



Diagnosing Usability Barriers and Conversion Quality Deficits: An Integrated Framework for Web Interface Evaluation

Dr. Joel Rodriguez

JR Training Design and Consulting, United States

ARTICLE INFO

KEYWORDS

heuristic evaluation; conversion rate optimization; LIFT Model; Nielsen's heuristics; usability; digital interface evaluation; user satisfaction; accessibility guidelines

HOW TO CITE

Rodriguez, D. J. (2026). Diagnosing Usability Barriers and Conversion Quality Deficits: An Integrated Framework for Web Interface Evaluation. International Journal of Automation and Digital Transformation, 5(2), 78-104.

<https://doi.org/10.54878/d75ny149>

ABSTRACT

Usability evaluation and conversion rate optimization examine the same interfaces through different methods, vocabularies, and specialists. This study integrates them. The app, referred to as App X, a standardized criterion-referenced protocol, scores a website against Nielsen's ten usability heuristics and the six factors of the LIFT conversion model in a single automated pass. The instrument was deployed publicly and produced 50 evaluations across 27 unique websites, paired with a 25-respondent satisfaction and acceptance survey. Usability fell in the good range ($M = 72.24$, $SD = 3.51$) and conversion quality likewise ($M = 72.0$, $SD = 5.90$); the composite Overall UX Score was invariant across three heuristic weighting schemes (maximum spread 0 points) and the Conversion Quality composite varied by at most 1 point across three factor weighting schemes. Diagnostic discrimination appeared at the construct level rather than the site level: the ten heuristics spanned 19 points ($M = 69.3$, $SD = 5.24$) and the six LIFT factors 13 points ($M = 71.0$, $SD = 4.20$), with the weakest dimensions - help and documentation (59), visibility of system status (62), clarity (65), and low anxiety (67) - forming a coherent, literature-consistent profile. Usability and conversion quality were weakly to moderately associated ($r = 0.40$, $r^2 \approx 0.16$), indicating complementary rather than redundant diagnostic coverage. Two findings bound every claim: the deterministic modeled scoring engine produced all 50 evaluations rather than the AI pipeline, and six repeat audited sites supplied 29 of the 50 evaluations. The contribution is therefore methodological - a reproducible, construct-grounded instrument answering a documented call for standardization -with criterion validation, AI-pipeline validation, and remediation effect testing reserved for future work.



Double-blind peer review by independent reviewers.

Received: 04 Jul 2026, Accepted: 15 Nov 2026, Published: 05 Dec 2026

*Corresponding author: rodriguezjoel@hotmail.com

Copyright: © 2026 by the author. Licensee Emirates Scholar Center for Research & Studies, UAE.

Open access under the Creative Commons Attribution (CC BY) license

<https://creativecommons.org/licenses/by/4.0>

Introduction

Every click on a website reflects a user decision: to continue engaging or to exit. When users encounter confusing navigation, unclear calls to action, slow performance, or frustrating interfaces, they are more likely to abandon the site before completing key actions such as purchases, registrations, donations, or inquiries [1]. Even minor usability barriers reduce conversion rates, constrain organizational growth, and undermine user trust and engagement [2].

In the digital economy, an organization's website serves as its primary storefront, communication channel, and service-delivery platform. For many businesses, nonprofits, and institutions, the website is the primary point of contact and often the only opportunity to convert visitors into customers, members, donors, or clients. Despite substantial investment in digital marketing and website development, many organizations continue to experience suboptimal conversion rates because usability barriers go undetected or unaddressed. These barriers prevent users from completing intended tasks, diminish trust, and lead to lost revenue and reduced engagement [3].

Usability is a foundational determinant of website effectiveness. Nielsen's ten usability heuristics [4], [5] have long served as a practical, expert-based framework for identifying interface problems. Heuristic evaluation, a widely used inspection method, enables systematic identification of issues related to visibility of system status, match between the system and the real world, user control and freedom, consistency and standards, error prevention, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design, error recognition and recovery, and help and documentation [6]. While heuristic evaluation is well established in human-computer interaction (HCI) research, few studies have empirically linked specific heuristic violations to measurable business outcomes, particularly conversion rates. Website prototypes, meanwhile, remain essential to user-centered design: they support early user testing, improve stakeholder collaboration, reduce development costs, and enhance digital experiences by providing interactive representations of a site's structure and behavior [7]. Most existing research has either identified usability issues in isolation or optimized conversion through A/B testing and analytics, with limited integration of structured heuristic evaluation and prototyping.

This gap is significant because usability issues are not merely inconveniences; they introduce friction that

directly impedes progression toward conversion. When users encounter confusing navigation, unclear calls to action, inconsistent layouts, or accessibility barriers, they often abandon tasks before completing them [1]. Understanding the relationship between heuristic-identified usability barriers and conversion behavior is therefore both a theoretical and a practical priority.

The present study addresses that gap by developing and applying App X, a standardized automated instrument that evaluates a live website against Nielsen's ten heuristics [4], [5] and the six factors of the LIFT conversion model in one integrated pass. Rather than treating usability and conversion as separate concerns, App X scores both on the same interface and pairs each report with a short satisfaction survey. Established work on prototyping and design thinking workflows informs how its diagnostic output is organized for iterative redesign [7]. The instrument was deployed to a distributed participant sample, yielding 50 evaluations across 27 websites drawn from a wide range of industries.

Hypothesis (long-term program goal). This research forms the first stage of a larger program whose guiding hypothesis is causal: systematically remediating the usability barriers a structured heuristic evaluation identifies, using Nielsen's ten heuristics, will meaningfully improve holistic website performance, specifically a measurable, practically meaningful increase in conversion rate alongside a user satisfaction score of 70% or greater. Testing that claim demands a pre-post remediation design in which a site is evaluated, changed, and re-measured; it is therefore the goal of the broader program rather than a claim tested here.

Research question (this study). The present study addresses the diagnostic precondition: to what extent can an integrated heuristic and conversion protocol, delivered through a single standardized automated instrument, reliably identify and prioritize the usability and conversion barriers on a live website, and to what extent do those two classes of weakness coincide on the same interface? The emphasis falls on reliability, prioritization, and co-occurrence rather than on a measured change in conversion. Establishing this diagnostic foundation is what makes the later causal trial possible.

The study is timely for three reasons. First, it bridges user-centered design principles with measurable organizational performance metrics, moving beyond subjective usability assessment toward evidence-based connections with conversion outcomes. Second, it examines the emerging role of generative artificial intelligence (AI) as an assistive evaluator in heuristic analysis, assessing its potential to improve the efficiency,

consistency, and scalability of usability inspection while preserving human judgment. Third, the findings bear directly on UX design practice, digital marketing strategy, organizational decision-making, and web accessibility initiatives.

The remainder of the paper is organized as follows. Section II reviews related work. Section III develops the conceptual framework, defining each scoring construct and its grounding in the literature. Section IV states the guiding hypothesis and research question. Section V describes the methods and Section VI the implementation of the instrument. Section VII reports results; Section VIII discusses them; Section IX states limitations; Section X the implications; Section XI future research; and Section XII concludes.

Related Work

Recent studies demonstrate that traditional usability frameworks such as Nielsen's heuristics require meaningful reinterpretation to remain effective across rapidly evolving technological domains. Specialized contexts conversational agents [6], decision support systems [8], financial platforms [1], and AI-driven environments [9], reveal limitations in static inspection methods and demonstrate the need for adaptive, data-driven, domain-specific heuristics. The literature consistently emphasizes the value of expert evaluators, the necessity of mobile and technical optimization, and the growing role of generative AI in both prototyping and user experience (UX) evaluation. Persistent gaps remain regarding scalability, demographic generalizability, longitudinal validation, and the ethical implications of AI-assisted design.

