage approached statistical significance, $F(1, 84) = 3.950$, $p = .050$, $\eta^2 = .045$.

The pattern shows that high-awareness individuals consistently produced few spelling errors regardless of tool usage ($M_{\text{Users}} = 1.70$ vs. $M_{\text{Non-users}} = 1.04$), whereas low-awareness individuals exhibited substantially higher error rates overall, with tool users performing the worst ($M = 7.74$) compared to non-users ($M = 3.00$).

The significant interaction was further examined through simple effects analysis using independent-samples t-tests to decompose the interaction:

Simple Effects of Tool Usage within Each Awareness Level:

- **High-Awareness Users:** No significant difference in Test2 spelling errors between tool users and non-users ($M = 1.70$ vs. $M = 1.04$ errors, $t(42) = 0.905$, $p = .372$, $d = 0.28$). This minimal effect suggests that metacognitively capable users maintain spelling competence regardless of tool usage patterns.
- **Low-Awareness Users:** Tool users showed significantly more spelling errors in Test2 than non-users ($M = 7.74$ vs. $M = 3.00$ errors, $t(42) = 2.526$, $p = .017$, $d = 0.77$) representing a large effect size with an average difference of 5.0 additional errors.

Simple Effects of Awareness Level within Each Tool Usage Condition:

- **Among Tool Users:** High-awareness users significantly outperformed low-awareness users in Test2 spelling errors ($M = 1.70$ vs. $M = 7.74$ errors, $t(41) = -3.403$, $p = .002$, $d = 1.07$).
- **Among Non-Users:** High-awareness individuals also outperformed low-awareness individuals in Test2 spelling errors ($M = 1.04$ vs. $M = 3.00$ errors, $t(43) = -2.064$, $p = .049$, $d = 0.63$).

The interaction is characterized by a large awareness effect among tool users (6.04 errors) but a smaller awareness effect among non-users (1.96 errors). The tool usage effect is minimal among high-awareness individuals (0.66 errors, $p = .372$) but substantial among low-awareness individuals (4.74 errors, $p = .017$).

The significant interaction ($p = .050$) directly supports H2: a significant interaction emerged between metacognitive awareness and spell-checking tool usage. Moreover, the simple effects pattern precisely supports H3: tool usage showed differential effects with minimal impact for high-awareness users ($p = .372$, $d = 0.28$) but performance degradation for low-awareness users ($p = .017$, $d = 0.77$).

5 Discussion

5.1 Breakthrough Progress in Metacognitive Measurement Methods

The core methodological contribution of this study lies in demonstrating the advantages of behavioral over self-report measures in assessing metacognitive awareness. Rather than relying on composite indices that risk circular dependency, we decomposed metacognitive performance into two distinct behavioral components: error detection rate (MAR—Monitoring Awareness Rate) and correction efficiency (CE—how effectively detected errors were corrected). Hierarchical regression analysis revealed striking differences in predictive validity between these approaches. Model 1, containing only baseline spelling ability (Test1 errors), explained 66.2% of Test2 performance variance ($R^2 = .662$, $p < .001$). Model 2, adding self-reported metacognitive awareness (Notice Mistakes), produced negligible improvement ($\Delta R^2 = .002$, $p = .443$), suggesting that participants' subjective assessments of their error monitoring contributed virtually no additional predictive power. In stark contrast, Model 3, incorporating the behavioral components MAR and CE, significantly enhanced prediction ($\Delta R^2 = .066$, $p < .001$; final $R^2 = .730$). The differential predictive contributions of MAR and CE are particularly informative. Correction efficiency (CE) emerged as a strong negative predictor ($\beta = -0.30$, $p < .001$), indicating that individuals who efficiently correct their errors achieve substantially better final performance. In contrast, error detection rate (MAR) showed only marginal effects ($p = .099$), suggesting that detecting errors is less predictive of performance than implementing effective corrections. This decomposition reveals an important insight: metacognitive regulation quality (correction efficiency) matters more than metacognitive monitoring accuracy (error detection).

