Quantifying the Novelty Bias when Evaluating Interactive Prototypes

Yumeng Ma
Computer Science and Engineering
University of Washington
Seattle, Washington, USA
yumengma@cs.washington.edu

Alexis Hiniker
Information School
University of Washington
Seattle, Washington, USA
alexisr@uw.edu

Jacob O. Wobbrock
University of Washington
Seattle, Washington, USA
wobbrock@uw.edu

Abstract

Experiments in human-computer interaction (HCI) often evaluate whether a prototype is "better," but novelty alone can affect users' judgments and possibly performance. To quantify this effect, we conducted a within-subjects study of 48 participants comparing four pairs of functionally identical prototypes (mice, keyboards, search engines, and AI chatbots). Each pair differed only in cosmetic features and a label marking one as "old" and the other as "new." Novelty labeling shifted preference: up to 77% favored the version labeled "new." Subjective ratings for the search engine increased under the "new" label by up to 7.1%. For the AI chatbot, ratings were driven by preference, with the preferred version rated up to 11.6% higher than the unpreferred one. Performance differences were modest and emerged for errors (e.g., 9.7% fewer misses with the "new" mouse, up to 7.2% lower error rates with the "new" keyboard). Technology readiness predicted baseline skill and occasionally moderated performance but did not protect judgments from novelty bias. These results show that novelty labeling reframes interpretation and preference more than performance, raising concerns for HCI evaluations relying on participant judgments.

CCS Concepts

• **Human-centered computing** → **Laboratory experiments**; **Empirical studies in HCI**.

Keywords

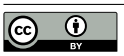
Novelty bias, Performance, Preference, Old, New, Version, Technology pairs

ACM Reference Format:
Yumeng Ma, Alexis Hiniker, and Jacob O. Wobbrock. 2026. Quantifying the Novelty Bias when Evaluating Interactive Prototypes. 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, 19 pages. https://doi.org/10.1145/3772318.3791685

1 Introduction

As new technologies rapidly emerge, researchers and practitioners regularly evaluate how users perform with and perceive them. In human-computer interaction (HCI), such evaluations anchor claims of improvement and guide design [16]. Whether assessing interaction techniques [61], adaptive interfaces [8], or intelligent systems [1], such studies ask whether a prototype is "better" than what came before, operationalized in some measurable way. HCI research that creates new inventions [40] can make participants aware—explicitly or implicitly—that one version being tested is "new" while the other is pre-existing. Signaling novelty can set expectations, alter engagement, and risk improper conclusions about new inventions. The *novelty effect* names this tendency to favor what is new over what is familiar [21]. In HCI, it has been described as a "first response to a new technology" [82]; however, responses might extend beyond qualitative impressions [14] to quantifiable preference and performance outcomes. For example, at venues like ACM UIST, where innovations are routinely compared against *de facto* baselines, it is unknown what quantifiable advantage, if any, new inventions have over their competitors *just by virtue of being new*. This risk is heightened in fields such as artificial intelligence (AI), where hype often outpaces demonstrable utility [15, 45]. At the same time, HCI has started to interrogate its own evaluation culture. Work on reproducibility [44, 49] highlights fragile findings and weak links between experimental setups and claims that travel into practice. Critiques around "innovation theater"[1] also describe pressures to showcase novelty that look impressive yet lack evidence for lasting benefit. In that landscape, evaluations that compare new systems against familiar baselines carry extra weight.

Psychology highlights cognitive mechanisms related to the novelty effect. Motivated reasoning [47] leads people to interpret evidence in line with prior beliefs. Expectation bias [88] and placebo effects [42] show that belief alone can shape outcomes. Yet in HCI, the novelty effect remains under-examined. Most HCI work has focused on demand characteristics, where participants have shown to align their behavior with perceived researcher expectations [14, 41]. Even less attention has been paid to *quantifying* the role of novelty as an isolated biasing force. Unlike traditional placebo effects that come from mechanisms such as conditioning or patient expectations [60], the novelty effect might be triggered by something as simple as a "new" label that implies superiority. This distinction motivates testing novelty via version labeling as a minimal cue. Moreover, responses to novelty can differ across individuals. Some people are more inclined to adopt, trust, or feel excited by new technologies than others, a construct called *technology readiness* (TR) [66]. Understanding how these traits affect responses to novelty is important for interpreting quantitative data in HCI. Whether higher TR makes people more susceptible to "new" labeling is unknown.



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791685

[1] https://www.innovativehumancapital.com/article/the-theater-of-innovation-why-companies-prioritize-appearances-over-innovation

To investigate the novelty effect in HCI for both performance and preference, we present a quantitative analysis of novelty framing using version labels as a placebo-like manipulation. Participants used four matched technology pairs—mice, keyboards, search engines, and AI chatbots—that were functionally identical but differed only in superficial features (color, font) and a "new" *vs.* "old" label. Input tasks (pointing, typing) yielded typical human performance metrics, while information tasks (searching, chatting) yielded participants' subjective ratings and interaction behavior. Technology order, cosmetic variant order, label assignment order, and position order were fully counterbalanced. After each comparison, participants reported a preference. We also measured participants' TR to see whether individual attitudes toward new technology influenced their performance in relation to perceived novelty.

Our findings show that novelty labeling alone drove strong preference for "new" versions across devices—up to 77% of participants favored the version labeled as "new"—and inflated subjective quality ratings in the search and AI chatbot tasks (search ratings increased by up to 7.1% for the "new" version, and preferred chatbots were rated up to 11.6% higher than unpreferred ones). By contrast, effects on human performance were smaller and uneven across measures and tasks, although they mildly affected accuracy in both pointing and text entry (the "new" mouse produced 9.7% fewer misses, and the "new" keyboard yielded up to 7.2% lower error rates). TR was associated with baseline performance and sometimes interacted with novelty, but it did not offset the dominance of the "new" label in shaping subjective ratings. Together, our results indicate that novelty shifts preference much more than performance, but neither is it entirely absent for the latter.

Our study makes three contributions to HCI research by providing: (1) a causal (experimental) demonstration that labeling a functionally identical system as "new" shifts preferences, inflates subjective ratings, and reduces input errors, (2) evidence that these perceptual shifts can outweigh or misalign with objective performance, and (3) clarification that TR influences baseline performance and can interact with novelty, but does not shield judgments from novelty-driven bias. These contributions delineate where framing—not function—drives effects, and provide a basis for study designs that separate novelty from genuine improvement.

2 Related Work

We pull on five threads of related work to inform our work. First, we examine how humans respond to innovation, especially in early encounters with new systems. Second, we look at how prototypes are typically assessed in HCI and where existing methods fall short when novelty is in play. Third, we focus on how the novelty effect has been studied in interactive systems. Fourth, we situate novelty within broader literature on placebo effects, since both describe expectation-driven shifts in evaluation when functionality is held constant. Lastly, we connect novelty framing to ongoing debates about replication, evaluation validity, and innovation theater in HCI and AI research.

2.1 Human Response to Novelty and Innovation

The human attraction to novelty is well documented by early psychology research [4, 12, 24] suggesting that this tendency is inherent. Literature in medicine further illustrates that when a treatment is presented as a novel, experimental agent, its perceived efficacy usually appears higher than when the same treatment is used as a control [3, 10, 75]. Enthusiasm for new treatments among researchers can also skew care and assessments between control and intervention groups [52].

Research in human behavior demonstrates that when faced with a novel innovation, users often react with a mixture of curiosity, excitement, and apprehension [73]. Although perceived novelty tends to elicit positive emotions such as fascination and enthusiasm [36], these initial reactions do not always correlate with sustained engagement or satisfaction [21].

Empirical studies with large participant samples have demonstrated that perceived novelty serves as an effective belief and influences the adoption of new technologies through user assessments of associated risks and rewards [85]. In addition, the perceived complexity, mysteriousness, and non-human nature of new technologies can contribute to implicit biases that affect decision-making processes [20]. Dispositional factors (e.g., optimism regarding technology's potential to provide greater control and flexibility, concerns about dependency and vulnerability to malicious activities, etc.) further influence adoption outcomes [70]. These dynamics illustrate the interplay between individual beliefs, situational factors, and perceived novelty in determining how users embrace and integrate innovations into their lives.

While users may exhibit increased interest and engagement with products they find unique or different from familiar alternatives [29], shortcomings that were previously masked by the allure of novelty may become apparent and lead to reassessment and disengagement [43]. Moreover, the reception of an innovation is susceptible to its alignment with current societal trends and endorsements from influential early adopters [2]. Therefore, innovations that challenge existing norms or require substantial changes in user behavior may encounter resistance despite offering considerable long-term benefits. User response in immediate and sustained situations is important for designing and implementing new technologies that capture initial interest and deliver lasting value.

These patterns highlight that novelty responses are shaped by situational cues and by individual dispositions. This is why we include technology readiness (TR) [67] in our study to examine whether baseline attitudes toward technology modulate how strongly participants respond to novelty.

2.2 User Assessments of Prototypes

Many HCI evaluations use a wide range of post-use self-report measures such as the Technology Acceptance Model (TAM) [13], NASA-TLX workload index [31], the System Usability Scale (SUS) [6], other Likert-type questionnaires [51], and qualitative interviews with participants. However, self-report instruments are susceptible to many different biases. Expectations can influence subjective usability and emotional responses toward systems [17, 46]. Demand characteristics may nudge participants toward what they believe researchers

want to hear. [14]. Social desirability can suppress negative feedback [25]. Acquiescence bias can disproportionally influence participants agreement with positive statements regardless of their true perceptions [58].

Some studies have turned to behavioral and physiological indicators to assess reactions to novelty. Horstmann and Herwig [37] used eye-tracking to capture fixation duration, time to first fixation, and gaze patterns as indicators of how attention shifts between familiar and unfamiliar stimuli. Poppenk et al. [69] employed recognition tasks that measured hit rates, false alarms, and source memory accuracy to assess how novelty shapes encoding and retrieval processes.

Considering both subjective and objective data can be useful when studying novelty effects. What people *say* about a system (their ratings or preferences) and what they *do* while using it (their behavior or performance) may diverge, and each reveals a different aspect of how novelty exerts influence. Our study therefore brings these together to test where novelty takes effect—whether in preference, performance, or both.

2.3 Novelty Effect in HCI

Research on novelty in HCI has examined a range of technologies, including activity trackers [77], virtual reality systems [59], large displays in work and public settings [39, 43], and gamified learning environments [72, 83]. Across these domains, engagement often spikes during early encounters and declines as the initial sense of novelty fades. Much of this work has therefore focused on long-term decline rather than on how novelty shapes first-use evaluations. Some studies have begun to address this early phase: Rutten and Geerts [74], for example, found that positive reactions to mid-air haptic feedback were strongest immediately after exposure, driven largely by perceived novelty. Dell et al. [14] showed that during initial evaluations, participants shifted their preferences to favor the option they believed was created by the researcher when both alternatives were identical. While such findings establish novelty as an influential factor in early impressions, its impact on baseline performance and preference during controlled first-use evaluations remains unexplored. Our study addresses this gap. In addition, although psychology has long documented how people respond differently when something is presented as new or improved, HCI still lacks quantitative evidence showing how this kind of framing alters interaction behavior and judgment outcomes. Our study provides this missing empirical link.

2.4 Novelty Effect as a Subset of Placebo Effect

Framing novelty as a subset of the placebo effect provides the basis for our research. HCI studies have illustrated how perceptions can strongly shape evaluations when the underlying realities are identical but labeled differently. Denisova and Cairns [15] found that basic instructions led players to believe that a game featured enhanced adaptive AI, even when no such features were present. Similarly, Pataranutaporn et al. [68] showed that altering participants' mental models of an AI system by framing it as caring, manipulative, or neutral, changed their interactions and perceptions. Thirty years ago, *The Media Equation* studies by Reeves and Nass showed, for example, that televisions showing news only were judged as more trustworthy and competent than televisions showing the same news but labeled as "generalists"[71]. These findings show how priming can lead users to perceive improvements in a system's functionality, even when those enhancements do not exist. Labeling a system "old" *vs.* "new" can likewise shift how identical systems are judged. Unlike prior work that relies on more elaborate primed features, our study isolates label-driven novelty to test its causal influence.