An emerging consensus holds that traditional frameworks such as Nielsen's heuristics [10] need substantial adaptation to remain efficient within varied technological domains [6], [11]. Abdul Aziz et al. [11] demonstrate the usefulness of aligning general principles to specific e-commerce features to enhance textbook marketplace transactions, while Langevin et al. [6] propose redefining heuristics to support the requirements of conversational agents, such as trustworthiness, context preservation, and the replacement of "help and documentation" with "help and guidance." A notable methodological departure comes from Genne [1], who argues that static inspection is inadequate for high-traffic environments where real-time behavioral analytics are needed to detect short-term friction and adapt to stringent regulation. This reframes UX optimization as a continuous, data-driven architectural function rather than a visual refinement phase [1].

Despite the rigor of multi-phase validation processes, considerable gaps remain regarding scalability and applicability across user demographics [6]. Abdul Aziz et al. [11] note that limited evaluator pools, often a single representative per user group, narrow generalizability, a limitation also reflected in the non-representative samples of Langevin et al. [6]. Genne [1] addresses the intersection of security and performance, but the infrastructure costs of event-based analytics represent a notable barrier for smaller organizations. Future research must examine the integration of AI to personalize user journeys and study the UX implications of emerging technologies such as blockchain and biometric authentication [1], [6], focusing on cross-platform consistency and adaptive security measures that do not compromise UX transparency [1].

Heuristic evaluation remains highly cost-efficient, particularly when three to five evaluators are used to identify 80% to 90% of interface violations [8], [10]. Studies show strong consensus that effectiveness is enhanced by at least two experts holding both usability expertise and domain knowledge [8], [10]. Where expert-based approaches focus on functional optimization and compliance with evidence-based design standards, quantitative user-based studies in e-commerce show that technical usability and mobile optimization are direct indicators of customer satisfaction and conversion [2], [8]. Cho et al. [8], Nielsen [10], and Nawir and Hendrawan [2] confirm that proactive usability adjustments are necessary to avoid unintended consequences, reduce cognitive load, and mitigate the costs of post-deployment fixes.

Gaps persist concerning the application of heuristic evaluation to paper prototypes versus live systems, since absent interface elements are harder to detect in non-interactive mockups [10]. Research underscores a need for pre-deployment testing to ensure safety, while e-commerce research shows a need for geographically specific studies of market dynamics [2], [8]. Limitations across studies include small expert samples in qualitative evaluations and dependence on self-reported Likert scales in quantitative research [2], [8]. Recent approaches propose multi-stage methodologies combining expert heuristic assessment with multivariate statistical techniques to measure the role of user satisfaction in organizational outcomes [2]. Future work should adopt likelihood-based approaches for hard-to-detect errors and longitudinal designs that track how usability improvements affect long-term behavior and adoption [2], [10].

Maguire [3] and Seethaladevi and Felisiya [12] agree that UX and usability are primary drivers of conversion, especially through optimized navigation and page-load speed. Maguire [3] differentiates functional usability from the emotional perceptions of UX; Seethaladevi and Felisiya [12] provide empirical validation, identifying a strong positive correlation ($r = 0.72$) between page-loading speed and conversion. Both align on mobile responsiveness: Maguire [3] emphasizes the technical integration of digital wallets and biometric features that enable impulse purchases, while Seethaladevi and Felisiya [12] find that younger users (18–25) are most influenced by these elements. An emerging interpretation is that usability can exist without positive UX; for example, when a site is functional but lacks the credibility or visual depth to inspire trust [3].

Limitations here include reliance on convenience sampling and insufficient data from emerging digital markets. Maguire [3] identifies tension between business-centric goals (e.g., upselling, data collection) and user-centric goals (e.g., efficiency, clutter-free interfaces), but little empirical research measures how that tension affects purchase decisions. Seethaladevi and Felisiya [12] acknowledge that brand reputation and security were not analyzed in depth, leaving a gap given that trust is central to e-commerce. Progress requires moving beyond descriptive surveys toward longitudinal randomized experiments and comparative prototype testing, within an integrated framework that balances universal accessibility against evolving social shopping behavior so that the transparency of AI-driven recommendations maintains consumer trust.

Across UI/UX, smart product design (SPD), and e-commerce, prototyping is identified as an essential venue for stakeholder communication, though the intended use of prototypes is growing more complex [7], [13], [14]. Researchers agree that traditional human-centered design processes need adjustment for software-intensive and data-driven systems, particularly as generative AI (GenAI) begins to automate rapid prototyping and visual generation [7], [14]. Applications differ: traditional e-commerce focuses on foundational usability, aesthetics, and satisfaction through established design-thinking stages, whereas SPD focuses on interactivity and environmental fidelity to address intangible service components [13], [14]. An emerging interpretation holds that prototypes are evolving from static representations of form into iterative objects that must accommodate AI augmentation and collaborative ecosystems [7], [14].

The long-term integration of these technologies, particularly AI-augmented prototyping, remains

insufficiently explored [7], [14]. A key limitation is the absence of longitudinal data and standardized assessment frameworks, as most studies provide only snapshot analyses or single-case applications [7], [13]. Gaps also exist in early-stage conceptualization, where GenAI is underused relative to late-stage asset generation, and designers report difficulty generating ideas because of the aggregate nature of these tools [7]. A further risk is that designers accept polished AI-generated output without understanding its algorithmic logic or systemic biases [7]. Future approaches should reconceptualize AI as both design material and social actor, supported by ethically grounded, collaborative frameworks that ensure transparency and diversity [7], [14].

The integration of advanced computation into UX evaluation marks a shift from manual, resource-intensive laboratory studies toward automated, scalable assessment. Ntoa et al. [9] and Liu [15] agree that growing contextual complexity demands multi-method approaches combining formative expert review with summative user testing. Ntoa et al. [9] propose the User Experience in Intelligent Environments framework, using environmental sensors to compute 103 metrics automatically; Liu [15] highlights the role of GenAI and large language models in simulating user personas and integrating qualitative feedback. Together, these lines of work suggest a reading of UX assessment in which multimodal data collection reduces dependence on subjective self-report while lessening evaluator cognitive load.

Significant gaps remain, particularly around the transparency of black-box models - systems whose internal operations are not readily interpretable although their inputs and outputs can be observed - and the absence of standardized evaluation frameworks for AI-assisted tools. A key methodological risk is AI-generated hallucination and contextual misinterpretation, which can produce misleading design decisions without rigorous human-in-the-loop validation [15]. Ntoa et al. [9] manage parameter volume through focused metric categorization, while Liu [15] identifies dataset bias and limited demographic generalizability as systemic limitations, proposing explainable AI techniques such as Shapley additive explanations and local interpretable model-agnostic explanations alongside domain adaptation and transfer learning. The unresolved challenge is a unified, cross-disciplinary benchmark balancing AI performance against privacy, data protection, and algorithmic accountability.

Collectively, the reviewed literature indicates that usability evaluation is transitioning from a discrete design

phase into an intelligence-supported function. Heuristic evaluation remains foundational and cost-efficient, but its effectiveness increasingly depends on hybrid approaches integrating behavioral analytics, AI-driven insight, and domain-specific adaptation, and on addressing dataset bias, infrastructure barriers, and the absence of standardized frameworks for AI-augmented evaluation.

Conceptual Framework

A. Nielsen's heuristics. The Overall UX Score evaluates each interface against Nielsen's ten usability heuristics, a set of broad principles for interaction design that Nielsen originally formalized [4] and later illustrated with concrete examples [5]. The first heuristic, visibility of system status, holds that a design should keep users informed through timely feedback, much as "You Are Here" indicators on mall maps show people where they are and help them decide where to go next [5]. The second, match between the system and the real world, requires that a design speak users' language using familiar words, phrases, and concepts and follow real-world conventions so that information appears in a natural order; the canonical illustration is a stovetop whose controls match the physical layout of the burners [5]. The third, user control and freedom, recognizes that users act by mistake and therefore need a clearly marked emergency exit - an undo or cancel path out of an unwanted state [5]. The fourth, consistency and standards, invokes Jakob's Law in advising that users should not have to wonder whether different words, situations, or actions mean the same thing, since people spend most of their time on other products and expect a new interface to work in familiar ways [5]. Across all ten principles - further encompassing error prevention, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design, error recognition and recovery, and help and documentation - the heuristics provide an established framework for diagnosing usability problems, and they constitute the dimensions aggregated into the Overall UX Score [4], [5].