Beyond predictive validity, behavioral analysis revealed that high-awareness users were significantly more likely to utilize review functions during task completion (click rate: $M = 0.98$ vs. $M = 0.19$, $t(86) = 16.99$, $p < .001$, $d = 3.62$). They also reported higher confidence in their answers ($t(86) = 2.69$, $p < .01$, $d = 0.57$) and perceived tasks as less difficult ($t(86) = 2.09$, $p < .05$, $d = 0.45$).This behavioral-perceptual pattern provides convergent evidence that our behavioral metacognitive indicators (MAR and CE) capture individual differences that manifest across multiple measurement domains—objective behavior, subjective confidence, and task perception—supporting its construct validity as a comprehensive measure of metacognitive processes.

Traditional metacognitive research has largely relied on self-report instruments such as the Metacognitive Awareness Inventory (MAI) [58] and the Motivated Strategies for Learning Questionnaire (MSLQ) [49]. Although widely used, these tools are subject to methodological limitations such as introspective bias and social desirability effects. In contrast, our behavioral decomposition—Monitoring Awareness Rate (MAR) and Correction Efficiency (CE)—provides a task-specific behavioral assessment of metacognitive monitoring and regulation during real performance. Unlike Schraw and Dennison's [58] distinction between metacognitive knowledge and regulation, TAR captures both detection and correction processes, thereby reflecting metacognitive regulation accuracy. While conceptually related to Nelson' [43] online measures (e.g., Judgments of Learning, Feelings of Knowing), MAR and CE exhibit stronger ecological validity, as they assess not only metacognitive judgment but also whether judgments are successfully translated into effective regulation.

However, this comparison also reflects differences in construct specificity. CE measures metacognitive effectiveness within the experimental task, while the self-report item reflects retrospective evaluations of participants' everyday writing habits. Thus, the superiority of CE may stem from genuine advantages of behavioral assessment, or from differences in task specificity. To more directly evaluate H1, future studies should compare CE with task-specific self-report items (e.g., "How accurately did you identify your errors in this task?"), or with established metacognitive scales (MAI, MSLQ) administered immediately after task completion to capture task-relevant metacognitive awareness. Together, these findings highlight the value of behavioral decomposition approaches for metacognitive assessment. Behavioral indicators offer stronger predictive validity and ecological relevance than general self-report measures, supporting the shift from introspective questionnaires toward task-embedded behavioral measures in metacognitive research.

5.2 Individual Difference Moderation Mechanisms of Cognitive Tool Effects

Our findings provide strong support for both H2 and H3. A two-way ANOVA examining Test 2 spelling errors revealed a significant interaction between metacognitive awareness and tool usage ($F(1, 84) = 3.950$, $p = .050$). More importantly, simple effects analysis revealed exactly the differential pattern predicted by H3: - For high-awareness individuals, tool usage had minimal impact.

High-awareness tool users (M = 1.70, SD = 2.83) performed virtually identically to high-awareness non-users (M = 1.04, SD = 1.76), with no significant difference ($t(42) = 0.905$, $p = .372$, $d = 0.28$). This suggests that for cognitively skilled individuals, spell-checking tools are neither dependence-inducing nor harmful. For low-awareness individuals, tool usage produced substantial performance degradation. Low-awareness tool users (M = 7.74, SD = 7.95) performed significantly worse than low-awareness non-users (M = 3.00, SD = 4.02), showing a difference of 4.74 spelling errors ($t(42) = 2.526$, $p = .017$, $d = 0.77$). This large effect size indicates that for lower-metacognitive individuals, spell-checking tools are associated with notably worse performance.

This interaction pattern is neither balanced nor incidental. Instead of exerting uniform positive or negative effects, the tool disproportionately harms low-awareness users while exerting little to no unfavorable impact on high-awareness users. This asymmetric effect carries substantial theoretical implications.

The finding that spell-checking tools impair low-awareness users' performance raises a key mechanistic question: why does tool availability specifically harm these users? We propose a Reduced Metacognitive Engagement ("Passive Correction") mechanism. When presented with tool-generated corrections, low-awareness users may accept suggestions passively without processing the underlying rule or reasoning. Rather than understanding why the correction is correct, they may simply click "accept", creating an illusion of competence—errors are corrected, but no cognitive understanding or knowledge updating occurs.

In contrast, when these users encounter errors without tool support, they are forced to engage more actively with problem solving. Even if difficult, this effort may elicit deeper cognitive engagement and potentially better learning. Tools that offer ready-made solutions may thus bypass the metacognitive effort required for learning, particularly for users with low metacognitive capacity.