2.5 Replication, Evaluation Validity, and Innovation Theater

Recent methodological critiques in HCI argue that many empirical evaluations rest on limited evidence and overextended interpretations. Some studies use small samples, idiosyncratic tasks, or loosely justified measures that make it hard to determine what a finding truly establishes. For example, Ortloff et al. [63] identified how under-powered designs, inconsistent effect-size reporting, and unclear measurement-construct links lead to claims that exceed what the data can support. In human–robot interaction, Leichtmann and Rohlfing [49] show that replication attempts frequently fail because foundational constructs, such as social presence or trust, are operationalized inconsistently, making it difficult to build theory across studies.

These concerns have become more visible in AI-mediated interaction research, where evaluations rely heavily on short, laboratory-style tasks. Several recent papers show that outcomes in these settings are sensitive to framing, expectations, and study instructions over system behavior. Kosch et al. [45] demonstrate that describing a system as having sophisticated adaptive AI—despite providing no functional support—shifts users' expectations and subjective evaluations through placebo effects. Olszewski [62] found that anthropomorphic identity cues change agreement with the same AI-generated recommendations. Scholars have also critiqued the cultural forces that reflect how new technologies are presented. Granados et al. [30] show that organizations can perform "innovation theater," where high-visibility demonstrations, symbolic prototype showcases, and staged signs of progress attempt to signal technological advancement even when little substantive improvement exists. This practice reinforces the cultural narratives that newer systems are inevitably better. When evaluation studies take these narratives for granted, they can unintentionally reproduce hype, especially when a "new" version is positioned as the expected site of improvement.

When outcomes lean toward the new version, it is hard to know how much credit belongs to interaction design versus novelty framing. Our study takes this tension as a starting point and treats novelty framing as a threat to validity for evaluations that anchor narratives about innovation.

3 Experiment Design

Our study sought to examine how perceived novelty influences objective performance and subjective preference when empirically testing interactive technologies. We presented four matched technology pairs to participants: mice, keyboards, search engines, and AI chatbots. Each pair was functionally identical and only differed in

superficial cosmetic features (color, font) and, importantly, by having an "old" or "new" label. Participants engaged with all four technology pairs in structured tasks (pointing, typing, searching, chatting) designed to capture performance and preference. In addition, we measured participants' technology readiness to explore whether individual attitudes toward technology influenced responses to novelty framing.

3.1 Participants

We recruited 48 participants through on-campus and local mailing lists, flyers, community Discord servers, and word-of-mouth. We used Spiel et al.'s [81] method of surveying gender. There were 24 women, 22 men, one non-binary, and one who preferred not to disclose, aged 19 – 55 years (M = 28.5, SD = 8.4).

Our sample size aligns with prior work showing this number yields reliable performance estimates [76], and our trial-level analyses produced a total trial count (mouse: 1728; keyboard: 4800) known to provide sufficient power in input studies [80].

Participants represented a broad mix of occupational backgrounds, involving administration (e.g., leasing management, payroll), coordination (e.g., studio, operations), technical fields (e.g., engineering, information science, physical security), science (e.g., physics, astronomy, biology, biomedical informatics), education (e.g., teaching, linguistics), physical and mental health, business, military, the arts, and caregiving. For highest education level attained, of 48 participants, nine had high school diplomas, three had associate's degrees, 18 had bachelor's degrees, 13 had master's degrees, and three had doctorates or professional degrees. Each participant received $35 compensation.

Our study was reviewed and approved by the university's Institutional Review Board (IRB). All participants provided informed consent prior to participation and were told that they could withdraw at any time without penalty. At the conclusion of the study, participants were fully debriefed regarding the true purpose of the research and informed that the "old" and "new" technologies were identical apart from cosmetic differences. Consent was re-obtained following this disclosure. No participants chose to withdraw their data after debriefing. However, one participant chose not to fill out the technology readiness (TR) questionnaire due to a lack of time.

3.2 Apparatus

Study sessions were conducted in a university laboratory. We selected four everyday interactive technologies to cover a range of input and information experiences of the kind that HCI studies frequently evaluate. Mice and keyboards are low-level input devices emphasizing psychomotor control, where performance is captured by speed, accuracy, and throughput [79, 80, 93]. By comparison, search engines and AI chatbots require higher-order cognitive work than mere psychomotor control. Together the four technologies we tested span hardware and software that differ in interaction modality, user familiarity, and potential susceptibility to novelty effects.

Cosmetic differences were introduced to facilitate the experimental manipulation. Work on placebo and expectation effects shows that presentation cues sustain the belief that conditions are distinct [22]. In addition, all technologies were de-branded to avoid associations with familiar products or companies. Manufacturer logos on the mice and keyboards were covered with tape. The search engines were simplified to display only a search bar, a list of results, and minimal preview snippets under each result. The AI chatbot used a basic text interface with AI messages on the left and user messages on the right.

3.2.1 Input Devices. For our mice and keyboard pairs, we used sticky notes labeled "old" and "new" to signal novelty. These labels were counterbalanced, so half of the participants saw the black mouse labeled "old" and the blue mouse labeled "new," and half saw the opposite (Figure 1). The same was true for the black and white keyboards (Figure 2).

The two mice were *Seenda Wired RGB Backlit Optical Mice, Model MS201*. These devices differed only in their color. The two keyboards were *Macally Wireless Keyboards, Model RFJJKEY*. These devices also differed only in their color.

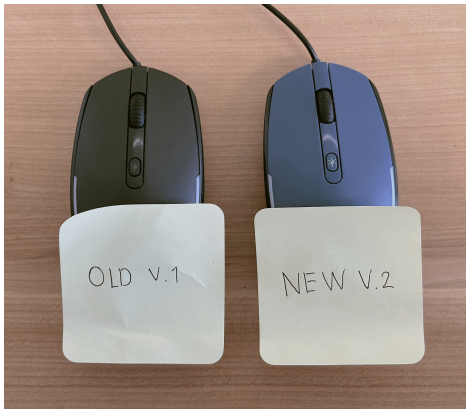


Figure 1: Two identical mice—black (*left*) and blue (*right*)—with "Old V.1" and "New V.2" labels.

Figure 2: Two identical keyboards—black (*bottom*) and white (*top*)—with "Old V.1" and "New V.2" labels.

3.2.2 Information Interfaces. The other two technologies were information interfaces, a search engine and an AI chatbot. We built stripped-down interfaces that connected to backend services and differed only in cosmetic aspects and their "Old X V.1" and "New X V.2" labels, which, like the input devices, were fully counterbalanced among participants.

Specifically, the two search engines (Figure 3) were implemented using the Bing Web Search API and differed only in font. (One was serif and the other was sans serif.) Both engines selected 10 random search terms from the same set of 200 queries sourced from Google Trends[2] in February 2025. The system was configured to automatically select and pre-fill one query into the search bar for each of 10 rounds. A 90-second timer was activated by query initiation, giving participants a comfortable time limit to peruse search results. (A green "Continue" button enabled early completion.) Pilot studies confirmed that 90 seconds provided sufficient browsing time without disengagement. Hyperlinks and typing were both disabled to shift the participants' focus purely on the search result snippets.

The two AI chatbots (Figure 4) were implemented using Together AI's API[3] and differed only in chat bubble colors and font. (One chatbot rendered AI messages in cool-toned bubbles with a serif font. The other rendered AI messages in warm-toned bubbles with a sans-serif font.) We changed both color and font for the AI chatbot to reinforce version differences, since pilot participants noted that conversational agents offered fewer persistent visual cues than the input devices and the search pages. The chatbots began each round with the prompt: *"Welcome to our conversation. I'm an AI here to chat with you. What's on your mind?"* Participants then had 90 seconds to engage in a chat of their choosing. Pilot studies of the chatbot condition led us to shorten the chatbots' responses, add short-term memory for coherence, and confirm 90 seconds per exchange.

3.3 Measures

We collected four types of measures: behavior, judgment, preference, and technology readiness (TR). These are explained below.

3.3.1 Behavior. For pointing tasks, we used FittsStudy [89, 91] (Figure 5), a testbed that evaluates pointing devices according to the ISO 9241-9 standard [18, 80]. The primary metric was throughput (TP) [54], a combined speed-accuracy measure of efficiency, computed as the effective index of difficulty divided by average movement time ($TP = ID_e/MT$) [80]. Movement time (MT) was taken per target selection from acquisition onset to selection attempt. Errors were analyzed as instances per ring-of-circles condition [89].

For typing, we used TextTest++ [92] (Figure 6), a short-phrase transcription testbed [55]. Text entry speed was measured as words per minute (WPM), with timing calculated from the first-entered to last-entered character in a phrase [53]. Accuracy was captured as both error rates and counts. Rate metrics included uncorrected error rate (UER), corrected error rate (CER), and their sum, total error rate (TER) [79, 92]. We also computed text-entry throughput (TP), a combined speed-accuracy measure of efficiency [93].

For searching, we measured search result browsing time (BT), which was the time participants spent reviewing a set of search results. We also measured rating time (RT), which was the time participants took to fill out Likert-type scales evaluating the quality of the search results.

For chatting, we used transcripts and timestamps to compute per-round user message count, per-round number of user characters sent, and rating time.

Task	Behavior Measures
Pointing 	Throughput (TP); Movement time (MT); Errors (counts per A×W condition).
Typing 	Throughput (TP); Words per minute (WPM); Uncorrected error rate (UER); Corrected error rate (CER); Total error rate (TER).
Searching 	Browsing time (BT); Rating time (RT).
Chatting 	Number of messages sent; Number of characters sent; Rating time (RT).

Table 1: Behavioral measures by task.

Table 1 summarizes the behavioral measures by task.

3.3.2 Judgment. Judgment measures captured subjective ratings of search results and AI chatbot responses using 7-point Likert scales. In each round, the order of rating dimensions was randomized to reduce sequence familiarity.

For the search engines, participants responded to the prompt: *"The search results were..."* (1 = strongly disagree, 7 = strongly agree), evaluating five dimensions: Comprehensive, Specific, Satisfactory, Trustworthy, and Relevant (Figure 7). These constructs map closely to core constructs in established evaluation models—relevance, satisfaction, and utility—as identified in Palanisamy's [64] conceptual framework for search evaluation.

For the AI chatbots, participants responded to the prompt: *"The AI's responses were..."* (1 = strongly disagree, 7 = strongly agree), evaluating nine dimensions: Clear, Empathetic, Professional, Helpful, Creative, Friendly, Relevant, Knowledgeable, and Articulate (Figure 8). Pilot studies helped us narrow an initial pool of chatbot evaluation measures [50] to nine dimensions across content fit, readability, and tone, which participants found most essential and discriminative for judging reply quality.

3.3.3 Preference. After completing both versions in a technology pair, participants indicated which version they preferred. We coded *Preferred* as a binary outcome for each technology, old or new.

3.3.4 Technology Readiness. We administered the 16-item Technology Readiness Index (TRI) [67] to account for individual differences. The TRI distinguishes between early adopters and technology-reluctant individuals [66] and comprises four constructs: Optimism (positive beliefs about technology), Innovativeness (tendency to adopt early), Discomfort (feelings of being overwhelmed), and Insecurity (distrust or skepticism).