B. Number of evaluators. The number of evaluators required for heuristic evaluation derives from the method's origins and its underlying mathematics. Nielsen and Molich [16] introduced heuristic evaluation as a discount inspection method in which a small set of experts independently judge an interface against established principles. Because no single evaluator detects every problem, evaluator count is consequential: Nielsen and Landauer [17] formalized the relationship in a model showing that the proportion of problems found follows $1 - (1 - \lambda)^n$, where n is the number of evaluators and λ is the average proportion of problems detected by one evaluator

(approximately 0.31). The model demonstrates diminishing returns, and the authors concluded that aggregating three to five evaluators identifies the majority of problems cost-effectively; a single evaluator typically uncovers about 35% of issues, whereas five surface approximately 75% [17]. Nielsen [18] popularized the related observation that small samples suffice and later reaffirmed that the same diminishing returns logic governs how many participants or evaluators a study needs [19]. Practitioner guidance from Moran and Gordon [20] maintains the three-to-five recommendation. The present study deliberately departs from this norm, substituting a single deterministic automated evaluator for the expert panel; it trades inter-rater triangulation for perfect procedural consistency, a trade-off taken up in Sections V and IX.

C. Overall UX Score. The Overall UX Score is operationalized as a composite index on a 0-100 scale, computed as a weighted average of sub-scores assigned to each of Nielsen's ten heuristics [4], [16]. Expressing usability on a 0-100 scale follows established convention, as Brooke [21] demonstrated with the System Usability Scale, which likewise condenses multidimensional usability data into a single comparable figure. Grounding the index in heuristic evaluation reflects that method's standing among inspection techniques: Jeffries et al. [22] found heuristic evaluation especially efficient at surfacing many usability problems at low cost, and Lynch [23] applied it alongside usability testing to assess real-world websites. Because aggregating ten heuristics into one value requires a weighting decision, the index was subjected to a sensitivity analysis recomputing every record under an equal scheme, an author-assigned scheme, and a Nielsen-1994 scheme in which each heuristic is weighted by the proportion of total variance its corresponding factor accounted for in Nielsen's principal components analysis of 249 usability problems [4], ranging from 6.1% for visibility of system status to approximately 2% for the lowest ranked factors, with a documented floor assigned to error recovery and help and documentation. Following the original characterization of these factor labels as a subjective abstraction of each factor's usability thrust [4], the scheme is treated as empirically informed rather than empirically fixed.

The index further incorporates a 0-5 explanatory rating for each finding, a measure of how completely a heuristic accounts for an observed problem, through an optional evidence-weighting procedure in which each heuristic's contribution is scaled by its rating ($\exp/5$), so that better evidenced findings exert greater influence on the composite. Finally, because a composite of expert judgments is sensitive to who conducts the evaluation,

the score must be interpreted in light of the evaluator effect documented by Hertzum and Jacobsen [24], whereby different evaluators applying the same method to the same interface identify substantially different problem sets. Taken together, these procedures render the Overall UX Score transparent, empirically grounded, and demonstrably robust, while acknowledging that it is best understood as a structured comparative estimate of usability grounded in an established inspection method rather than as a psychometrically validated instrument.

D. Heuristic weighting schemes. Because the Overall UX Score aggregates ten heuristics into a single value, the weight assigned to each is a consequential modeling choice; the study therefore compares three schemes. The equal scheme assigns every heuristic a weight of 1.0, not assuming relative importance. The author-assigned scheme applies face-valid weights that modestly emphasize the heuristics judged most consequential to task success. The Nielsen-1994 scheme derives its weights empirically from the proportion of total variance each corresponding factor accounted for in Nielsen's principal-components analysis of 249 usability problems [4], with a documented floor assigned to error recovery and help and documentation, which fell outside the reported top-ten factors. Following the original caution that the factor labels represent a subjective abstraction [4], the factor-to-heuristic mapping is treated as empirically informed rather than fixed. Each scheme's weights are normalized to a mean of 1.0 so that only their relative distribution differs, since absolute scale is immaterial to a weighted mean. Weight values are reported in Section V.

E. The LIFT conversion model. The LIFT Model, Landing Page Influence Function for Tests, is a conversion optimization framework created by Chris Goward in 2009 while he was founder and CEO of the optimization agency WiderFunnel [25], [26]. Goward developed the model as a structured, heuristic alternative to ad hoc best practices, organizing evaluation of any web or mobile experience around six conversion factors (e.g., value proposition, clarity, relevance, anxiety, distraction, and urgency) viewed from the perspective of the page visitor [25]. He formalized the framework in *You Should Test That!*, which remains the primary published source [26].

F. Conversion Quality. The Conversion Quality index operationalizes the six LIFT factors [26]. Because that framework is a practitioner heuristic rather than a psychometrically validated instrument, each factor is grounded in a construct validated in the relevant literature, so that scoring rests on an established theoretical foundation rather than practitioner assertion.

App X operationalizes these constructs as expert-inspected interface properties expected to influence the corresponding user-perception construct, not as direct measurements of users' perceptions; the validated scales cited below therefore establish theoretical warrant for each factor, not measurement equivalence with the index's 0-100 scores. Value proposition is grounded in perceived value, operationalized through the multidimensional PERVAL scale [27] and the perceived-usefulness construct of the Technology Acceptance Model (TAM) [28], with the value-to-behavior link reinforced by meta-analytic evidence that perceived value drives satisfaction and repurchase intention [29]. Relevance maps onto information scent within Information Foraging Theory [30], whose influence on goal completion is corroborated by research on the determinants of e-commerce conversion [31] and by evidence that a corporate website's contextual relevance shapes purchase intention [32]. Clarity is grounded in the perceived ease-of-use construct of TAM [28], a dimension distinct from the perceived usefulness anchoring value proposition, and in cognitive load theory [33], consistent with evidence that legible, well-structured retail design measurably raises conversion [34]. Anxiety, scored inversely as low anxiety, is grounded in online trust and perceived risk, shown to operate as mediating and moderating mechanisms on the path from website quality to purchase intention [35], [36]. Distraction, also scored inversely, is grounded in visual complexity, experimentally linked to users' affect and approach behavior [37], with the underlying attentional-interference mechanism corroborated meta-analytically [38]. Urgency rests on the scarcity principle of persuasion [39] and on the motivation-ability-trigger behavior model [40]. Each factor's 0-100 score is interpreted against criterion bands rather than absolute conversion benchmarks.

G. Criterion-referenced scorecard. Because the LIFT Model is a practitioner framework rather than a validated scoring standard [26], App X operationalizes Conversion Quality through a criterion-referenced scorecard mapping each factor to a construct validated in the usability, trust, persuasion, and e-commerce-conversion literature. The calibration matrix pairs each factor with its validated construct, a representative scale or measure, observable example indicators, and its expected direction of effect: value proposition in perceived value and perceived usefulness [27], [28]; relevance in information scent and message match [30], [31]; clarity in perceived ease of use and readability [28], [34]; anxiety in online trust and perceived risk [35]; distraction in visual complexity [37]; and urgency in the scarcity principle and

the motivation-ability-trigger model [39], [40]. The scorecard scores each factor as an expert-inspected interface property expected to influence its corresponding construct rather than as a direct measurement of users' perceptions. Each factor is scored on 0-100 against criteria derived from its anchoring construct and interpreted by band, from excellent (80-100) through poor (0-19). Factor sub-scores are combined into the composite through a weighted mean, $LIFT = (\sum w_i s_i) / (\sum w_i)$, in which each sub-score s_i is multiplied by its factor weight and the result is clamped to 0-100; weights are configurable and reported under sensitivity analysis. Because the framework references each factor to criteria derived from validated constructs rather than to industry-average conversion rates, the resulting estimates are expressed in relative, criterion-referenced terms, an expected competitive position rather than an absolute predicted conversion rate.