An alternative explanation involves selection effects. Lower-ability individuals may preferentially adopt spell-checking tools because they struggle more with spelling. The observed performance difference might then reflect pre-existing ability differences rather than tool-caused degradation. However, our statistical control for Test1 baseline ability, combined with the specificity of the interaction pattern, somewhat mitigates this concern. If selection effects alone were responsible, we would expect to see tool-use differences across both awareness groups. Instead, tools appear selectively harmful only to low-awareness users. Additionally, H1's finding that behavioral correction efficiency (CE) independently predicts Test2 performance suggests that metacognitive capacity, beyond baseline ability, influences tool effects. Nevertheless, we acknowledge that distinguishing selection from causal effects requires stronger designs, as noted in the limitations section.

Risko and Gilbert's [54] cognitive offloading theory suggests that individuals tend to transfer cognitive burden to external tools to optimize overall performance. However, this theory mainly focuses on conditions for offloading occurrence, paying less attention to individual differences in offloading consequences. In the context of spelling-correction tools, our findings suggest that cognitive offloading effects are not universally beneficial but depend on individuals' metacognitive awareness. This finding, while limited to spelling tasks, suggests the need to incorporate individual metacognitive characteristics into offloading effect prediction models at least for text-editing and spelling-correction tools. Whether this pattern generalizes to other domains of cognitive offloading (e.g., search engines, calculators, AI writing assistants) requires further research.

5.3 Implications for CHI Research and Practice

5.3.1 From Demographics to Cognitive Characteristics in User Modeling. Traditional HCI research has predominantly relied on demographic variables to understand individual differences in technology adoption and usage [27, 62]. Digital divide studies focus on differences across social groups, while user modeling research primarily employs task and system characteristics as main variables. However, this study demonstrates a critical limitation of demographic-based approaches: participants were highly homogeneous on traditional variables (90.9% identical education levels, no significant differences in age, gender, or tool usage habits), yet grouping based on metacognitive awareness revealed substantial tool effect differences.

This finding suggests that cognitive characteristics, specifically metacognitive awareness, may be an important predictor of technology effects, at least for text-correction tasks. This does not necessarily imply that cognitive characteristics are more powerful than demographic factors in all technology contexts; rather, it highlights that cognitive ability assessment could complement demographic approaches in understanding individual differences in tool use particularly for cognitive assistance tools.

5.3.2 The Metacognitive Awareness-Tool Usage Interaction Model. Based on empirical findings, this study proposes the "Metacognitive Awareness-Tool Usage Interaction Model," categorizing users into four types: high awareness tool users, high awareness non-users, low awareness tool users, and low awareness non-users. Two-factor ANOVA confirmed significant interaction effects (F(1,84) = 3.950, p = .050), validating that tool effects are moderated by users' metacognitive characteristics.

This interaction pattern provides a new theoretical framework for understanding human-computer collaboration. Unlike traditional technology acceptance models emphasizing universal adoption, this model emphasizes the match between user cognitive characteristics and technology functions. High awareness users achieve "synergistic enhancement" where human-computer collaboration exceeds pure human or machine capabilities; low awareness users may fall into "collaboration traps" where over-reliance reduces overall performance.

5.3.3 Methodological Innovation: Behavioral Measurement in HCI Evaluation. This study's methodological contribution extends beyond metacognition research to HCI evaluation practices. The superiority of behavioral measurement (TAR: MAR and CE) over self-report measures provides a template for evaluating cognitive assistance tools. Traditional HCI evaluation relies heavily on subjective indicators like user satisfaction and usability scales, which may not capture deeper cognitive impacts of technology use.

A practical implication for CHI researchers is that task-based objective measures can meaningfully complement subjective assessments, especially when evaluating cognitive assistance tools.

For such technologies, understanding potential impacts on users' underlying capabilities can provide a fuller picture of tool effects not only immediate satisfaction or usability. Behavioral measures may reveal performance changes that users do not explicitly report, offering an additional lens for interpreting technology–user interactions.

5.3.4 Design Philosophy: From Universal to Personalized Assistance in the context of spell-checking tools. The findings from this spelling-task study suggest a refined approach to intelligent assistance design, at least for text-correction and spell-checking tools. The traditional "one-size-fits-all" approach assumes all users benefit from identical technology functions, but this study demonstrates such assumptions may produce counterproductive results for specific user groups.