We calculated the TR score for each participant by subtracting the average of the inhibitor dimensions (Discomfort and Insecurity) from the average of the contributor dimensions (Optimism and Innovativeness) [56]. Positive scores indicate *Technology Ready*

[2]https://trends.google.com/trends
[3]https://www.together.ai

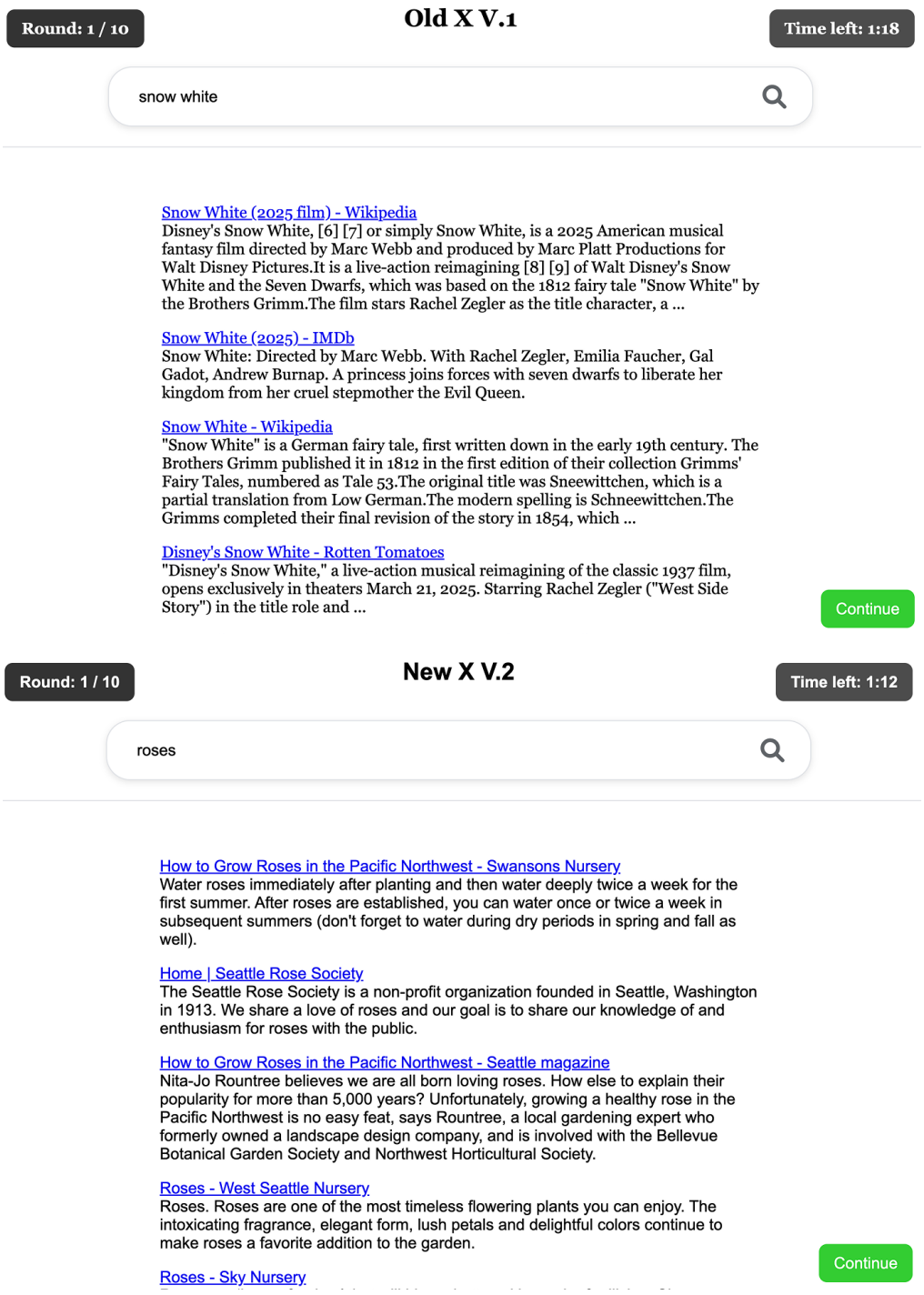


Figure 3: Two identical search engines—serif (*top*) and sans serif (*bottom*)—with on-screen labels "Old X V.1" and "New X V.2".

and negative scores indicate *Technology Not-Ready*. We obtained permission to use the full version of the scale from its author [67].

3.4 Procedure

Upon arrival, participants were informed that they would interact with two versions of each technology. Within each pair, one technology was described as an "old product" and the other as a

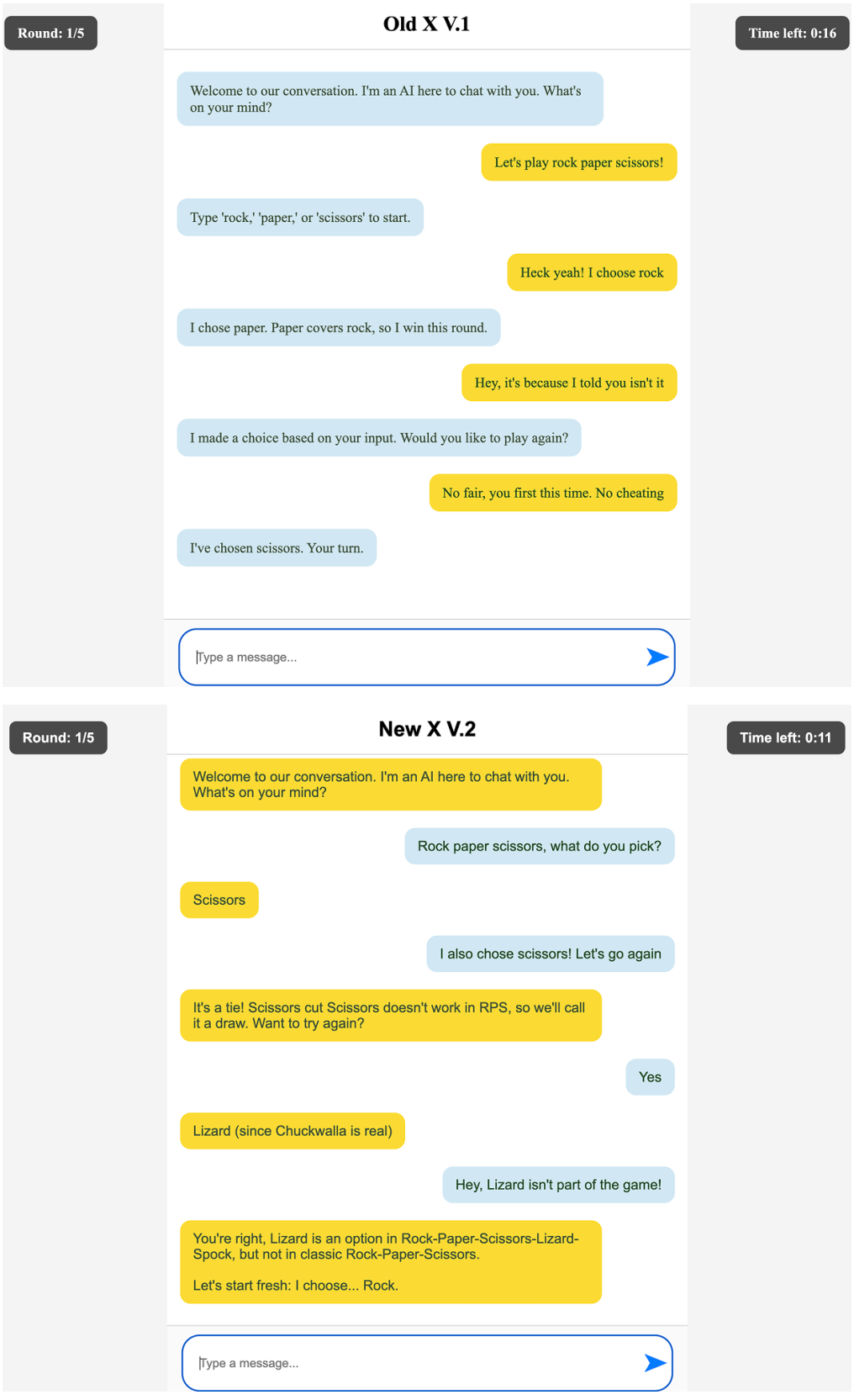


Figure 4: Two identical AI chatbots—AI messages in cool-toned bubbles with serif font (*top*) and AI messages in warm-toned bubbles with sans serif (*bottom*)—with on-screen labels "Old X V.1" and "New X V.2".

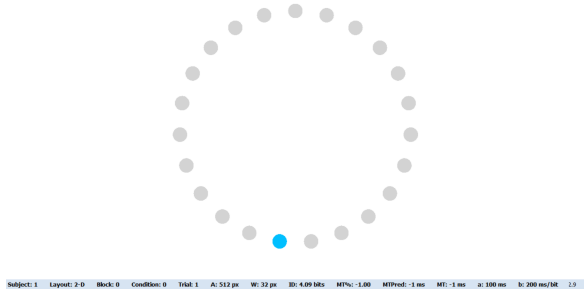


Figure 5: FittsStudy displaying a circular target layout following ISO 9241-9.

Figure 6: TextTest++ displaying a short-phrase transcription prompt.

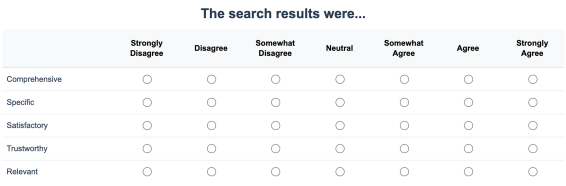


Figure 7: Likert-type scales for rating search engine results on five dimensions. Dimension order was randomized each round.

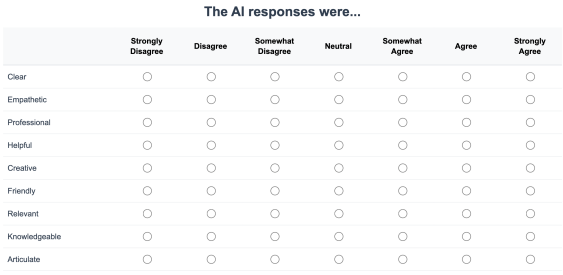


Figure 8: Likert-type scales to rate AI chatbot responses on nine dimensions. Dimension order was randomized each round.

"newly released version incorporating brand new technology." Each participant completed the study individually and filled out the TRI [67] at the end. Study sessions lasted about 80 minutes.

Participants worked with four technology types—mice, keyboards, search engines, and AI chatbots—each paired with their respective tasks: pointing, typing, searching, and chatting. Figure 9 illustrates a participant completing tasks with the version labeled "new" across all four technology pairs. Table 2 shows the corresponding protocols for each task.

As noted above, our experiment employed extensive counterbalancing to ensure order effects would not occur. We counterbalanced the following:

- *Technology order.* The four technologies (mice, keyboards, search engines, AI chatbots) appeared in all 4! = 24 orders. With $N = 48$, each order was assigned to two participants.
- *Cosmetic variant order.* For each technology, its two cosmetic variants (e.g., two colors or two fonts) were shown in both first/second orders equally often (variant A→B and B→A).
- *Label assignment order.* "New"/"old" labels were crossed with cosmetic variants. Each variant appeared as "new" for half the participants and as "old" for the other half.
- *Position order.* Position was also counterbalanced, yielding equal counts of A-new first, A–new second, A–old first, A–old second, and analogously for B.

Our counterbalancing scheme is graphically depicted in Figure 10. Statistical tests showed no effects of presentation order on any of our measures, indicating effective counterbalancing.

We used standardized prompts at five points: (1) introducing the technology pair, (2) after a brief demo, (3) before the first version, (4) before the second version, and (5) after both versions to collect preferences. Table 3 lists the exact scripts, which were followed for all participants, repeated across sessions according to the counterbalanced order.

We explained to participants that the sticky notes labeled "Old V.1" and "New V.2" for the mice and keyboards were simplified labels used in place of the actual model names. For each mouse and keyboard, participants were free to place the corresponding sticky note anywhere visible during the pointing or typing tasks. For the search engines and AI chatbots, we explained that the labels "Old X V.1" and "New X V.2" were used to simplify the actual model numbers, with X serving as a placeholder for the model name. These labels remained fixed graphically in the interface and were always visible, even when participants scrolled.