H. User satisfaction. Customer satisfaction (CSAT) was operationalized as the percentage of survey respondents expressing satisfaction with the usefulness of their evaluation, measured by a single five-point item ranging from "very dissatisfied" to "very satisfied" and scored using a top-two-box rule. This situates CSAT within the established customer satisfaction measurement tradition formalized in the American Customer Satisfaction Index (ACSI) [41] and reflects satisfaction theory's conceptualization of satisfaction as a post-experience evaluative judgment [42]. Although a single-item measure precludes estimation of internal consistency reliability, single-item measures of global, concrete constructs perform acceptably relative to multi-item alternatives [43]. The reported CSAT is accompanied by the mean rating, response count, and full distribution, allowing the collapsed percentage to be interpreted against its underlying spread. A 70% threshold was adopted as a researcher-specified target; this criterion is not an empirically validated cutoff but is contextualized against national norms, which have remained in the mid-to-high 70s on a 0-100 scale [41], [44]. Because the ACSI is a latent-variable index rather than a top-two-box proportion, it serves as a contextual reference rather than a direct equivalent, and the 70% threshold is treated as a conservative operational criterion given the tendency of top-two-box measures toward the upper end of the scale.

I. Accessibility guidelines. The App X interface was designed in accordance with the Web Content Accessibility Guidelines (WCAG) 2.1 [45], reflecting evidence that implementing accessibility guidelines improves the experience of both nondisabled users and users with visual impairments [46]. Evaluation scores are encoded using a five-band color scale mapped to defined ranges; consistent with Success Criterion 1.4.1 (Use of

Color), color is used redundantly with numeric values and text labels so that no information depends on color perception alone. Text and interface elements were designed to meet the Level AA minimum contrast ratio of 4.5:1 for normal text [47] and 3:1 for non-text and user-interface components [48], with selected high-emphasis surfaces meeting the Level AAA ratio of 7:1; these choices are supported by research linking legibility to luminance contrast [49] and to the readability needs of visually impaired readers [50], [51], as well as by work clarifying how WCAG contrast requirements map onto perceptual color space [52]. The score-to-color mapping is an author-defined visualization convention informed by, but not prescribed by, WCAG, and because a full multi-criterion conformance audit was not conducted, accessibility claims are limited to the contrast and use-of-color provisions described. Red-green encodings pose a known risk, since red-green color-vision deficiency affects approximately 8% of men of Northern European ancestry [53]; App X mitigates this through redundant numeric and textual encoding consistent with Success Criterion 1.4.1, an approach aligned with broader guidance on color schemes that remain interpretable under color-vision deficiency [54]. The palette is additionally cross-checked against WCAG 2.2 [55], whose minimum-contrast thresholds are unchanged from those of WCAG 2.1.

J. Integrating usability and conversion evaluation. Two methodologically distinct traditions govern the evaluation of digital interfaces. Usability inspection is anchored by the heuristic evaluation framework of Nielsen and Molich [16], in which a small set of experts judges an interface against established principles, the ten heuristics subsequently codified by Nielsen [4], to identify conditions that obstruct task completion. Conversion rate optimization is exemplified by practitioner frameworks such as Goward's LIFT Model [26], which isolates the conditions obstructing a visitor's progress toward a conversion goal. These traditions address related but non-equivalent problems, and yet they are commercially intertwined: Kornichuk and Boryczka [56] demonstrated that landing page text readability can predict conversion rates, underscoring that subtle usability properties shape commercial outcomes even where the two literatures remain conceptually distinct. Despite this interdependence, the literature contains no formally validated protocol integrating Nielsen's heuristics with the LIFT Model, a striking absence given the maturity of both traditions. This fragmentation reflects two documented shortfalls: Zimmermann and Auinger [57] found that no established guidelines existed for identifying and evaluating the touchpoints that influence conversion, prompting their call for a standardized

optimization framework, while Liu [15], in a systematic review of automated and remote UX evaluation, identified standardization as critically underrepresented. By formally integrating the ten heuristics with the six LIFT factors, the present work proposes an academically grounded unified protocol addressing both usability inspection and conversion suppression within a single diagnostic framework, advancing the standardization agenda documented by Zimmermann and Auinger [57] and Liu [15].

Hypothesis and research question

Website usability is central to how people behave and how satisfied they feel, because modern organizations depend on digital platforms to drive interaction and conversion [3], [9], [58]. Barriers such as unclear navigation create friction that stops users short of the actions a site depends on [3], [59]. Heuristic evaluation offers an efficient, low-cost way to surface those barriers without the overhead of full user testing [11], [60], [61]. This framing rests on evidence that usability quality tracks engagement and conversion [62], [63] and on the established efficiency of heuristic methods for producing such diagnoses without extensive testing resources [64], [65], [66]. Because organizations that act on these diagnoses tend to see stronger retention and business outcomes [67], [68], establishing a reliable, reproducible diagnosis is itself a contribution as well as the groundwork for the causal trial that follows.

H (research program; tested in future remediation work). Systematically remediating the usability barriers identified through a structured heuristic evaluation using Nielsen's ten heuristics will improve holistic website performance, specifically, a measurable, practically meaningful increase in conversion rate together with a satisfaction score of 70% or higher. This causal claim requires a pre-post remediation design and is not tested in the present cross-sectional study.

RQ (addressed in the present study). To what extent can an integrated protocol that applies Nielsen's ten usability heuristics and the LIFT model's six conversion factors, delivered through a single standardized automated instrument, reliably identify and prioritize a website's usability and conversion barriers, and to what extent do those two classes of weakness co-occur on the same interface?

Answering this question establishes the diagnostic foundation on which the program's hypothesis can subsequently be tested. Three sub-questions structure the analysis: RQ₁, is the instrument's output reproducible and robust to its author-defined modeling choices? RQ₂, does

the instrument discriminate, does it produce differentiated, actionable prioritization rather than an undifferentiated verdict? RQ₃, are usability and conversion findings related but non-redundant?

Methods

A. Research design. The study employed a cross-sectional, empirical expert-evaluation design to examine the diagnostic utility and complementarity of two established interface evaluation frameworks, Nielsen's ten usability heuristics [4], [16], and the LIFT conversion model [26], when applied jointly to live websites. Rather than treating usability and conversion as separate evaluation traditions, the design operationalized both within a single standardized automated instrument (the App X evaluation system) and applied them to every submitted website, producing a comparative dataset of usability and conversion findings across industry contexts. Each website was assessed in one integrated session in which both frameworks were applied to the same interface in a single pass. Integrated application within a single session, rather than as sequentially administered separate sessions, was adopted to maximize ecological validity [69] and to avoid the inconsistency and fatigue that repeated, multi-session inspection of the same interface is known to introduce [24], while enabling direct comparison of the findings each framework produced.

B. Website sampling and recruitment. The instrument was deployed as a publicly accessible web application, and participants were recruited through a single professional network announcement and direct outreach. Participants selected the website to be evaluated (e.g., their own, an employer's, or a site of interest), which makes the resulting site set a purposive, participant-generated convenience sample rather than a probability sample. This is appropriate for small-sample, theory-driven evaluation research in which information-rich cases are selected deliberately rather than at random [70], [71], but it precludes statistical generalization to any defined population of websites. Sites were required to be publicly accessible without authentication, to present an identifiable primary conversion goal (e.g., enrollment inquiry, purchase, booking, or sign-up), and to offer sufficient interface complexity to permit assessment across all ten heuristics and all six LIFT factors.

Data collection produced 50 completed evaluations across 27 unique websites over a ten-day window. The realized site set spanned higher education, retail and travel e-commerce, developer and payments infrastructure, productivity software, creative and professional services,

hospitality, and nonprofit categories. Because participants were free to resubmit a site, six sites were evaluated more than once; the resulting departure from independence is quantified in Section VII-L and treated as a limitation in Section IX.

C. The evaluator: a standardized automated instrument. Evaluation was performed by a single standardized automated evaluator rather than a panel of human experts. App X operationalizes both frameworks as a rule-based scoring protocol, applying the same weighted heuristic matrix, the same construct-referenced LIFT scorecard, and the same criterion thresholds to every website. This design controls, rather than measures, the between-evaluator variability documented in the usability literature, the evaluator effect, in which different human evaluators inspecting the same interface with the same method detect markedly different problem sets [24]. By substituting a single deterministic protocol for a human panel, the study trades the multi-rater triangulation of traditional heuristic evaluation [16] for procedural consistency and reproducibility across the sample; the implications for criterion validity are addressed in Section IX.