A more tailored design philosophy should emphasize "cognitive adaptivity"—matching system functions with user cognitive characteristics. For the CHI community, this means moving from "technology serving everyone" to "technology serving people with different cognitive characteristics differentially." While these findings are specific to spell-checking, the principle of matching tool design to user cognitive characteristics may apply more broadly to cognitive assistance tools. However, validating this principle across different task domains remains an important direction for future work.

Practical implications for spell-checking tool design include:

- During user research phases, considering metacognitive awareness—specifically, users' ability to recognize and effectively correct their own errors—as a user characteristic alongside traditional demographics. Simple assessment tasks (measuring error detection and correction efficiency) could help identify users who may particularly benefit from enhanced tool support.
- Providing differentiated support strategies: progressive guidance for low-awareness users, enhanced control for high-awareness users:
  - ✓ For low-awareness users, the challenge involves two distinct deficits: First, low-awareness users show limited error detection rate (MAR). Simply underline errors passively may not be sufficient if users fail to notice the highlighted information. Tool design should consider: 1. More salient error highlighting (e.g., color contrast, animations); 2. Progressive disclosure that guides users' attention to errors they are most likely to miss. Second, low-awareness users show limited correction efficiency (CE). Even when errors are detected, they struggle to implement effective corrections. Tool design should provide: 1. Structured guidance (e.g. step-by-step correction review); 2. Explicit rule explanations that help users understand why an error is wrong; 3. Suggested corrections coupled with explanations of the underlying rules
  - ✓ For high-awareness users who both detect and correct errors effectively, providing greater control and transparency in tool function (e.g., detailed feedback, options to adjust sensitivity, ability to toggle explanations on/off) may be preferable, as they can efficiently utilize detailed correction information and don't require extensive guidance.

While the present study employed a relatively basic spell-checking paradigm, the evidence provides an essential foundation for personalized intelligent assistance in more sophisticated cognitive offloading technologies. The metacognitive individual-difference mechanisms identified here offer methodological and theoretical support for future adaptive interaction systems capable of dynamically adjusting assistance based on users' cognitive profiles.

5.4 Research Limitations and Future Research Directions

5.4.1 Limitations in Causal Interpretation. As with most cross-sectional HCI studies, our design does not allow strong causal claims. The observed interaction pattern—where low-awareness tool users performed worse than low-awareness non-users—may reflect either a causal mechanism (e.g., reduced metacognitive engagement when using tools) or a selection effect (i.e., lower-skill individuals being more likely to use tools).

Although we statistically controlled for baseline ability, the strong Test1–Test2 correlation ($R^2 = .662$) indicates that pre-existing differences remain a substantial factor. Therefore, our findings should be interpreted as documenting associations and individual differences, rather than definitive causal effects.

Future work using randomized tool-assignment and longitudinal designs is needed to determine whether tool use itself produces differential learning outcomes across awareness levels.

5.4.2 Methodological Limitations. The limitations are listed as follows:

1. Short-Term vs. Long-Term Learning Effects A key limitation of our study concerns the distinction between short-term performance and genuine internalization. Test 2 was administered immediately (≈5 minutes) after the error-review phase, making it more sensitive to short-term working-memory effects than to long-term consolidation. While this immediate post-test aligns with our primary goal—examining how individuals with different metacognitive profiles respond to error feedback in the short term—it prevents us from determining whether the observed improvements reflect durable learning or merely transient recall. Internalization is conceptually distinct from temporary performance gains and establishing it typically requires delayed retention tests or transfer assessments.

Moreover, our single-session design with only one brief review opportunity provides a snapshot of participants' metacognitive awareness and spelling performance rather than a full picture of their learning trajectories. Such designs inherently impose ceiling-effect constraints, especially for high-awareness individuals with limited improvement space, making change-score analyses less interpretable.

To address these interconnected limitations, future research should adopt extended learning designs that span multiple days or weeks. Specifically, we recommend: (1) multiple review cycles to derive meaningful change scores less constrained by ceiling effects; (2) intermediate assessments to capture learning trajectories; (3) delayed retention tests to evaluate the durability of learning; and

(4) transfer tasks to assess whether metacognitive gains generalize beyond the trained domain. Together, these extensions would provide converging evidence for differential learning effects, clarify whether metacognitive awareness influences long-term internalization, and determine whether tool use leads to lasting cognitive change.