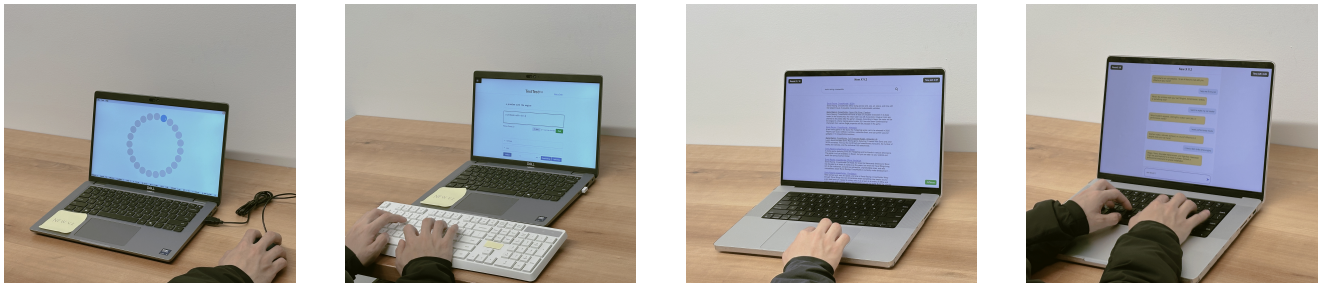
3.5 Design & Analysis

We conducted four within-subjects experiments evaluating interactions with mice, keyboards, search engines, and AI chatbots. Across all four experiments, the technologies were functionally identical.

The following factors and predictors were entered into our statistical models:

- *Version* {old, new}, the primary treatment.
- *Preferred* {0, 1}, a binary choice made after comparing both versions within each technology pair.
- *Technology Readiness (TR)*, a continuous covariate in $[-4, +4]$.

We used two interaction terms to examine *Version* further:



(a) (b) (c) (d)

Figure 9: A participant interacting with the "new" version of each technology pair: (a) pointing with the mouse, (b) typing with the keyboard, (c) reviewing search result snippets with the search engine, and (d) chatting with the AI chatbot.

Task	Protocols
Pointing	Click on targets of different sizes and distances. Timing begins only after clicking the first target. Each target is attempted once, including if missed. Perform the task quickly and accurately, but not recklessly. There are 18 conditions, each presented as a ring of circles. Within each condition, there are 23 trials (three practice trials followed by 20 test trials).
Typing	Transcribe a series of phrases. Timing begins after typing the first character and ends with the last character. The task should be completed quickly and accurately, but not recklessly. Errors near the text cursor should be corrected, others can be ignored. Press the **Enter** key to move to the next phrase. There are 50 phrases to transcribe with each keyboard.
Searching	A pre-filled search term appears in the search bar. Timing begins after clicking the search button. There are 10 searches per engine, each followed by a Likert-type survey. Each set of search results can be reviewed for up to 90 seconds. Hyperlinks are disabled. No additional input can be entered in the search bar.
Chatting	Engage in an open-ended conversation with the AI chatbot. Timing begins after typing the first character. If a topic is difficult to think of, the AI can help generate suggestions. There are five chat sessions, each followed by a Likert-type survey. Each chat session lasts 90 seconds.

Table 2: Protocols by task.

- *Version×Preference.* This term tests whether novelty labeling interacts with a participant's technology preference. For example, if novelty is influential, advantages for "new" should be larger among participants who preferred "new," and absent (or reversed) among those who preferred "old."
- *Version×TR.* This term tests whether novelty labeling interacts with a participant's TR score. For example, perhaps more tech-ready participants show stronger sensitivity to the "new" label than less tech-ready participants.

We analyzed most outcomes at the trial level to match how the measures are produced and to retain within-participant variance. A "trial" is task-specific (Table 4). For the mouse, it was one ring-of-circles condition; for the keyboard, it was one transcribed phrase; for search, it was one set of search results; and for the AI chatbot, it was one 90 second chat. *Trial* and *Participant* were modeled as random factors in all models to account for repeated measures and within-subject variance.

Two additional outcomes were participant-level and were analyzed as such:

- *Keyboard Throughput (TP)*: summarized per participant and version (per TextTest++) [93], then compared across versions using a within-subjects model.
- *Preference*: each participant made a single forced choice between the "new" and "old" versions, analyzed as a binomial test of proportions [11].

All analyses were implemented in R (v4.4.3). Performance and behavioral measures were modeled according to their statistical properties:

- *Mouse.* Throughput (TP) and movement time (MT) were analyzed using linear mixed-effects models (LMMs;

Technology	Introduction	After Demonstration	Before First Version	Before Second Version	After Both Tasks
Mouse 	*"You will be working with two different mice. The first mouse is an old mouse V.1. The second mouse is a new release V.2 that incorporates brand new technology. You will complete a series of pointing tasks with both mice."*	*"Now we are going to begin testing with the two mice. The first mouse is an old mouse V.1. The second mouse is a new release V.2 that incorporates brand new technology. You will complete a series of pointing tasks with both mice."*	*"The first mouse is an old mouse V.1. Let's begin."*	*"The second mouse is a new release V.2 that incorporates brand new technology. Let's begin."*	*"Which mouse did you prefer? Why?"*
Keyboard 	*"You will be working with two different keyboards. The first keyboard is an old keyboard V.1. The second keyboard is a new release V.2 that incorporates brand new technology. You will complete a series of typing tasks with both keyboards."*	*"Now we are going to begin testing with the two keyboards. The first keyboard is an old keyboard V.1. The second keyboard is the new release V.2 that incorporates brand new technology. You will complete a series of typing tasks with both keyboards."*	*"The first keyboard is an old keyboard V.1. Let's begin."*	*"The second keyboard is a new release V.2 that incorporates brand new technology. Let's begin."*	*"Which keyboard did you prefer? Why?"*
Search Engine 	*"You will be working with two different search engines. The first search engine is an old search engine V.1. The second search engine is a new release that incorporates brand-new technology. You will complete a series of search tasks with both search engines."*	*"Now we are going to begin testing with the two search engines. The first search engine is an old search engine V.1. The second search engine is a new release V.2 that incorporates brand new technology. You will complete a series of search tasks with both search engines."*	*"The first search engine is an old search engine V.1. Let's begin."*	*"The second search engine is a new release V.2 that incorporates brand new technology. Let's begin."*	*"Which search engine did you prefer? Why?"*
AI Chatbot 	*"You will be working with two different AI chatbots. The first AI chatbot is an old AI chatbot V.1. The second AI chatbot is a new release that incorporates brand-new technology. You will complete a series of conversing tasks with both AI chatbots."*	*"Now we are going to begin testing with the two AI chatbots. The first AI chatbot is an old AI chatbot V.1. The second AI chatbot is a new release V.2 that incorporates brand new technology. You will complete a series of conversing tasks with both AI chatbots."*	*"The first AI chatbot is an old AI chatbot V.1. Let's begin."*	*"The second AI chatbot is a new release V.2 that incorporates brand new technology. Let's begin."*	*"Which AI chatbot did you prefer? Why?"*

Table 3: Scripted instructions used during each technology session. The same scripted templates were followed for all participants, with order adjusted to reflect the counterbalancing scheme.

Technology	Trial Level
Mouse 	18 A×W conditions per mouse (36 total per participant, N = 1728 total)
Keyboard 	50 transcription phrases per keyboard (100 total per participant, N = 4800 total)
Search 	10 search result sets per engine (20 total per participant, N = 960 total).
AI Chatbot 	5 conversations per chatbot (10 total per participant, N = 480 total)

Table 4: Definition of a trial by technology.

`lme4::lmer`) [86]. Error counts showed strong overdispersion and excess zeros, and were therefore modeled using zero-inflated mixed negative binomial regression (`glmmTMB::glmmTMB`) [28, 35].

- *Keyboard.* Throughput (TP) and words per minute (WPM) were analyzed using LMMs. Uncorrected, corrected, and total error rates (UER, CER, and TER) violated normality and were analyzed using nonparametric aligned rank transform (ART) ANOVAs (`ARTool::art`) [19, 90].
- *Search engine.* Browsing time (BT) and rating time (RT) were analyzed using LMMs on log-transformed responses, which is customary for time-based measures [48]. Likert-item ratings for Comprehensive, Specific, Satisfactory, Trustworthy, and Relevant were ordinal and modeled using cumulative link mixed models (CLMMs; `ordinal::clmm`) [32].
- *AI chatbot.* Rating time (RT) was analyzed using log-transformed LMMs [48]. Message counts were analyzed with Poisson generalized linear mixed-effects models (GLMMs; `lme4::glmer`) and negative binomial GLMMs when overdispersion was present [5]. Likert-item ratings for Clear, Empathetic, Professional, Helpful, Creative, Friendly, Relevant, Knowledgeable, and Articulate were analyzed with CLMMs.

Associations between technology readiness (TR) and performance were examined using nonparametric Spearman rank correlations [78], computed separately for the "old" and "new" versions. Differences in these associations indicated where *TR* interacted with *Version*.

Binary preference choices ("old" vs. "new") were analyzed separately using exact binomial tests (`stats::binom.test`).

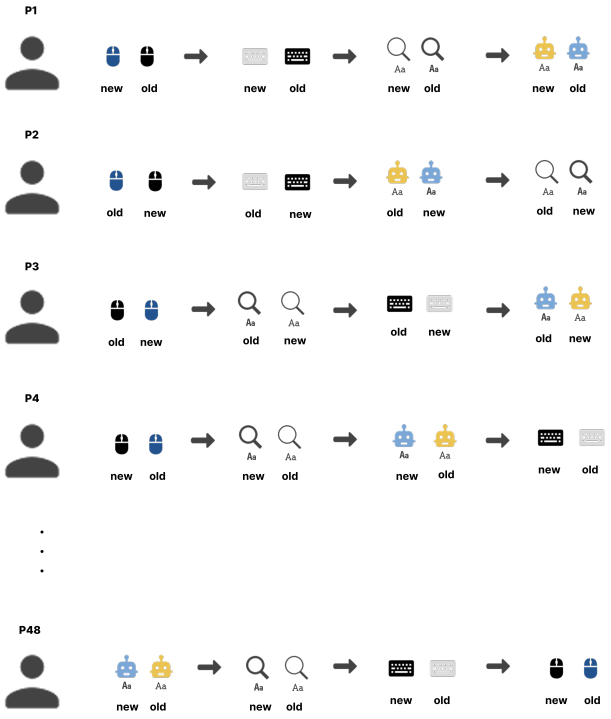


Figure 10: Counterbalancing scheme across technology order, cosmetic variant order, label assignment order, and position order.

4 Results

We report results for the four technologies in turn (mice, keyboards, search engines, AI chatbots). For each, we present behavioral outcomes, and for search and chatbot we also report subjective ratings. We then analyze stated preferences and test moderation by technology readiness (TR).

4.1 Behavioral Outcomes

4.1.1 Mouse. There were no significant effects on TP and MT. In contrast, error analysis revealed a significant main effect of *Version* on error count ($\chi^2(1) = 7.56$, $p = .006$), with participants accumulating more missed targets across trials on the "old" mouse ($M = 20.00$, $SD = 37.55$) compared to the "new" mouse ($M = 18.06$, $SD = 31.61$). There was also a main effect of *Preferred* on error count ($\chi^2(1) = 4.22$, $p = .040$) such that participants accumulated fewer missed targets on the unpreferred mouse ($M = 18.38$, $SD = 30.27$) than the preferred mouse ($M = 19.68$, $SD = 38.66$).