The architecture provides two scoring paths: a generative AI path, in which a language model inspects retrieved page content and returns per-heuristic and per-factor scores with findings, and a deterministic modeled path that applies the same scoring matrix to extracted page signals when the AI path is unavailable. Which path produced each record is logged. As reported in Section VII-K, every record in this dataset was produced by the

modeled path; that fact conditions the interpretation of all subsequent results and is carried explicitly through Sections VIII and IX.

D. Evaluation instrument: the unified assessment protocol. The primary instrument was a purpose-developed unified protocol that integrated Nielsen's ten heuristics with the LIFT model's six conversion factors into a single structured assessment, enabling simultaneous coverage of usability inspection and conversion-condition analysis. For each website, the protocol followed the two-pass procedure recommended by Nielsen [4]: a first pass to develop a holistic understanding of the interface's structure, content, and primary conversion pathway, and a second pass to conduct a systematic criterion-referenced inspection of each element.

For the usability component, each of the ten heuristics received a 0-100 quality score accompanied by a specific finding, a recommendation, and a 0-5 explanatory rating indicating how completely the heuristic accounted for the observed issue, following Nielsen's emphasis on the explanatory power of individual heuristics [4]. The ten scores were aggregated into an Overall UX Score using a weighted mean. Because the weighting of heuristics is a consequential modeling choice, three schemes were computed and compared in a sensitivity analysis; their weights are given in Table I. Weights are normalized to a mean of 1.0 so that only their relative distribution differs.

TABLE I. Heuristic Weighting Under Each Scheme (Normalized to Mean 1.0)

Heuristic	Nielsen 1994 raw (% var.)	Nielsen 1994 norm.	Author	Equal
H1 Visibility of system status	6.1	1.76	1.2	1.0
H2 Match with the real world	5.9	1.71	1.0	1.0
H3 User control & freedom	4.6	1.33	1.0	1.0
H4 Consistency & standards	4.2	1.21	1.2	1.0
H5 Error prevention	3.7	1.07	1.1	1.0
H6 Recognition rather than recall	3.1	0.90	1.0	1.0
H7 Flexibility & efficiency	2.0	0.58	0.9	1.0

H8 Aesthetic & minimalist design	2.0	0.58	0.9	1.0
H9 Error recovery & diagnosis	1.5a	0.43	1.0	1.0
H10 Help & documentation	1.5a	0.43	0.7	1.0
Sum / mean	34.6 / 3.46	10.0 / 1.0	10.0 / 1.0	10.0 / 1.0

a H9 and H10 fall outside the top-ten factors reported in Nielsen's principal-components analysis [4]; both were assigned a documented author-determined floor value. The Nielsen-1994 scheme spreads weight more widely (0.43–1.76) than the author scheme (0.70–1.20).

For the conversion component, because the LIFT model is a practitioner framework rather than a validated scoring standard [26], each factor was scored on 0–100 against criteria derived from a construct validated in the relevant literature: value proposition in perceived value and perceived usefulness [27], [28]; relevance in information scent [30], [31]; clarity in perceived ease of use and cognitive load [28], [33]; anxiety in online trust and

perceived risk [35]; distraction in visual complexity [37]; and urgency in the scarcity principle and behavior model triggers [39], [40]. Each factor was scored as an expert-inspected interface property expected to influence its corresponding user perception construct rather than as a direct measurement of that construct. Factor scores were combined through a weighted mean and interpreted against criterion bands, yielding relative, criterion-referenced estimates rather than absolute predicted conversion rates. The three factor weighting schemes compared in the sensitivity analysis are given in Table II.

TABLE II. LIFT Factor Weighting Under Each Scheme

Factor	Equal	LIFT (Goward)	Evidence-informed
Value proposition	1.00	0.30	0.28
Clarity	1.00	0.20	0.24
Relevance	1.00	0.18	0.16
Low anxiety	1.00	0.14	0.16
Urgency	1.00	0.10	0.08
Low distraction	1.00	0.08	0.08

Goward's framework [26] establishes only the priority ordering - value proposition primary, the remaining five modulating it - and publishes no numeric weights; the decimal values in the LIFT column are therefore author-assigned. The evidence-informed column redistributes weight toward factors with peer-reviewed predictive support (clarity and readability, trust, value).

E. User satisfaction and acceptance measures. A survey was administered following report generation. Satisfaction was measured with a single five-point item and summarized as a top-two-box CSAT score, the percentage of respondents selecting "satisfied" or "very satisfied" [41], [42], reported alongside the mean rating and full distribution. Although single-item measures preclude internal consistency estimation, they are considered acceptable for global, clearly defined constructs [43]. A 70% threshold was specified a priori as a researcher-defined criterion, contextualized against but

not equated with ACSI national norms, which have held in the mid-to-high 70s on a 0–100 scale [44]. A ten-item acceptance questionnaire on a seven-point agreement scale accompanied the satisfaction item, grouped into three blocks (e.g., core evaluation, application quality and trust, and report and decision support), and reported per item. Respondents additionally reported field of work, years of experience, gender, age, U.S. region, and highest education level to permit descriptive segmentation.

F. Ethical procedure. Participation was gated by an on-screen informed consent notice stating the study's purpose, the data recorded, the voluntary nature of participation, and the researcher's contact details. Only sessions in which the participant selected "I Agree" proceeded to evaluation; declining sessions terminated without recording an evaluation. The study evaluated publicly accessible websites and collected no data from those sites' end users. Respondent identifiers were

collected for response integrity verification and are reported here only in aggregate; the non-anonymous collection is treated as a potential source of social desirability bias in Section IX.

G. Reliability and data analysis. Consistent with the single automated evaluator design, reliability was established through reproducibility rather than inter-rater agreement. Because the modeled scoring path is deterministic, resubmitting an identical website under an identical configuration returns identical scores; inter-rater statistics such as the intraclass correlation coefficient and Cohen's κ are therefore reported as not applicable. Three reproducibility checks were computed instead: invariance of the composite across weighting schemes, repeat audit agreement for sites evaluated more than once, and stability of the cumulative mean as the sample accumulated.

Analysis was descriptive. Per-heuristic and per-factor means, medians, standard deviations, variances, and ranges were computed. Overall UX Scores were compared across the three weighting schemes and the relationship between Overall UX Score and Conversion Quality was examined with the Pearson product-moment correlation, reported with its coefficient of determination. Satisfaction data were summarized as CSAT, mean rating, and distribution. No inferential hypothesis testing was conducted: the sample was a non-probability convenience sample containing repeated measures on six sites, so p-values computed on it would rest on assumptions of independence and random sampling that the design does not satisfy. All reported statistics are descriptive of this sample.

Implementation

This study employed a structured, multi-phase methodology designed to systematically investigate usability barriers and their impact on conversion potential within live website environments. Each phase integrated qualitative and quantitative methods, expert-criterion evaluation, and AI-augmented analysis to ensure methodological rigor, replicability, and comprehensive insight. Implementation proceeded in three phases.

A. Phase 1: Application development. This phase produced an AI-augmented web-based evaluation system that applies the unified inspection protocol to live websites. The application comprised three components: a landing page with a URL submission interface; an evaluation dashboard presenting heuristic and conversion

scores, findings, recommendations, strengths, weaknesses, and the satisfaction survey; and a researcher dashboard aggregating all evaluation and survey records. Heuristic evaluation was implemented against Nielsen's ten heuristics [4], [16] with a standardized weighted-scoring matrix. Conversion scoring was derived from the six LIFT factors [26], each scored against criteria drawn from a validated construct. Standardized scoring logic, criterion bands, and a deterministic rule set ensured that every website was evaluated under identical conditions. Before deployment, the application underwent internal pilot testing to verify links, scoring routines, exports, and survey forms.