2. Measurement Limitations

While our Discussion section notes that the behavioral measure (CE) and self-report measure reflect different constructions with different levels of specificity, an additional measurement consideration concerns the design of the self-report item. The self-report measure used in H1 was a single yes–no question; even when administered post-task, single-item measures lack the reliability and validity evidence found in established multi-item scales (e.g., MAI, MSLQ). Future validation of behavioral decomposition approaches would benefit from comparisons with task-adapted multi-item instruments or with single-item measures that have been independently validated for assessing task-specific metacognitive awareness.

Additionally, tool usage was measured via self-report rather than behavioral logging. Participants may inaccurately report daily tool use frequency, introducing measurement error that could affect the observed interaction pattern. Future studies incorporating log-based or passive behavioral tracking methods would help improve the accuracy and reliability of tool-use measurement.

6 Conclusion

Through spelling task experiments with 88 native English speakers, this study reveals the critical moderating role of metacognitive awareness in cognitive tool effects. Task-specific behavioral indicators of metacognitive awareness—particularly correction efficiency—predicted spelling performance more strongly than a self-report measure ($\Delta R^2 = .066$, $p < .001$ vs. $\Delta R^2 = .002$, $p = .443$), indicating that behavioral measures captured metacognitive processes more closely tied to task performance. More importantly, two-factor ANOVA confirmed a significant interaction effect between metacognitive awareness and spell-checking tool usage habits ($F(1, 84) = 3.950$, $p = .050$), specifically manifesting as substantial detrimental effects on low metacognitive awareness users (an average increase of 4.7 errors), while being essentially neutral for high awareness users, suggesting that technology effects may not be universal across all user populations.

These findings offer several key contributions to HCI research and practice: Methodologically, this work demonstrates the value of task-based behavioral assessment for estimating metacognitive processes, complementing traditional self-report approaches; Theoretically, they suggest an initial framework for understanding metacognitive awareness-tool usage interactions, offering new perspectives on the role of individual differences in technology effects; Practically, the results motivate exploring more personalized forms of assistance, particularly for text-correction and spell-checking systems. For the CHI community, this study advocates incorporating user cognitive characteristics into system design considerations, shifting from "one-size-fits-all" approaches toward technology that adapts to diverse cognitive profiles.

Future research should validate this framework across diverse cognitive tools and cultural contexts, advancing the development of truly personalized intelligent assistance technologies.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP24KJ2244 and JP24K02976.

References

[1] Phillip L. Ackerman and Margaret E. Beier. 2007. Further explorations of perceptual speed abilities in the context of assessment methods, cognitive abilities, and individual differences during skill acquisition. *Journal of experimental psychology. Applied* 13, 4 (2007), 249–272. https://api.semanticscholar.org/CorpusID:2493520

[2] Hewa Fouad Ali, Lisa Jamal Nakshbandi, Fatima Saadi, and Sami Hussein Hakeem Barzani. 2022. The effect of spell-checker features on spelling competence among EFL Learners: An empirical study. *International journal of social sciences & educational studies* 9, 3 (2022), 101–111.

[3] Hewa Fouad Ali, Lisa Jamal Nakshbandi, Fatima Saadi, and Sami Hussein Hakeem Barzani. 2022. The effect of spell-checker features on spelling competence among EFL Learners: An empirical study. *International journal of social sciences & educational studies* 9, 3 (2022), 101–111.

[4] Roger Azevedo and Allyson F. Hadwin. 2005. Scaffolding Self-regulated Learning and Metacognition – Implications for the Design of Computer-based Scaffolds. *Instructional Science* 33 (2005), 367–379. https://api.semanticscholar.org/CorpusID:58144594

[5] Roger Azevedo, Amy Johnson, Amber Chauncey, and Candice Burkett. 2010. Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In *New science of learning: Cognition, computers and collaboration in education*. Springer, 225–247.

[6] Hunter Ball, Phil Peper, Durna Alakbarova, Gene Brewer, and Sam J Gilbert. 2022. Individual differences in working memory capacity predict benefits to memory from intention offloading. *Memory* 30, 2 (2022), 77–91.