4.1.2 Keyboard. There were no significant effects on TP. However, the average speed of the preferred keyboard was 69.71 WPM ($SD = 25.16$), while the average speed of the unpreferred keyboard was 68.74 WPM ($SD = 24.40$). This 1.4% advantage for preferred keyboards was statistically significant ($F(1, 4701) = 5.22$, $p = .022$, $\eta_p^2 < .01$). A *Version* × *Preferred* interaction also emerged ($F(1, 46) = 4.69$, $p = .036$, $\eta_p^2 = .09$). On the "old" keyboard, participants typed faster when it was preferred at 77.03 WPM ($SD = 29.50$) than when it was unpreferred at 64.58 WPM ($SD = 20.52$). On the "new" keyboard, the pattern reversed. Participants typed faster on the unpreferred keyboard at 75.09 WPM ($SD = 28.20$) than on the preferred one at 64.91 WPM ($SD = 20.49$).

A main effect of *Version* on UER was present ($F(1, 4701) = 7.48$, $p = .006$), with higher UER on the "old" keyboard ($M = 0.83\%$, $SD = 2.21\%$) than the "new" ($M = 0.77\%$, $SD = 3.49\%$). Further, UER was higher on the unpreferred keyboard ($M = 0.86\%$, $SD = 3.11\%$) compared to preferred one ($M = 0.74\%$, $SD = 2.72\%$) ($F(1, 4701) = 12.71$, $p < .001$). *Version* had a main effect on CER, also. The average CER of the "old" keyboard was 4.52% (SD = 6.50%), while the average CER of the "new" keyboard was 4.21% (SD = 6.69%). This difference was statistically significant ($F(1, 4701) = 4.04$, $p = .045$). TER was higher on the "old" keyboard ($M = 5.35\%$, $SD = 6.79\%$) than on the "new" ($M = 4.99\%$, $SD = 7.43\%$) ($F(1, 4701) = 8.11$, $p = .004$).

4.1.3 Search Engine. For BT by *Version*, the average time was 42.59 s ($SD = 29.10$) on the "old" search engine and 42.24 s ($SD = 28.77$) on the "new" one. By *Preferred*, the averages were 42.39 s ($SD = 29.03$) for the unpreferred and 42.44 s ($SD = 28.85$) for the preferred engine. For RT, participants averaged 16.71 s ($SD = 14.81$) on the "old" engine and 17.39 s ($SD = 14.17$) on the "new" one. By *Preferred*, the average time was 17.10 s ($SD = 15.24$) for the unpreferred and 17.00 s ($SD = 13.72$) for the preferred engine. None of these differences were statistically significant. A *Version* × *Preferred* interaction was marginal $F(1, 46) = 3.11$, $p = .084$.

4.1.4 AI Chatbot. The number of user messages sent and number of characters sent were not significantly different by *Version*. There was a significant *Version* × *Preference* interaction for RT ($F(1, 46) = 4.20$, $p = .046$, $\eta_p^2 = .08$). Participants who preferred the "old" chatbot gave ratings more quickly ($M = 24.24s$, $SD = 12.73$) than when it was unpreferred ($M = 32.76s$, $SD = 18.04$), an advantage of about 8.5 s (~26%). In contrast, participants who preferred the "new" chatbot gave ratings more slowly when it was preferred ($M = 33.61s$, $SD = 21.91$) than when it was unpreferred ($M = 26.01s$, $SD = 12.55$), a disadvantage of about 7.6 s (~29%). Figure 11 shows this interaction.

4.2 Judgment Ratings

4.2.1 Search Engine. Across dimensions, participants consistently favored the "new" search engine over the "old" one. *Comprehensive*, *Specific*, *Satisfactory*, *Trustworthy*, and *Relevant* were all rated higher for the "new" version. These *Version* effects are illustrated in Figure 12 and detailed means, standard deviations, and test statistics are provided in Table 5. *Trustworthy* also revealed a *Preferred* effect, where participants rated their preferred engine as more trustworthy than their unpreferred one ($M = 5.64$, $SD = 1.25$ *vs.* $M = 5.39$, $SD = 1.33$, $\chi^2(1) = 4.05$, $p = .044$). No interaction effects were observed.

4.2.2 AI Chatbot. Across dimensions, participants consistently rated their preferred chatbot more favorably than their unpreferred one. *Clear*, *Empathetic*, *Professional*, *Helpful*, *Creative*, *Friendly*, *Relevant*, *Knowledgeable*, and *Articulate* all had higher scores for the preferred chatbot. These *Preferred* effects are illustrated in Figure 13, and detailed means, standard deviations, and test statistics are provided in Table 6. No main effects of *Version* or interactions with *Preferred* were observed.

Dimension	Old		New		$\chi^2(1)$	p
	M	***SD***	***M***	***SD***		
Comprehensive	5.37	1.42	5.74	1.26	18.03	$< .001^{***}$
Specific	5.37	1.42	5.74	1.26	10.92	$.001^{**}$
Satisfactory	5.46	1.33	5.85	1.14	18.68	$< .001^{***}$
Trustworthy	5.41	1.33	5.62	1.26	4.15	$.042^{*}$
Relevant	5.85	1.19	6.11	1.10	14.87	$< .001^{***}$

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 5: Search engine ratings by *Version*. Means (*M*), standard deviations (*SD*), and significance levels are reported.

Dimension	Unpreferred		Preferred		$\chi^2(1)$	p
	M	*SD*	*M*	*SD*		
Clear	5.23	1.47	5.99	1.02	29.45	$< .001^{***}$
Empathetic	4.25	1.50	4.73	1.48	12.58	$< .001^{***}$
Professional	5.31	1.50	5.83	1.21	31.89	$< .001^{***}$
Helpful	5.00	1.68	5.65	1.44	26.00	$< .001^{***}$
Creative	4.33	1.55	4.89	1.55	24.31	$< .001^{***}$
Friendly	4.85	1.54	5.43	1.33	26.85	$< .001^{***}$
Relevant	5.33	1.48	5.95	1.12	36.17	$< .001^{***}$
Knowledgeable	5.14	1.64	5.71	1.37	26.01	$< .001^{***}$
Articulate	5.37	1.45	5.87	1.12	34.83	$< .001^{***}$

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 6: AI Chatbot ratings by *Preference*. Means (*M*), standard deviations (*SD*), and significance levels are reported.

4.3 Stated Preference

Participants generally indicated a preference for the "new" technology versions. For the mouse, 33 of 48 participants (69%) preferred the "new" device over the "old" one, an advantage that was statistically significant ($p = .013$, 95% CI [.54, .81]). For the keyboard, 29 of 48 participants (60%) preferred the "new" version. This difference did not reach statistical significance ($p = .193$, 95% CI [.45, .74]). For the search engine, 36 of 48 participants (75%) preferred the "new" version ($p < .001$, 95% CI [.60, .86]). Preference was strongest for the AI chatbot, with 37 of 48 participants (77%) favoring the "new" version ($p < .001$, 95% CI [.63, .88]). Overall, participant preferences consistently tilted toward the "new" versions despite there being no functional difference between what was labeled "new" versus "old."

4.3.1 Counterbalancing Verification. To verify counterbalancing, we examined preference distributions by cosmetic assignment and presentation order. For the mouse, 27 participants chose blue and 21 chose black; 23 chose the first mouse they tried and 25 chose the second. For the keyboard, 29 preferred black and 19 preferred white; 25 preferred the first and 23 the second. For the search engine, preferences were split between serif (28) and sans serif (20) fonts, and between first (22) and second (26) presentation order. For the AI chatbot, preferences were divided between AI messages in cool-toned bubbles with serif font (27) and AI messages in warm-toned bubbles with sans serif font (21), and between first (23) and second (25). Exact binomial tests confirmed that none of these proportions were significantly different from 50/50 (all $p > .20$), indicating that neither cosmetic assignment nor presentation order systematically biased preferences.

4.3.2 Participant Justifications. Participants' justifications revealed three recurring themes about aesthetics, functionality, and experience.

Aesthetic Cues. Several participants based their choice on purely aesthetic reasons. P25 said they liked the "old" mouse simply *"because it was black."* P21 preferred the "old" keyboard for the same reason, remarking, *"The color is better."* P9 liked the "new" keyboard because *"it helped [them] find the right letters faster."* Fonts were also highlighted. P6 chose the "new" search engine because *"it was a sans serif font,"* and P12 felt the "new" engine *"looked more organized."* For the AI chatbot, P13 emphasized typography, favoring the "old" chatbot because *"the UI was very clean"* and the font was easier to read. The fact that some of these cosmetic features affected participants' judgments confirms the importance of having fully counterbalanced them in our study.

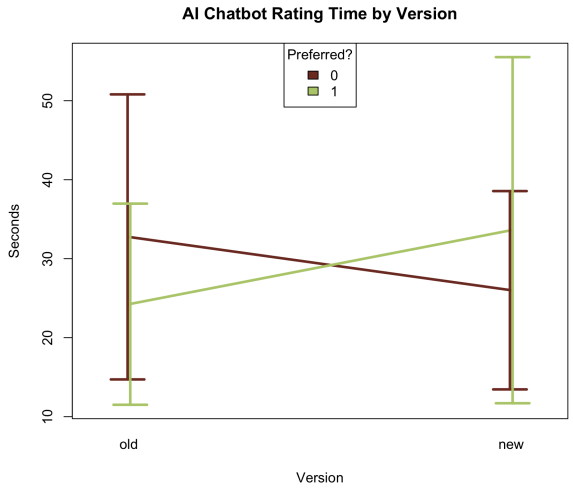


Figure 11: Interaction plot of rating time (RT) by *Version* and *Preferred*. Error bars are ±1 SD. Lower values indicate less time rating the chatbot responses.

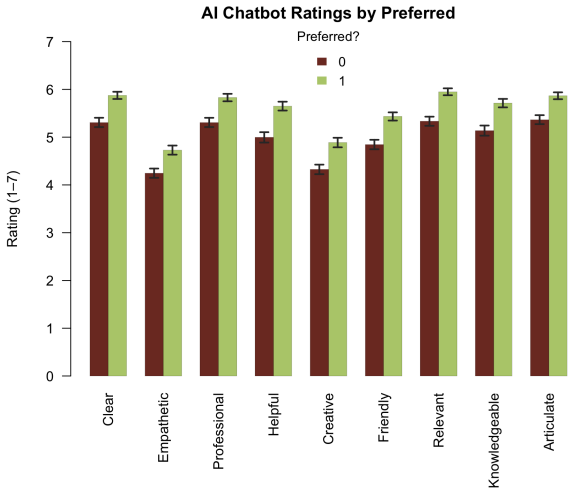


Figure 13: Mean ratings of preferred and unpreferred AI chatbots across 10 Likert items. Error bars are ±1 SE. Preferred chatbots were rated more positively across all dimensions.

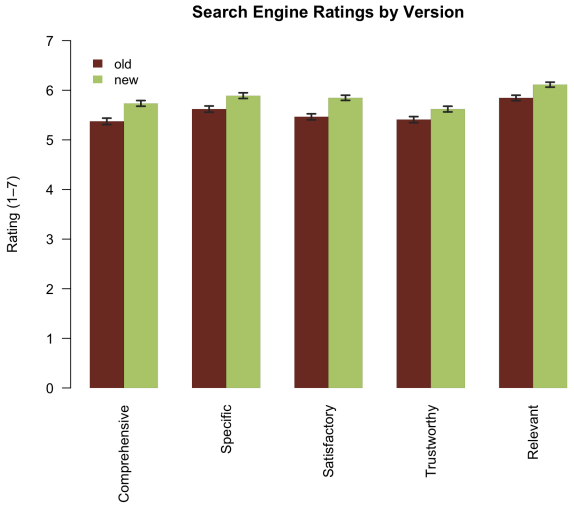


Figure 12: Mean ratings of the "old" and "new" search engines across the five dimensions. Error bars are ±1 SE. Higher values reflect more favorable evaluations.