B. Phase 2: Automated evaluation. Participants submitted URLs and the application generated a complete usability and conversion assessment in a single integrated session, applying both frameworks to the same interface in one pass to maximize ecological validity [69] and avoid the inconsistency of repeated multi-session inspection [24]. Because evaluation was performed by a single deterministic protocol rather than a panel of raters, the evaluator effect [24] was controlled by design rather than mitigated procedurally, and re-submitting an identical website returned identical scores. Following each evaluation, an optional survey captured perceptions of the report's usefulness together with demographic characteristics.

C. Phase 3: Data analysis and presentation. All results were compiled into a unified dataset and exported in CSV, JSON, and PDF formats for descriptive analysis. Average heuristic and conversion factor scores, score distributions, and summary statistics were computed across the sample, and the relationship between usability and conversion scores was examined to characterize cross-industry patterns.

Results

A. Consent disposition. Fifty-one consent decisions were recorded. Fifty participants agreed, and one declined; the declining session produced no evaluation, so no unconsented activity entered the dataset (Table III). All satisfaction, CSAT, and questionnaire statistics reported below derive exclusively from consented sessions. With one refusal, the figure is descriptive of this sample and supports no estimate of refusal behavior at scale.

TABLE III. Consent Disposition

Disposition	Count	% of decisions
Consented ("I agree")	50	98

Declined ("I do not agree")	1	2
Evaluations run by declining sessions	0	0
Total consent decisions	51	100

B. Descriptive summary. The consented dataset comprised 50 evaluations across 27 unique websites, with 25 survey responses for a 50% response rate. Table IV reports the three headline outcome measures. The Overall UX Score averaged 72.24 (Mdn = 73, SD = 3.51, range 62–79), with all 50 evaluations exceeding the instrument's minimum threshold of 50. Conversion Quality averaged 72.0 (Mdn = 70, SD = 5.90, range 60–87),

with all evaluations meeting the ≥ 60 criterion. CSAT was 100% ($M = 4.80$, $SD = 0.40$ on the five-point item, $n = 25$) against the 70% target. The Overall UX Score and Conversion Quality were weakly-to-moderately correlated, $r = 0.40$.

TABLE IV. Descriptive Statistics for the Consented Dataset

Measure	Value	Mdn	M $\pm$ SD	Var.	Range / threshold
Evaluations	50	—	—	—	27 unique sites
Consents recorded	50	—	—	—	—
Survey respondents	25	—	—	—	50% response rate
Overall UX Score	72.24	73	72.24 $\pm$ 3.51	12.34	62–79 (min. 50)
Conversion Quality	72.0	70	72.0 $\pm$ 5.90	34.68	60–87 (≥ 60 : 100%)
CSAT (top-two-box)	100%	5	4.80 $\pm$ 0.40	0.16	target $\geq 70\%$, $n = 25$
UX–conversion association	0.40	—	—	—	r ; $r^2 \approx 0.16$

C. Respondent characteristics. Table V summarizes the 25 survey respondents. The sample was highly educated: 80% held a bachelor's degree or higher (bachelor's 32%, master's 40%, doctorate 8%). It was geographically concentrated, with 64% in the Southeast and no respondents from the Midwest or West. Occupationally, it was not a technology-practitioner sample: education and training (32%) and sciences and medicine (32%)

together accounted for 64%, technology and computing for 16%, and engineering for none. Respondents were professionally senior: 72% reported six or more years of experience, 28% reported 21 or more, and 64% were aged 35 or older. Gender was 64% male and 36% female. Fig. 1 displays the three dimensions that most constrain generalization.

TABLE V. Respondent Characteristics ($N = 25$)

Dimension	Largest category	n (%)	Second	n (%)	Notable absence
Field of work	Education & training / sciences & medicine (tie)	8 (32)	Technology & computing	4 (16)	Engineering 0 (0%)
Years of experience	6–10 / 21+ (tie)	7 (28)	1–5	5 (20)	—
Gender	Male	16 (64)	Female	9 (36)	—
Age	25–34	8 (32)	45–54	7 (28)	18–24 only 1 (4%)
U.S. region	Southeast	16 (64)	Northeast	7 (28)	Midwest & West 0 (0%)

Education	Master's	10 (40)	Bachelor's	8 (32)	—
-----------	----------	---------	------------	--------	---

Fig. 1. Respondent characteristics by U.S. region (dark blue), field of work (teal), and highest education level (amber); counts out of N = 25. Two of five census regions and one of seven occupational categories are unrepresented.

D. Acceptance questionnaire. Ten items on a seven-point agreement scale produced 250 individual ratings from 25 respondents. The pooled mean was 6.55 (Mdn = 7, SD = 0.82, variance 0.67), with 91% of ratings at 6 or 7. Item

means ranged from 6.40 to 6.72, a spread of 0.32 points across ten distinct constructs (Table VI, Fig. 2). The correlation between a respondent's mean questionnaire score and the Overall UX Score assigned to the site that respondent evaluated was $r = 0.01$.

TABLE VI. Acceptance Questionnaire Item Means and Pooled Statistics

Block	Item construct	M / 7	n
Core evaluation	Effectiveness (issues identified)	6.48	25
	Effectiveness (recommendations)	6.40	25
	Efficiency	6.72	25
	Satisfaction	6.72	25
Quality & trust	Accuracy	6.40	25
	Explainability	6.52	25
	Trust	6.56	25
Report & decision support	Report quality	6.52	25
	Decision support	6.56	25
	Behavioral intention	6.60	25
Pooled (N = 250 ratings)	6.55	25	

Pooled Mdn = 7; SD = 0.82; variance = 0.67; 91% of ratings at 6–7; r with assigned Overall UX Score = 0.01.

Fig. 2. Acceptance questionnaire item means on the seven-point scale, $n = 25$ per item. The full 0.32-point spread occupies the top 5% of the scale, indicating a ceiling effect rather than construct-level discrimination.

E. Overall UX Score: weighting sensitivity. Every evaluation was re-scored under all three heuristic weighting schemes with evidence weighting enabled

(Table VII, Fig. 3). All three schemes returned an identical mean of 72; the maximum spread between schemes was 0 points, classified low by the instrument's own stability rubric. Medians differed by at most one point, standard deviations by 0.3, and observed ranges overlapped almost completely. The Nielsen-1994 scheme was the active configuration for all other reported results.

TABLE VII. Overall UX Score Under Three Heuristic Weighting Schemes

Scheme	M	Mdn	Var.	SD	Range
Equal	72	72	11.8	3.4	63–79
Author-assigned	72	72	9.9	3.2	64–78
Nielsen 1994 (active)	72	73	12.3	3.5	62–79
Maximum spread	0	1	2.4	0.3	—

Evidence weighting: on. Weights as specified in Table I.

Fig. 3. Overall UX Score central tendency and observed extremes under the three heuristic weighting schemes. The composite mean is identical across schemes (maximum spread 0 points).

F. Overall UX Score distribution. All 50 evaluations (100%) fell within the good band, 60–79. No evaluation

was classified as poor, weak, moderate, or excellent (Table VIII, Fig. 4). The observed 17-point range occupied 17% of the nominal 100-point scale.

TABLE VIII. Overall UX Score Distribution by Criterion Band

Band	Range	Count	%
Poor	0–19	0	0
Weak	20–39	0	0
Moderate	40–59	0	0
Good	60–79	50	100
Excellent	80–100	0	0
Total	—	50	100

M = 72.24; Mdn = 73; SD = 3.51; variance = 12.34; 100% at or above the minimum threshold of 50.

Fig. 4. Distribution of the 50 Overall UX Scores across the five criterion bands. Every evaluation fell in the good band; four of five bands were empty.

G. Score by heuristic. Mean scores for the ten heuristics ranged from 59 (H10, help and documentation) to 78 (H2, match with the real world), a 19-point spread, with M =

69.3, Mdn = 71, SD = 5.24, and variance 27.41 across the ten dimension means (Table IX, Fig. 5). H10 was the only heuristic below the good threshold of 60. Dispersion across heuristics (SD = 5.24) exceeded dispersion across sites (SD = 3.51).