[7] Nicole Bergen and Ronald Labonté. 2019. "Everything Is Perfect, and We Have No Problems": Detecting and Limiting Social Desirability Bias in Qualitative Research. *Qualitative Health Research* 30 (2019), 783 – 792. https://api.semanticscholar.org/CorpusID:209340442

[8] Saskia Brand-Gruwel, Iwan Wopereis, and Amber Walraven. 2009. A descriptive model of information problem solving while using internet. *Comput. Educ.* 53 (2009), 1207–1217. https://api.semanticscholar.org/CorpusID:15727688

[9] Jack Burston. 2001. Exploiting the potential of a computer-based grammar checker in conjunction with self-monitoring strategies with advanced level students of French. *Calico Journal* 18, 3 (2001), 499–515.

[10] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[11] Mengyu Chen, Andrew Yang, Seungchan Min, Kristy A Hamilton, and Emily Wall. 2025. A Novel Lens on Metacognition in Visualization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[12] Sherry Y. Chen and Robert D. Macredie. 2002. Cognitive styles and hypermedia navigation: Development of a learning model. *J. Assoc. Inf. Sci. Technol.* 53 (2002), 3–15. https://api.semanticscholar.org/CorpusID:30023121

[13] Anindya Damayanti, Sri Suning Kusumawardani, and Sunu Wibirama. 2023. A Review of Learners' Self-Regulated Learning Behavior Analysis Using Log-Data Traces. In *2023 IEEE 12th International Conference on Engineering Education (ICEED)*. IEEE, 90–95.

[14] Ali Darejeh, Nadine Marcus, Gelareh Mohammadi, and John Sweller. 2024. A critical analysis of cognitive load measurement methods for evaluating the usability of different types of interfaces: guidelines and framework for Human-Computer Interaction. *ArXiv* abs/2402.11820 (2024). https://api.semanticscholar.org/CorpusID:267750906

[15] Sholihatul Hamidah Daulay, Hotma Berutu, Muhammad Dalimunte, Muhammad Muslim Nasution, Eka Apriani, Muthmainnah Muthmainnah, and Shaumiwaty Shaumiwaty. 2024. What AI-Based Writing Assistant Actually Improved: Writing Quality or Writing Skills? In *Impacts of Generative AI on Creativity in Higher Education*. IGI Global, 423–442.

[16] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel P. Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024). https://api.semanticscholar.org/CorpusID:267751333

[17] Andrew Dillon and Charles Watson. 1996. User analysis in HCI - the historical lessons from individual differences research. *Int. J. Hum. Comput. Stud.* 45 (1996), 619–637. https://api.semanticscholar.org/CorpusID:18034610

[18] Kevin Doherty and Gavin Doherty. 2018. The construal of experience in HCI: Understanding self-reports. *Int. J. Hum. Comput. Stud.* 110 (2018), 63–74. https://api.semanticscholar.org/CorpusID:972790

[19] Anil R Doshi and Oliver P Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science advances* 10, 28 (2024), eadn5290.

[20] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*. Vol. 44. Elsevier, 247–296.

[21] Jennifer L. Dyck and Janan Al-Awar Smither. 1996. Older adults' acquisition of word processing : the contribution of cognitive abilities and computer anxiety. *Computers in Human Behavior* 12 (1996), 107–119. https://api.semanticscholar.org/CorpusID:144065668

[22] Lauren Figueredo and Connie K Varnhagen. 2005. Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Reading Psychology* 26, 4-5 (2005), 441–458.

[23] Gerhard Fischer. 2000. User Modeling in Human Computer Interaction. https://api.semanticscholar.org/CorpusID:1526097

[24] John H. Flavell. 1979. Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. *American Psychologist* 34 (1979), 906–911. https://api.semanticscholar.org/CorpusID:8841485

[25] Dennis F Galletta, Alexandra Durcikova, Andrea Everard, and Brian M Jones. 2005. Does spell-checking software need a warning label? *Commun. ACM* 48, 7 (2005), 82–86.

[26] Sam J Gilbert, Arabella Bird, Jason M Carpenter, Stephen M Fleming, Chhavi Sachdeva, and Pei-Chun Tsai. 2020. Optimal use of reminders: Metacognition, effort, and cognitive offloading. *Journal of Experimental Psychology: General* 149, 3 (2020), 501.