Perceived Functionality. Some participants justified their preferences by linking them to how well they thought they performed with each version. For the mouse, P14 preferred the "old" version because they *"missed more targets with the second one [new],"* whereas P27 described the "new" mouse as *"more sensitive, more accurate."* P47 echoed this sentiment, noting the "new" mouse *"felt less jerky ... [I] didn't have to move [my] hand as much."*

Keyboard judgments were often framed in terms of control and precision. P25 described the "old" keyboard as having firmer keys and the "new" one as feeling loose, while P33 judged the "new" keyboard as *"more tactile [with] clicky feedback"* that improved typing satisfaction.

In the search engine condition, P28 favored the "old" version because it *"brings more relevant results,"* while P37 credited the "new" engine for producing *"quicker, more relevant info."* P12 tied preference explicitly to perceived success, saying they preferred the "new" search engine because they *"rated higher scores on it."*

In the AI chatbot condition, P43 praised the "new" chatbot as *"more clear, more explanation of things,"* whereas P26 thought the "new" chatbot gave *"more second- or third-layer detail"* than the "old" one.

Experiential Impressions. Other participants described their choices in terms of overall feel, comfort, or tone with no correlation to performance. For the *mouse*, some preferred the "new" version because it felt *"easier to maneuver ... a little bit lighter"* (P3) or *"fit [their] hand better"* (P8), while others leaned toward the "old" mouse, which felt *"less harsh"* (P24).

For the keyboard, P10 chose the "new" one because the keys were *"tighter ...[had] more feedback,"* whereas P31 disliked the "new" one because it *"felt wobbly, the sound was messing [them] up."* P8 noted simply that the "new" keyboard was *"softer"* to type on, while the "old" was *"louder."*

For the search engine, some participants based impressions on how the top sources looked. For example, ".org" sites and Wikipedia were seen as more credible, while outlets such as Fox News prompted skepticism. P15, for instance, favored the "new" version because Wikipedia appeared more prominently at the top across results. Impressions at times also suggested minimal distinction. P8 said they *"could not tell a big difference,"* P38 felt the two search engines were *"pretty close,"* and P13 described their choice as shaped mostly by *"familiarity"* rather than functional differences.

For the *AI chatbot*, participants' impressions were formed by the tone of the chats. P4 thought the "old" chatbot was *"a little bit more*

lighthearted." P17 described the "new" one as *"more friendly"* with *"more personality,"* and P24 preferred the "new" one because the conversation felt *"more productive,"* while the "old" seemed *"more curious."*

4.4 Technology Readiness

4.4.1 Mouse. Spearman rank correlations indicated that TP correlated positively with TR for both versions ("old": $\rho = .241$, $p < .001$; "new": $\rho = .209$, $p < .001$), meaning participants who were technology ready had higher pointing throughput. There was also a main effect of *TR* on TP ($F(1, 45) = 5.07$, $p = .029$), as well as a *Version* × *TR* interaction on TP ($F(1, 1626) = 6.02$, $p = .014$), such that the positive TR–TP association was stronger for the "old" mouse than for the "new" one. MT correlated negatively with *TR* for both versions (old: $\rho = -.087$, $p = .012$; new: $\rho = -.124$, $p < .001$), indicating that participants with higher TR generally completed movements quicker with the "new" mouse.

4.4.2 Keyboard. WPM rose with TR in both versions of the keyboard ("old": $\rho = .167$, $p < .0001$; "new": $\rho = .141$, $p < .0001$), meaning participants who were more technology ready had faster typing speeds. At the same time, UER showed small associations with *TR* that differed by version ("old": $\rho = .08$, $p < .001$; "new": $\rho = -.05$, $p = .019$). TER further reflected a positive association with *TR* in the "new" keyboard condition (Spearman $\rho = .09$, $p < .001$). Overall, higher TR was linked with faster typing but also with a greater accumulation of errors, indicating the classic speed-accuracy tradeoff in human performance.

4.4.3 Search. TR also moderated subjective evaluations of search results. For *Comprehensive*, a *Version* × *TR* interaction was significant ($\chi^2(1) = 7.46$, $p = .006$): higher TR predicted lower comprehensiveness ratings for the "old" engine ($\rho = -.11$, $p = .019$), but showed no effect for the "new" ($\rho \approx 0$, $p = .97$). For *Specific*, the interaction was again significant ($\chi^2(1) = 8.30$, $p = .004$), with TR positively associated with ratings of the "new" engine ($\rho = .125$, $p = .007$) but not the "old" ($\rho = -.045$, $p = .33$). By contrast, *Satisfactory* ($\chi^2(1) = 21.29$, $p < .0001$), *Trustworthy* ($\chi^2(1) = 6.36$, $p = .012$), and *Relevant* ($\chi^2(1) = 16.85$, $p < .001$) showed main effects of *Version* only, with the "new" engine rated more favorably than the "old".

4.4.4 AI Chatbot. RT increased with *TR* for both versions ("old": $\rho = .250$, $p < .001$; "new": $\rho = .165$, $p = .011$; $F(1, 45) = 6.85$, $p = .012$), meaning higher tech-readiness was linked to taking longer to give their ratings. For subjective ratings, there was a consistent main effect of *Version* across dimensions. Participants judged the "new" chatbot higher on being *Clear* ($\chi^2(1) = 8.29$, $p = .004$), *Empathetic* ($\chi^2(1) = 5.44$, $p = .020$), *Professional* ($\chi^2(1) = 7.03$, $p = .008$), *Helpful* ($\chi^2(1) = 9.42$, $p = .002$), *Creative* ($\chi^2(1) = 8.89$, $p = .003$), *Friendly* ($\chi^2(1) = 8.76$, $p = .003$), and *Relevant* ($\chi^2(1) = 16.85$, $p < .001$), with no significant interactions involving *TR*.

5 Discussion

Although they are numerous and various, our findings point in a coherent direction: *Novelty labels and stated preferences shifted subjective judgments more than they shifted objective performance; where objective performance was affected, it emerged in error behavior, not speed or efficiency.* As participants' own justifications made clear, their preferences were sometimes rooted in aesthetics, perceived functionality, and experiential impressions aside from performance. Across domains, versions labeled "new" were preferred (up to 77%), but the central behavioral metrics scarcely budged. TP and MT for the mouse, TP (aside from WPM with a small 1.4% tilt for preferred keyboards), BT and RT for search, and number of messages and number of characters sent for AI all showed little change.

5.1 Errors Moved While Speed Held More Steady

As noted above, the one place were behavior *did* shift due to novelty was in the reduction in *errors*—fewer missed targets with the "new" mouse, and lower text entry error rates (UER, CER, TER) with the "new" keyboard. There are two mechanisms that might explain why errors were reduced due to perceived novelty: speed-accuracy regulation and user confidence.

5.1.1 Speed-Accuracy Regulation. Performance differences across devices were not driven by changes in underlying motor or typing capacity, but by shifts in how participants balanced speed and accuracy. In decision science, this balance can be modeled as a shift in the *response criterion*: a more conservative criterion reduces errors by requiring slightly more certainty before committing to an action, whereas a liberal criterion favors speed at the cost of accuracy [33, 87]. Our results fit this account. For the mouse, TP and MT showed no detectable differences, yet participants committed fewer misses with the "new" device. For keyboards, error rates were broadly lower for the "new" device, while text entry speed showed only modest differences and even reversed depending on preference. Such patterns—stable speed but fewer errors—are hallmarks of criterion adjustments. Participants appear to have protected their overall pace while deploying subtle safeguards (e.g., slightly steadier cursor control, marginally more careful keystrokes) that reduced errors without producing visible gains. This finding matches participants' explanations, where error experience and subtle control differences were described as reasons for preferring one technology version over another.

5.1.2 User Confidence. Users' confidence with the "old" and "new" input devices might not have been equal. For example, errors with a technology labeled "old" may have been read as evidence of inadequacy, whereas the same errors with a "new" device were discounted as incidental. This asymmetry aligns with attribution theory, which shows that people selectively assign causes depending on contextual cues [57, 84]. Novelty may therefore receive the benefit of the doubt, as users attribute errors to themselves instead of to the system. Confidence has been shown to influence the accuracy of judgments [34, 65], and in motor domains, speed–accuracy tradeoffs provide a parallel account of how people regulate their execution in rapid movements [26]. For mice, preferred devices appeared to encourage riskier, tighter trajectories, producing additional misses consistent with over-confidence. For keyboards, preference was instead associated with cleaner, more decisive execution, reducing errors. In both cases, perceived novelty and preference may have recalibrated how participants treated mistakes—either as opportunities to push limits or as signals to refine their control—depending

on the geometry of the task (continuous trajectories *vs.* discrete keystrokes).

5.2 Search and Chatbot Judgments Shifted without Faster Work

In search, bowsing time (BT) was stable, but judgments favored the "new" versions, suggesting that perceived informational quality shifted without additional time taken. In search ratings, each round gave participants a fixed set of results with little room to build an ongoing sense of "fit." That might have made participants sensitive to the "new"/"old" labels when rating the results. The label sets an expectation of quality, and with no strong personal tie to the system, ratings fall in line with that expectation.

With AI chatbots, preference captured the ratings across 10 dimensions, and rating time (RT) even *increased* when the preferred "new" AI chatbot was judged—suggesting that preference can *invite more engagement*. The absence of effects on number of user messages exchanged or the number of user characters typed reinforces this. In AI chatbot ratings, preference mattered. Chatbot interactions unfold over multiple turns. Participants could see style, tone, and responsiveness accumulate. These cues create a personal sense of which chatbot feels more usable or satisfying. Participant ratings then anchor to this emerging preference, because the ongoing dialogue may have provided richer signals than the "new"/"old" labels.

These patterns suggest that judgments were shaped less by the amount of work participants put in and more by how they interpreted the task. In search, people did not need longer browsing to perceive greater quality, because informational judgments usually rely on quick cues—labels, surface features, or early impressions—that can decisively tip evaluations [27, 38]. With AI chatbots, once a version felt "better," participants expanded their appraisals across many dimensions and took longer to explain why. The added time reflected a shift in how carefully participants justified a judgment they already leaned toward. In both tasks, novelty and preference influenced the interpretive frame participants brought to the system, changing what counted as evidence of quality more than the mechanics of searching or chatting. Put simply, judgments deepened because the system felt worth evaluating, not because it required more work to use. Participants' verbal accounts similarly emphasized surface cues with search results and chatbot conversational tone, underscoring that the novelty effect operated through interpretation more than effort.

5.3 Technology Readiness as "Decorator"

We found that technology readiness (TR) acted as something of a "decorator," meaning that higher TR pointed to better baseline performance, such as higher pointing TP and text entry WPM. TR also pointed to differences across *Version* in some tasks. Participants with higher TR made fewer persistent errors on the "new" keyboard, and TR showed different patterns with search comprehensiveness and specificity. However, TR did not reduce the effect of labeling. Search engine and AI chatbot ratings were still elevated by the "new" label. In this sense, TR pointed to performance differences but left subjective judgments open to novelty-driven bias.

5.4 Implications for HCI Technology Evaluations

The overarching implication of our study is that novelty primarily *reweights participant interpretations*—how users *frame* outcomes and *justify* preferences—rather than upgrading human performance. That said, the one arena where human performance might be altered by novelty is in people's propensity to make errors. We summarize three implications for HCI experiments that compare interaction techniques or interactive systems:

- *Expect misalignment between subjective ratings and objective performance.* Inflated subjective ratings without faster behavior (search, chat), or fewer errors without higher throughputs (mouse, keyboard), show that users may feel improvement when the overall performance data do not show clear gains.
- *Preference can both help and hurt behavior.* With keyboards, preference aligned with fewer errors (confidence as "cleanliness"). With mice, preference coincided with more misses (confidence as "risk taking"). Thus, "feels better" is not a reliable proxy for "performs better," even within the same participant across trials.
- *Novelty labeling is sufficient to cause an effect.* With all else held constant, the simple labels "old" and "new" shifted judgments, confidence, and error tolerance. New HCI techniques and systems are at risk of being favorably misjudged if evaluations simply signal something as "new," perhaps even without explicit labeling. Crucially, this signaling is hard to avoid when comparing a novel invention to a recognizably *de facto* standard.