TABLE IX. Mean Score by Heuristic, Ranked (N = 10 Dimensions)

Rank	Heuristic	Mean	Band
1	H2 Match between system & real world	78	Good
2	H3 User control & freedom	72	Good
2	H5 Error prevention	72	Good
2	H7 Flexibility & efficiency of use	72	Good
5	H6 Recognition rather than recall	71	Good
5	H9 Error recovery & diagnosis	71	Good
7	H8 Aesthetic & minimalist design	70	Good
8	H4 Consistency & standards	66	Good
9	H1 Visibility of system status	62	Good
10	H10 Help & documentation	59	Moderate
Across the ten dimension means		69.3	SD 5.24; range 19

Fig. 5. Mean score by heuristic across all evaluations. The 19-point spread between H2 (78) and H10 (59) is the study's

principal evidence of diagnostic discrimination; H10 alone falls below the good threshold.

H. Finding quality: explanatory ratings. The 50 evaluations generated 500 findings, each carrying a 0-5 explanatory rating recording how completely the flagged check accounted for the identified problem. The mean rating was 3.59 with a modal rating of 4; 289 findings (58%) were rated 4 or 5 and 211 (42%) were rated 2 or 3.

No finding was rated 0 or 1 (Table X, Fig. 6). Evidence weighting was applied, so lower-rated findings contributed proportionally less to the composite scores reported in Sections VII-E and VII-F.

TABLE X. Explanatory Rating Distribution (500 Findings)

Rating	Count	%	Interpretation
0-1	0	0	Heuristic does not explain the finding
2	62	12	Partial or ambiguous explanation
3	149	30	
4	223	45	Well explained
5	66	13	
Total	500	100	M = 3.59; 58% rated ≥ 4

Fig. 6. Distribution of 0-5 explanatory ratings across all 500 findings. Unlike the satisfaction measures, this distribution is graded rather than at ceiling (M = 3.59, mode 4).

I. Conversion Quality: weighting sensitivity and factor profile. Conversion Quality was recomputed under three factor weighting schemes (Table XI, Fig. 7). Means differed by at most 1 point (71-72), and all three schemes shared a median of 70; the maximum spread was 1 point.

The evidence-informed scheme, constructed by redistributing weight toward factors with peer-reviewed predictive support, returned the same mean as Goward's original ordering. Dispersion on the conversion side exceeded that on the usability side by a factor of roughly 1.7 in standard deviation and 2.8 in variance (Fig. 8).

TABLE XI. Conversion Quality Under Three Factor Weighting Schemes

Scheme	M	Mdn	Var.	SD	Range
Equal	71	70	28.3	5.3	62-86
LIFT (Goward), active	72	70	34.7	5.9	60-87
Evidence-informed	72	70	33.3	5.8	60-86
Maximum spread	1	0	6.4	0.6	—

Fig. 7. Conversion Quality central tendency and observed extremes under the three factor weighting schemes (maximum spread 1 point).

Fig. 8. Dispersion of the two composite measures. Conversion Quality varied across sites substantially more than the Overall UX Score, indicating greater between-site discrimination.

At the factor level, the six LIFT dimensions averaged 71.0 (Mdn = 71.5, SD = 4.20, variance 17.67), spanning 65

(clarity) to 78 (value proposition), a 13-point range; all six fell within the good band (Table XII, Fig. 9). Low anxiety (67) was the second-weakest factor.

TABLE XII. LIFT Factor Calibration and Mean Scores (N = 6 Dimensions)

Rank	LIFT factor	Validated construct	Measure	Mean	Band
1	Value proposition	Perceived value / usefulness	PERVAL [27]; TAM usefulness [28]	78	Good

2	Low distraction	Visual complexity / cognitive load	Visual-complexity and order measures [37], [38]	73	Good
3	Relevance	Information scent / message match	Information Foraging Theory [30]; relevance-fit items [31]	72	Good
4	Urgency	Scarcity / motivation	Scarcity-perception [39] and motivation items [40]	71	Good
5	Low anxiety	Online trust / perceived risk	E-trust and perceived-risk scales [35], [36]	67	Good
6	Clarity	Perceived ease of use / readability	TAM ease [28]; readability [34]	65	Good
Across the six dimension means	71.0	SD 4.20; range 13			

Each factor is scored as an expert-inspected interface property expected to influence its anchoring construct. The cited scales establish theoretical warrant, not measurement equivalence; none was administered to users.

Fig. 9. Mean score by LIFT conversion factor. Clarity (65) and low anxiety (67) were the weakest dimensions; the 13-point spread is narrower than the 19-point heuristic spread in Fig. 5.

J. Association between usability and conversion quality. Overall UX Score and Conversion Quality were positively

associated, $r = 0.40$, $r^2 \approx 0.16$, indicating that approximately 16% of the variance in Conversion Quality was shared with the Overall UX Score (Table XIII). Both variables were range-restricted: the Overall UX Score spanned 62-79 and Conversion Quality 60-87, and no evaluation fell below 60 on either measure. Fig. 10 shows the observed measurement envelope against the nominal scale.

TABLE XIII. Association Between the Two Composite Measures

Statistic	Value
Evaluations	50
Mean Overall UX Score	72.24
Mean Conversion Quality	72.0
Pearson r	0.40
Shared variance r^2	0.16
Observed UX range	62–79
Observed Conversion Quality range	60–87

Descriptive; no significance test is reported because the sample is a non-probability sample containing repeated measures on six sites.

Fig. 10. Observed measurement envelope for the two composites against the nominal 0-100 scale, with the sample centroid and the direction of the observed association ($r = 0.40$). The envelope occupies roughly 5% of the nominal measurement plane, which

attenuates the correlation and precludes observation of the relationship at the low end of either scale.

K. Satisfaction. All 25 respondents rated their evaluation positively: 20 (80%) selected "very satisfied" and 5 (20%) "satisfied," with no neutral or negative responses (Table XIV, Fig. 11). The mean was 4.80 (Mdn = 5, SD = 0.40, variance 0.16) and CSAT was 100% against the 70% target.

Because all 50 evaluations fell in the good band, the planned cross-tabulation of satisfaction against assigned score band populated only one of five cells and yielded no between-band comparison. Within that band, satisfaction was uncorrelated with the assigned score, $r = -0.04$,

corroborating the $r = 0.01$ obtained for the acceptance questionnaire in Section VII-D.

TABLE XIV. Satisfaction Distribution and Cross-Tabulation by Score Band

Rating	Count	%	UX band	M satisf.	n
Very satisfied	20	80	Poor	—	0
Satisfied	5	20	Weak	—	0
Neutral	0	0	Moderate	—	0
Dissatisfied	0	0	Good	4.80	25
Very dissatisfied	0	0	Excellent	—	0
Total	25	100	CSAT 100%; M 4.80 $\pm$ 0.40; r vs. score = -0.04		

Fig. 11. Satisfaction distribution ($n = 25$). The negative tail is empty, and the measure is at ceiling, leaving no headroom to discriminate degrees of satisfaction.

L. Collection profile, evaluation source, and repeat audits. Collection spanned ten calendar days (12–21 August) with activity on eight of them, a mean of 6.25 evaluations per active day (Mdn = 5, SD = 7.0, variance 48.94). The distribution was strongly asymmetric: the opening day accounted for 24 of the 50 evaluations (48%), and the

remaining seven active days averaged 3.7 (Table XV, Fig. 12). The standard deviation exceeded the median, the signature of a distribution dominated by a single high-volume day. The cumulative mean of the Overall UX Score converged within approximately the first ten evaluations and remained flat thereafter, ending at 72 with no visible drift.

TABLE XV. Data Collection Profile

Statistic	Value
Calendar span	12–21 Aug (10 days)
Active collection days	8
Total evaluations	50
Mean per active day	6.25
Median per active day	5
SD / variance	7.0 / 48.94
Peak day	12 Aug: 24 (48%)
Mean excluding peak day	3.7
Days with no activity	16 and 18 Aug
Approximate convergence point	~10 evaluations

Fig. 12. Evaluations recorded per calendar day. The front-loaded profile is characteristic of single-announcement recruitment in which response decays with feed visibility.