[27] Juan María González-Anleo, Luca Delbello, Jose-Maria Martinez-Gonzalo, and Andres Gómez. 2024. Sociodemographic Impact on the Adoption of Emerging Technologies. *Journal of Small Business Strategy* (2024). https://api.semanticscholar.org/CorpusID:272378808

[28] Sandra Grinschgl, Frank Papenmeier, and Hauke S Meyerhoff. 2021. Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology* 74, 9 (2021), 1477–1496.

[29] Toru Ishikawa, Hiromichi Fujiwara, Osamu Imai, and Atsuyuki Okabe. 2008. Wayfinding with a GPS-based mobile navigation system: A comparison with maps and direct experience. *Journal of environmental psychology* 28, 1 (2008), 74–82.

[30] Melina Klepsch and Tina Seufert. 2020. Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science* 48 (2020), 45–77. https://api.semanticscholar.org/CorpusID:213485521

[31] Asher Koriat. 2008. Subjective confidence in one's answers: the consensuality principle. *Journal of experimental psychology. Learning, memory, and cognition* 34 4 (2008), 945–59. https://api.semanticscholar.org/CorpusID:207700747

[32] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[33] HaeJin Lee, Frank Stinar, Ruohan Zong, Hannah Valdiviejas, Dong Wang, and Nigel Bosch. 2025. Learning Behaviors Mediate the Effect of AI-powered Support for Metacognitive Calibration on Learning Outcomes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[34] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*. 1–22.

[35] Qiujie Li, Rachel Baker, and Mark Warschauer. 2020. Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education* 45 (2020), 100727.

[36] Po-Han Lin, Tzu-Chien Liu, and Fred Paas. 2017. Effects of spell checkers on English as a second language students' incidental spelling learning: a cognitive load perspective. *Reading and Writing* 30, 7 (2017), 1501–1525.

[37] Richard E. Mayer and Roxana Moreno. 2003. Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist* 38 (2003), 43 – 52. https://api.semanticscholar.org/CorpusID:13667935

[38] Hauke S Meyerhoff, Sandra Grinschgl, Frank Papenmeier, and Sam J Gilbert. 2021. Individual differences in cognitive offloading: A comparison of intention offloading, pattern copy, and short-term memory capacity. *Cognitive Research: Principles and Implications* 6, 1 (2021), 34.

[39] Christina Miles. 2008. The use or non-use of calculators affects on student's ability to perform basic mathematics problems. (2008).

[40] Kouider Mokhtari and Carla A Reichard. 2002. Assessing students' metacognitive awareness of reading strategies. *Journal of educational psychology* 94, 2 (2002), 249.

[41] Alexandra B Morrison and Lauren L Richmond. 2020. Offloading items from memory: Individual differences in cognitive offloading in a short-term memory task. *Cognitive Research: Principles and Implications* 5, 1 (2020), 1.

[42] Nicholas V. Mudrick, Roger Azevedo, and Michelle Taub. 2019. Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying multimedia learning. *Comput. Hum. Behav.* 96 (2019), 223–234. https://api.semanticscholar.org/CorpusID:150311732

[43] Thomas O. Nelson. 1990. Metamemory: A Theoretical Framework and New Findings. Psychology of Learning and Motivation, Vol. 26. Academic Press, 125–173. doi:10.1016/S0079-7421(08)60053-5

[44] Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. 2002. Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In *Nordic Conference on Human-Computer Interaction*. https://api.semanticscholar.org/CorpusID:11320034

[45] Richard E. Nisbett and Timothy D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84 (1977), 231–259. https://api.semanticscholar.org/CorpusID:7742203

[46] Anna-Marie Ortloff, Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas, and Matthew Smith. 2025. Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025). https://api.semanticscholar.org/CorpusID:278063119

[47] Delroy L. Paulhus. 1984. Two-component models of socially desirable responding. *Journal of Personality and Social Psychology* 46 (1984), 598–609. https://api.semanticscholar.org/CorpusID:144339882

[48] Ji-Lun Peng and Su-Ling Yeh. 2025. Cognitive Offloading in Short-Term Memory Tasks: Trust Toward Tools as a Moderator. *International Journal of Human–Computer Interaction* 41, 21 (2025), 13382–13393. doi:10.1080/10447318.2025.2474449

[49] Paul R Pintrich et al. 1991. A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). (1991).