That said, these three implications describe the immediate effects of novelty framing in our data. Their significance for HCI evaluation practice also depends on (1) how novelty cues arise in typical study designs and (2) how lab-style first exposure relates to long-term use "in the wild." It may be that novelty effects wear off during prolonged use [43], highlighting the value of field deployments [7, 9, 23].

5.4.1 Novelty Cues and Best Practices. Our use of direct "old" and "new" labels did not make our evaluation unrealistic. Explicit labels heighten the clarity of the signal, but they operate on the same expectation mechanism that underlies many implicit novelty cues. In many UIST-style evaluations, participants can easily tell which version is the *de facto* and which is the "new" system without explicit labeling [14]. The structure of these studies make novelty obvious. For example, one interface is clearly presented as the experimental prototype, conditions appear in a way that highlights the contribution, or the narrative itself frames one version as the improved design. These implicit signals function much like explicit labels because participants easily know which system represents the novel concept. Our manipulation therefore acts as a controlled and transparent way to isolate the novelty signal that already exists in many HCI evaluations. Cases where novelty cues are subtle (situations where participants cannot confidently tell which version is new and must infer it from weak or ambiguous hints) fall outside the scope of our research question.

Researchers designing evaluations can take steps to make these cues less discernible, although they can be hard to fully remove.

Neutral condition naming can use balanced terms such as "A" and "B" so no version conveys an implied baseline. Symmetric introductions can give both versions the same level of detail and the same setup script, with no version presented as the system of interest. Randomized framing can assign the "new" status to different versions across participants—for example, telling half the sample that version A is the one under development and telling the other half that version B fills that role. Reduced experimenter involvement during transitions can also help—for example, on-screen instructions for switching conditions prevent researchers from giving hints through tone, gestures, or emphasis. Visual polish should be matched across conditions so participants do not pick up cues from differences in device encasing, interface layout, or prototype refinement.

Our results primarily show just how susceptible subjective ratings are to novelty cues. Evaluations relying on subjective impressions may be more vulnerable to novelty framing than those grounded in objective behavioral outcomes. Subjective ratings should be interpreted with caution and accompanied by behavioral measures, particularly in evaluations where novelty cues cannot be fully removed. Making this pairing standard practice would strengthen evaluations' reliability and support meta-level discussions about replication, contribution claims, and the evidentiary bar for system improvement in HCI.

5.4.2 Ecological Validity and First-Use Contexts. Our findings speak to early, structured encounters, the setting that underlies many controlled lab evaluations in CHI and UIST. Our participants engaged with each technology pair in short, goal-driven tasks, which mirrors the way many prototype comparisons create an immediate contrast between versions. Novelty framing pushed participants to judge the labeled "new" versions more favorably during this first exposure window. This pattern clarifies how quickly novelty expectations enter evaluation settings and how they affect the earliest interpretations researchers draw about system quality.

Outside the lab, novelty operates alongside longer-term forces. Marketing pushes the idea of improvement through launch events and announcements (e.g., Apple's annual iPhone releases), peer influence reinforces desirability when people see others adopt the latest device, and accumulated experience can shift early excitement toward more tempered judgments. Prior work has documented similar early surges in interest and later declines [43]. Our findings complement this by capturing the very first comparison point—before habits, fatigue, or social pressures enter the picture. This first-use moment matters for HCI because many evaluations rely on short, single-session encounters with a new technique or system.

At the same time, these observations raise concerns for evaluations that expect short sessions to speak for longer-term use. When novelty labels push ratings upward while interaction quality stays the same, early results may not hold once the excitement of "newness" fades. Many studies, especially prototype comparisons, tutorials, and demo-style evaluations, depend on first impressions. If novelty cues remain visible, those first impressions can give an inflated picture of how strong a system actually is. Acknowledging this gap helps researchers present the reach of their findings more carefully and helps reviewers read early-stage evidence with the right level of caution.

5.5 Limitations and Future Work

This work examined novelty effects under highly controlled conditions with four technologies: keyboards, mice, search engines, and AI chatbots. Technologies were stripped of brand identity, reduced to core functionality, and distinguished only by cosmetic features and labels, all fully counterbalanced. This control was necessary to isolate novelty labels, but it also means that our results speak most directly to early, structured encounters of the kind used in many CHI and UIST evaluations. Our design captures the opening moment of comparison, not the dynamics that unfold as people continue to use a system. Future studies could examine richer or higher-stakes technologies—such as productivity suites, AR/VR systems, or AI-mediated tools—where novelty expectations may be stronger and carry greater consequences than with familiar, conventional technologies such as the ones we tested. Studies that rely on higher-level tasks like planning, collaboration, creative work, or multi-step workflows may reveal different sensitivities to novelty.

Our experiments adopt a deliberately constrained setting. Participants encountered each technology pair in a single lab session, with focused tasks and clear timing. This format fits many controlled studies, yet it does not present contexts where novelty builds through social influence, marketing, or extended use. More widely, novelty can gain value through social exposure and uncertainty. Outside technology, consumer trends like Labubu and blind boxes show how anticipation, peer visibility, and short-lived excitement can increase desirability—mechanisms that our lab setting did not model. Field deployments or longer-term use introduce forces our design did not capture, including support channels, community norms, and evolving expectations. Those settings may dampen framing effects through habituation or amplify them through commitment dynamics. Future work using longitudinal or in-the-wild methods would clarify whether novelty is a fleeting bias or a persistent factor in evaluation.

A further consideration involves how short-term performance unfolds within a session. Our design minimized systematic learning differences across labels due to counterbalancing and brief tasks, but future work should analyze how performance changes over time to determine whether novelty raises initial performance or alters how performance evolves. Such analyses, especially when extended into longitudinal or repeated-use settings, would help determine whether the level differences observed at first exposure persist, shrink, or reverse as people continue to use the two labeled variants across time.

Another limitation is the form of our manipulation. Sticky-note labels and scripted introductions signal the novel version more prominently in degree than what many evaluations use. At the same time, as discussed earlier, newness is usually easy to spot in practice—a prototype is introduced as the contribution, appears in a manner that highlights it, or has a lack of visual polish that signals its role. We do not claim that our study covers every such situation, but that it treats explicit labeling as a boundary case to make the novelty signal visible and measurable. We acknowledge that many real evaluations rely on cues that are less overt yet still noticeable. Future work can benchmark these signals, such as condition names or social endorsements, to determine whether they produce similar or weaker shifts in evaluation. It may also help to explore designs

that vary expectations more gradually. For example, studies that work within a single technology domain could support manipulations of expectation and incorporate measures of participants' beliefs before and after exposure. In addition, although cosmetic features were fully counterbalanced, inherent preferences for certain colors or fonts may have influenced judgments independently of labeling. Future studies should better disentangle visual preference from novelty framing.

Finally, our study prioritized breadth across four technology types rather than going into greater depth with any one system. This breadth demonstrated the robustness of novelty framing across both hardware and software, but left open finer-grained analyses within each domain (e.g., kinematic analyses of mouse trajectories, conversational quality in AI chatbots). Future studies could concentrate on a single technology class to uncover domain-specific mechanisms in more detail.

6 Conclusion

As the human-computer interaction (HCI) community continues to rely on user studies to evaluate new systems, it is important to recognize how novelty framing can effect performance and preference. This paper brings the phenomenon of label-driven novelty bias to the attention of the HCI community and examines its effects across hardware and software technologies. Through controlled experiments with 48 participants, we showed that: (1) participants consistently preferred and rated higher the versions labeled as "new," even though the technologies were functionally identical, (2) these appraisal effects extended across multiple dimensions in judgment tasks, (3) while small reductions in errors appeared for "new" input devices, other aspects of skill—efficiency (throughput), browsing time, number of user characters typed, and number of messages sent—remained stable, and (4) willingness to adopt new technologies (technology readiness) was linked to stronger baseline performance but did not diminish the overall novelty bias.

The influence of novelty labeling in HCI evaluations has been acknowledged [14] but not, until now, quantified. Our results make clear that novelty does not enhance performance much but it does alter preference and perspective a great deal. Recent discussions on replication, evaluation validity, and hype show why accounting for novelty bias is increasingly important when interpreting early study results. For HCI, this raises a methodological challenge. Evaluation practices that rely on short-term preference or subjective appraisal risk mistaking framing effects for genuine progress. If our field is to draw reliable conclusions about what counts as "better," we should design studies that can separate real advances in interaction from the premise and allure of novelty.

Acknowledgments

The authors thank Ed Cutrell for conversations related to this work. This work was supported in part by a National Science Foundation Graduate Research Fellowship, and by Baidu, Microsoft, and the University of Washington Information School. Any opinions, findings, conclusions or recommendations expressed in our work are those of the authors and do not necessarily reflect those of any supporter.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Richard P Bagozzi. 2007. The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the association for information systems* 8, 4 (2007), 3.

[3] Corrado Barbui, Andrea Cipriani, Paolo Brambilla, and Matthew Hotopf. 2004. " Wish bias" in antidepressant drug trials? *Journal of clinical psychopharmacology* 24, 2 (2004), 126–130.

[4] Daniel E Berlyne. 1950. Novelty and curiosity as determinants of exploratory behaviour. *British journal of psychology* 41, 1 (1950), 68.

[5] Norman E Breslow and David G Clayton. 1993. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* 88, 421 (1993), 9–25.

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] Barry Brown, Stuart Reeves, and Scott Sherwood. 2011. Into the wild: challenges and opportunities for field trial methods. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1657–1666.

[8] Daniel Buschek and Florian Alt. 2021. Building Adaptive Touch Interfaces—Case Study 6. In *Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice*. 379–406.

[9] Parmit K Chilana, Amy J Ko, Jacob O Wobbrock, and Tovi Grossman. 2013. A multi-site field study of crowdsourced contextual help: usage and perspectives of end users and software teams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 217–226.

[10] Andrea Cipriani, Toshi A Furukawa, Georgia Salanti, Anna Chaimani, Lauren Z Atkinson, Yusuke Ogawa, Stefan Leucht, Henricus G Ruhe, Erick H Turner, Julian PT Higgins, et al. 2018. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet* 391, 10128 (2018), 1357–1366.

[11] Charles J Clopper and Egon S Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 4 (1934), 404–413.

[12] Leslie B Cohen, Eric R Gelber, and Marilee A Lazar. 1971. Infant habituation and generalization to differing degrees of stimulus novelty. *Journal of Experimental Child Psychology* 11, 3 (1971), 379–389.

[13] Fred D Davis et al. 1989. Technology acceptance model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption* 205, 219 (1989), 5.

[14] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. " Yours is better!" participant response bias in HCI. In *Proceedings of the sigchi conference on human factors in computing systems*. 1321–1330.

[15] Alena Denisova and Paul Cairns. 2015. The placebo effect in digital games: Phantom perception of adaptive artificial intelligence. In *Proceedings of the 2015 annual symposium on computer-human interaction in play*. 23–33.

[16] Alan Dix. 2009. Human-computer interaction. In *Encyclopedia of database systems*. Springer, 1327–1331.

[17] Onur Dönmez, Yavuz Akbulut, Gözde Zabzun, and Berrin Köseoğlu. 2025. Effects of survey order on subjective measures of cognitive load: A randomized controlled trial. *Applied Cognitive Psychology* 39, 2 (2025), e70039.

[18] Sarah A Douglas, Arthur E Kirkpatrick, and I Scott MacKenzie. 1999. Testing pointing device performance and user assessment with the ISO 9241, Part 9 standard. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 215–222.

[19] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An aligned rank transform procedure for multifactor contrast tests. In *The 34th annual ACM symposium on user interface software and technology*. 754–768.

[20] Kimberly D Elsbach and Ileana Stigliani. 2019. New information technology and implicit bias. *Academy of Management Perspectives* 33, 2 (2019), 185–206.