[50] Jim Ranalli. 2021. L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing* 52 (2021), 100816. https://api.semanticscholar.org/CorpusID:233578537

[51] Lauren L Richmond, Lois K Burnett, Julia Kearley, Sam J Gilbert, Alexandra B Morrison, and B Hunter Ball. 2025. Individual differences in prospective and retrospective memory offloading. *Journal of Memory and Language* 142 (2025), 104617.

[52] Hazelynn Rimbar. 2017. THE INFLUENCE OF SPELL-CHECKERS ON STUDENTS'ABILITY TO GENERATE REPAIRS OF SPELLING ERRORS. *Journal of Nusantara Studies (JONUS)* 2, 1 (2017), 1–12.

[53] Evan F Risko and Timothy L Dunn. 2015. Storing information in-the-world: Metacognition and cognitive offloading in a short-term memory task. *Consciousness and cognition* 36 (2015), 61–74.

[54] Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in cognitive sciences* 20, 9 (2016), 676–688.

[55] Kenneth Ruthven. 1990. The influence of graphic calculator use on translation from graphic to symbolic forms. *Educational studies in mathematics* 21, 5 (1990), 431–450.

[56] Gavriel Salomon, David N. Perkins, and Tamar Globerson. 1991. Partners in Cognition: Extending Human Intelligence with Intelligent Technologies. *Educational Researcher* 20, 3 (1991), 2–9. doi:10.3102/0013189X020003002

[57] Gregory Schraw. 2009. A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning* 4 (2009), 33–45. https://api.semanticscholar.org/CorpusID:144787216

[58] G Schraw and RS Dennison. 1994. Assessing metacognitive awareness Contemporary Educational Psychology, 19 (4), 460-475.

[59] Judith S Shockley, Wealtha C McGurn, Carolyn Gunning, Elaine Graveley, and Delight Tillotson. 1989. Effects of calculator use on arithmetic and conceptual skills of nursing students. 402–405 pages.

[60] Betsy Sparrow, Jenny Liu, and Daniel M Wegner. 2011. Google effects on memory: Cognitive consequences of having information at our fingertips. *science* 333, 6043 (2011), 776–778.

[61] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2023. The Metacognitive Demands and Opportunities of Generative AI. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2023). https://api.semanticscholar.org/CorpusID:266359743

[62] Ali Tarhini. 2016. The Effects of Cultural dimensions and Demographic Characteristics on E-learning Acceptance. *ArXiv* abs/1607.01492 (2016). https://api.semanticscholar.org/CorpusID:24025195

[63] Keith W Thiede and John Dunlosky. 1999. Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory and Cognition* 25 (1999), 1024–1037. https://api.semanticscholar.org/CorpusID:143536719

[64] Peter M Todd and Gerd Gigerenzer. 2007. Environments that make us smart: Ecological rationality. *Current directions in psychological science* 16, 3 (2007), 167–171.

[65] Tamara van Gog and Halszka Jarodzka. 2013. Eye Tracking as a Tool to Study and Enhance Cognitive and Metacognitive Processes in Computer-Based Learning

Environments. https://api.semanticscholar.org/CorpusID:140418033
[66] Marcel V.J. Veenman, Bernadette H. A. M. van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning* 1 (2006), 3–14. https://api.semanticscholar.org/CorpusID:146690540
[67] Gerrit Cornelis Van Der Veer. 1989. Individual differences and the user interface. *Ergonomics* 32 11 (1989), 1431–1449. https://api.semanticscholar.org/CorpusID:38416716
[68] V. Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *Institutions & Transition Economics: Microeconomic Issues eJournal* (2003). https://api.semanticscholar.org/CorpusID:14435677
[69] Ashley Marie Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela J. Wisniewski. 2019. Moving beyond 'one size fits all'. *Interactions* 26 (2019), 34 – 39. https://api.semanticscholar.org/CorpusID:207906707
[70] Ping Wei, Xiaosai Wang, and Hui Dong. 2023. The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology* 14 (2023), 1249991.
[71] Philip H Winne and Dianne Jamieson-Noel. 2002. Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology* 27, 4 (2002), 551–572.
[72] Philip H Winne and Nancy E Perry. 2000. Measuring self-regulated learning. In *Handbook of self-regulation*. Elsevier, 531–566.