[21] Dirk M Elston. 2021. The novelty effect. *Journal of the American Academy of Dermatology* 85, 3 (2021), 565–566.

[22] Laura Enax, Bernd Weber, Maren Ahlers, Ulrike Kaiser, Katharina Diethelm, Dominik Holtkamp, Ulya Faupel, Hartmut H Holzmüller, and Mathilde Kersting. 2015. Food packaging cues influence taste perception and increase effort provision for a recommended snack product in children. *Frontiers in Psychology* 6 (2015), 882.

[23] Abigail Evans and Jacob Wobbrock. 2012. Taming wild behavior: the input observer for obtaining text entry and mouse pointing measures from everyday computer use. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1947–1956.

[24] Robert L Fantz. 1964. Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science* 146, 3644 (1964), 668–670.

[25] Robert J Fisher. 1993. Social desirability bias and the validity of indirect questioning. *Journal of consumer research* 20, 2 (1993), 303–315.

[26] Paul M Fitts. 1992. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology: General* 121, 3 (1992), 262.

[27] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*. 1–15.

[28] William Gardner, Edward P Mulvey, and Esther C Shaw. 1995. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological bulletin* 118, 3 (1995), 392.

[29] John T Gourville. 2006. Eager sellers and stony buyers: Understanding the psychology of new-product adoption. *Harvard business review* 84, 6 (2006), 98–106.

[30] Cristian Granados, Yarid Ayala, and Monica Ramos-Mejia. 2024. Is it substantive or just symbolic? Understanding innovation theater in organisations: The case of technology-based innovation. *Technovation* 129 (2024), 102880.

[31] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[32] Donald Hedeker and Robert D Gibbons. 1994. A random-effects ordinal regression model for multilevel analysis. *Biometrics* (1994), 933–944.

[33] Richard P Heitz. 2014. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience* 8 (2014), 150.

[34] Vivian Allen Charles Henmon. 1911. The relation of the time of a judgment to its accuracy. *Psychological review* 18, 3 (1911), 186.

[35] Joseph M Hilbe. 2011. *Negative binomial regression*. Cambridge University Press.

[36] Elizabeth C Hirschman. 1980. Innovativeness, novelty seeking, and consumer creativity. *Journal of consumer research* 7, 3 (1980), 283–295.

[37] Gernot Horstmann and Arvid Herwig. 2016. Novelty biases attention and gaze in a surprise trial. *Attention, Perception, & Psychophysics* 78 (2016), 69–77.

[38] Dongchen Huang, Yige Zhu, and Eni Mustafaraj. 2019. How Dependable are" First Impressions" to Distinguish between Real and Fake NewsWebsites?. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. 201–210.

[39] Wen Huang. 2020. *Investigating the novelty effect in virtual reality on stem learning*. Ph. D. Dissertation. Arizona State University.

[40] Scott E Hudson and Jennifer Mankoff. 2014. Concepts, values, and methods for technical human–computer interaction research. In *Ways of Knowing in HCI*. Springer, 69–93.

[41] Olga Iarygina, Kasper Hornbæk, and Aske Mottelson. 2025. Demand characteristics in human–computer experiments. *International Journal of Human-Computer Studies* 193 (2025), 103379.

[42] Ted J Kaptchuk and Franklin G Miller. 2015. Placebo effects in medicine. *New England Journal of Medicine* 373, 1 (2015), 8–9.

[43] Michael Koch, Kai von Luck, Jan Schwarzer, and Susanne Draheim. 2018. The novelty effect in large display deployments–Experiences and lessons-learned for evaluating prototypes. In *Proceedings of 16th European conference on computer-supported cooperative work-exploratory papers*. European Society for Socially Embedded Technologies (EUSSET).

[44] Thomas Kosch and Sebastian Feger. 2024. Risk or chance? Large language models and reproducibility in HCI research. *Interactions* 31, 6 (2024), 44–49.

[45] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The placebo effect of artificial intelligence in human–computer interaction. *ACM Transactions on Computer-Human Interaction* 29, 6 (2023), 1–32.

[46] Sari Kujala, Ruth Mugge, and Talya Miron-Shatz. 2017. The role of expectations in service evaluation: A longitudinal study of a proximity mobile payment service. *International Journal of Human-Computer Studies* 98 (2017), 51–61.

[47] Ziva Kunda. 1990. The case for motivated reasoning. *Psychological bulletin* 108, 3 (1990), 480.

[48] Raymond J Lawrence. 2018. The lognormal as event-time distribution. In *Log-normal Distributions*. Routledge, 211–228.

[49] Benedikt Leichtmann, Verena Nitsch, and Martina Mara. 2022. Crisis ahead? Why human-robot interaction user studies may have replicability problems and directions for improvement. *Frontiers in Robotics and AI* 9 (2022), 838116.

[50] Hongru Liang and Huaqing Li. 2021. Towards standard criteria for human evaluation of chatbots: A survey. *arXiv preprint arXiv:2105.11197* (2021).

[51] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[52] Yan Luo, Carl Heneghan, and Nav Persaud. 2023. Catalogue of bias: novelty bias. *BMJ Evidence-Based Medicine* 28, 6 (2023), 410–411.

[53] I Scott MacKenzie. 2002. A note on calculating text entry speed. *Unpublished work. Available online at http://www. yorku. ca/mack/RN-TextEntrySpeed. html* (2002).

[54] I. Scott MacKenzie and Poika Isokoski. 2008. Fitts' throughput and the speed-accuracy tradeoff. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1633–1636. doi:10.1145/1357054.1357308

[55] I Scott MacKenzie and R William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI'03 extended abstracts on Human factors in computing systems*. 754–755.

[56] Wyatt MacNevin, Eric Poon, and Thomas A Skinner. 2021. Technology readiness of medical students and the association of technology readiness with specialty interest. *Canadian medical education journal* 12, 2 (2021), e31–e41.

[57] Bertram F Malle. 2006. The actor-observer asymmetry in attribution: a (surprising) meta-analysis. *Psychological bulletin* 132, 6 (2006), 895.

[58] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437.

[59] Ines Miguel-Alonso, Bruno Rodriguez-Garcia, David Checa, and Andres Bustillo. 2023. Countering the novelty effect: a tutorial for immersive virtual reality learning environments. *Applied Sciences* 13, 1 (2023), 593.

[60] Swapna Munnangi, Joshua Henrina Sundjaja, Karampal Singh, Anterpreet Dua, and Lambros D Angus. 2018. Placebo effect. (2018).

[61] Brad A. Myers. 2024. *Pick, Click, Flick! The Story of Interaction Techniques* (1 ed.). Vol. 57. Association for Computing Machinery, New York, NY, USA.

[62] Samuel Aleksander Sánchez Olszewski. 2024. Designing human-ai systems: Anthropomorphism and framing bias on human-ai collaboration. *arXiv preprint arXiv:2404.00634* (2024).

[63] 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. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.

[64] Ramaraj Palanisamy. 2013. Evaluation of search engines: a conceptual model and research issues. *International Journal of Business and Management* 8, 6 (2013), 1.

[65] Gerry Pallier, Rebecca Wilkinson, Vanessa Danthiir, Sabina Kleitman, Goran Knezevic, Lazar Stankov, and Richard D Roberts. 2002. The role of individual differences in the accuracy of confidence judgments. *The Journal of general psychology* 129, 3 (2002), 257–299.

[66] Ananthanarayanan Parasuraman. 2000. Technology Readiness Index (TRI) a multiple-item scale to measure readiness to embrace new technologies. *Journal of service research* 2, 4 (2000), 307–320.

[67] Ananthanarayanan Parasuraman and Charles L Colby. 2015. An updated and streamlined technology readiness index: TRI 2.0. *Journal of service research* 18, 1 (2015), 59–74.

[68] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence* 5, 10 (2023), 1076–1086.

[69] Jordan Poppenk, Stefan Köhler, and Morris Moscovitch. 2010. Revisiting the novelty effect: when familiarity, not novelty, enhances memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36, 5 (2010), 1321.

[70] Mark Ratchford and Michelle Barnhart. 2012. Development and validation of the technology adoption propensity (TAP) index. *Journal of Business Research* 65, 8 (2012), 1209–1215.

[71] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996), 19–36.

[72] Luiz Rodrigues, Filipe D Pereira, Armando M Toda, Paula T Palomino, Marcela Pessoa, Leandro Silva Galvão Carvalho, David Fernandes, Elaine HT Oliveira, Alexandra I Cristea, and Seiji Isotani. 2022. Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study. *International Journal of Educational Technology in Higher Education* 19, 1 (2022), 13.

[73] Everett M Rogers, Arvind Singhal, and Margaret M Quinlan. 2014. Diffusion of innovations. In *An integrated approach to communication theory and research*. Routledge, 432–448.

[74] Isa Rutten and David Geerts. 2020. Better because it's new: The impact of perceived novelty on the added value of mid-air haptic feedback. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[75] Georgia Salanti, Sofia Dias, Nicky J Welton, AE Ades, Vassilis Golfinopoulos, Maria Kyrgiou, Davide Mauri, and John PA Ioannidis. 2010. Evaluating novel agent effects in multiple-treatments meta-regression. *Statistics in medicine* 29, 23 (2010), 2369–2383.

[76] Jeff Sauro and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

[77] Grace Shin, Yuanyuan Feng, Mohammad Hossein Jarrahi, and Nicci Gafinowitz. 2019. Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use. *JAMIA open* 2, 1 (2019), 62–72.

[78] Sidney Siegel. 1957. Nonparametric statistics. *The American Statistician* 11, 3 (1957), 13–19.

[79] R William Soukoreff and I Scott MacKenzie. 2003. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 113–120.

[80] R William Soukoreff and I Scott MacKenzie. 2004. Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *International journal of human-computer studies* 61, 6 (2004), 751–789.

[81] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. 2019. How to do better with gender on surveys: a guide for HCI researchers. *Interactions* 26, 4 (June

2019), 62–65. doi:10.1145/3338283

[82] JaYoung Sung, Henrik I Christensen, and Rebecca E Grinter. 2009. Robots in the wild: understanding long-term use. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 45–52.

[83] Crystal Han-Huei Tsay, Alexander K Kofinas, Smita K Trivedi, and Yang Yang. 2020. Overcoming the novelty effect in online gamified learning systems: An empirical evaluation of student engagement and performance. *Journal of Computer Assisted Learning* 36, 2 (2020), 128–146.

[84] Bernard Weiner. 1985. An attributional theory of achievement motivation and emotion. *Psychological review* 92, 4 (1985), 548.

[85] John D Wells, Damon E Campbell, Joseph S Valacich, and Mauricio Featherman. 2010. The effect of perceived novelty on the adoption of information technology innovations: a risk/reward perspective. *Decision Sciences* 41, 4 (2010), 813–843.

[86] Brady T West, Kathleen B Welch, and Andrzej T Galecki. 2022. *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC.

[87] Wayne A Wickelgren. 1977. Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica* 41, 1 (1977), 67–85.

[88] Janet B Williams, D Popp, KA Kobak, and MJ Detke. 2012. P-640-The power of expectation bias. *European Psychiatry* 27, S1 (2012), 1–1.

[89] Jacob O Wobbrock, Edward Cutrell, Susumu Harada, and I Scott MacKenzie. 2008. An error model for pointing based on Fitts' law. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1613–1622.

[90] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963

[91] Jacob O. Wobbrock, Kristen Shinohara, and Alex Jansen. 2011. The effects of task dimensionality, endpoint deviation, throughput calculation, and experiment design on pointing measures and models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1639–1648. doi:10.1145/1978942.1979181

[92] Mingrui Ray Zhang and Jacob O Wobbrock. 2019. Beyond the input stream: Making text entry evaluations more flexible with transcription sequences. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 831–842.

[93] Mingrui Ray Zhang, Shumin Zhai, and Jacob O Wobbrock. 2019. Text entry throughput: Towards unifying speed and accuracy in a single performance metric. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.