Provenance logging established that all 50 evaluations were produced by the deterministic modeled scoring engine and none by the generative AI path (Table XVI,

Fig. 13). Because this conditions the interpretation of several earlier results, Table XVI states the revision each requires.

TABLE XVI. Evaluation Provenance and Its Consequences for Interpretation

Source	Count	%	Mean score
Generative AI path	0	0	—
Modeled path (deterministic fallback)	50	100	72

Affected result	Reading if AI-scored	Reading given modeled provenance
Band concentration (100% good)	Site quality clustered, or instrument compression	Consistent with the modeled engine's narrower output range operating on limited page signals
Explanatory ratings ($M = 3.59$)	Model self-assessment of finding completeness	Output of the modeled engine's rating logic, not of a generative model reflecting on its reasoning
Weighting invariance (0 and 1 point)	Composite scoring robust to weighting choice	Modeled engine robust to weighting; untested for the AI path, which would add a source of variance absent here
Repeat-audit agreement	Test–retest reliability of the instrument	Expected property of a deterministic engine; does not establish reliability for non-deterministic generation
Satisfaction and questionnaire results	Reception of the tool	Unaffected. These measure report format, workflow, and decision support

Fig. 13. Evaluation provenance. Every record in the dataset was produced by the deterministic modeled scoring path; the generative AI path contributed none.

Six sites were evaluated more than once, contributing 29 of the 50 evaluations (58%); one site alone contributed 15 (30% of the dataset). Five of the six returned identical scores on every repeat audit; median net change was 0, and mean net change was +1.5 ($SD = 3.35$). The single substantive movement was +9 points on one site (68 to

77), alongside one 75 among fifteen otherwise identical scores of 73 on another (Table XVII, Fig. 14). Under a deterministic engine, movement of this magnitude on the same URL implies either genuine change in page content between audits or a stochastic component in the pipeline. The findings module applies a seeded jitter to explanatory ratings, so some controlled variability exists. The dataset cannot distinguish these explanations.

TABLE XVII. Repeat-Audited Sites

Site	Audits	Score sequence	Net change
airbnb.com	15	73 ×9, 75, 73 ×4	0
stripe.com	6	72 ×6	0
ncl.com	2	68, 77	+9
notion.so	2	79, 79	0
thefoodiebunch.com	2	71, 71	0
heurlift.netlify.app	2	64, 64	0
6 sites	29	5 of 6 unchanged	$M +1.5$

Median net change 0; $SD = 3.35$; variance 11.25. The remaining 21 evaluations covered 21 singly-audited sites.

Fig. 14. Concentration of the 50 evaluations across 27 sites. Six repeat-audited sites supplied 29 evaluations, so the effective number of independent observations is nearer 27 than 50.

Discussion

A. Principal findings. The study asked what happens when two evaluation traditions that rarely meet are applied to the same interface in a single pass. Three findings answer the research question's sub-parts, and Table XVIII maps each to its evidence.

On reproducibility (RQ₁), the instrument's output proved insensitive to the modeling choices that a reviewer would most reasonably contest. The composite Overall UX Score was identical under equal, author-assigned, and variance-derived heuristic weighting (maximum spread 0 points); Conversion Quality varied by at most 1 point across three factor weighting schemes, five of six repeat-audited sites returned identical scores, and the cumulative mean stabilized within approximately ten evaluations. These are the appropriate response to the study's largest author-assigned components: rather than defending a particular weighting as correct, the analysis demonstrates that the conclusion does not depend on getting it right. Notably, the evidence-informed conversion scheme, built by shifting weight toward factors with peer-reviewed predictive support, converged on the same mean as Goward's practitioner ordering.

On discrimination (RQ₂), the instrument discriminated at the construct level but not at the site level. Every one of the 50 evaluations fell in a single one of five criterion bands, and the 17-point observed range occupied 17% of the nominal scale; a composite that assigns essentially the same verdict to every site cannot separate a well-designed interface from a poorly designed one. The heuristic and factor profiles behaved differently: the ten heuristics spanned 19 points (SD = 5.24) and the six LIFT factors 13 points (SD = 4.20), both wider in relative terms than the between-site dispersion of the composite (SD = 3.51). The diagnostic value therefore resides in the profile rather than the score. A practitioner told only that a site scores 72 learns little; one told that help and documentation averages 59, visibility of system status 62, clarity 65, and low anxiety 67 receives an actionable prioritization. The conversion composite was the more sensitive of the two at the site level, discriminating across sites roughly 1.7 times as widely as the usability composite (Fig. 8), while

the usability profile discriminated more widely across dimensions, a division of labor worth carrying into instrument design.

The profile itself is substantively coherent with the published literature, which strengthens its credibility. Help and documentation is routinely minimal on production websites, and system status feedback is commonly absent from marketing-oriented pages that have little interactive state to report; conversely, professionally built sites generally adopt familiar language and conventions, which is what the highest heuristic score (H₂, 78) records. On the conversion side, clarity and trust signaling are the dimensions conversion research most often finds under-addressed relative to messaging, and they are precisely the two weakest factors here (65 and 67) against the strongest (value proposition, 78). The instrument is not producing arbitrary rankings; it reproduces a recognizable, independently explicable pattern. That correspondence is weak convergent evidence for construct validity, not proof, but a result a reviewer can sanity-check against domain knowledge.

On complementarity (RQ₃), the two composites were related but far from redundant: $r = 0.40$ leaves roughly 84% of the variance in conversion quality unexplained by usability. This is the more useful result for the instrument's rationale. Were the correlation near unity, measuring both would be superfluous, and the combined audit would add nothing over a usability audit alone. A moderate association is what justifies assessing both dimensions in a single pass. At the construct level, the weak dimensions align in a conceptually interpretable way - clarity (65) with help and documentation (59) and consistency and standards (66), and low anxiety (67) with the absence of status and recovery feedback - while the strongest dimensions likewise align (H₂ at 78 with value proposition at 78). The alignment is at the level of constructs, not interface elements; the present data do not record which specific element triggered each flag, so element-level co-occurrence remains an inference rather than a measurement.

TABLE XVIII. Research Sub-Questions, Evidence, and Boundaries

Sub-question	Evidence in this study	What it does not establish
RQ ₁ Is output reproducible and robust to author-defined modeling choices?	Weighting spread 0 points (UX) and 1 point (conversion); 5 of 6 repeat sites identical; cumulative mean stable after ~10 evaluations	That the score is correct. Invariance is reliability, not validity; a systematically biased instrument converges just as smoothly as an accurate one
RQ ₂ Does the instrument discriminate and prioritize?	19-point heuristic spread and 13-point factor spread produce a specific, named prioritization; explanatory	Site-level discrimination. 100% of composites fell in one of five bands; the design cannot separate sample

	ratings graded rather than at ceiling (M = 3.59)	composition from instrument compression
RQ3 Are usability and conversion findings related but non-redundant?	$r = 0.40$, $r_2 \approx 0.16$; construct-level alignment of the weakest and strongest dimensions on both sides	An association with realized conversion. Both scores are outputs of one instrument reading one page; shared method variance cannot be excluded

B. Interpreting the ceiling in the reception data. Satisfaction and acceptance results were uniformly high - CSAT 100%, satisfaction $M = 4.80$ of 5, questionnaire $M = 6.55$ of 7, with 91% of ratings at 6 or 7 - and two independent instruments agreeing lends the reception finding weight. The 0.32-point spread across ten distinct constructs, however, means the questionnaire did not discriminate between them: respondents who rated one item highly rated all items highly, the signature of a ceiling effect. The consequence is specific. The reception data support a general acceptance claim but cannot support fine-grained comparative claims that explainability (6.52) is a relative weakness against efficiency (6.72), for instance, is well inside the noise.

Three plausible contributors should be acknowledged rather than argued away: a 50% response rate, in which participants who found the report unhelpful are precisely those most likely to abandon the survey; recruitment through the researcher's own professional network with responses tied to name and email, which raises the likelihood of acquiescence and social-desirability effects; and a respondent profile dominated by education and health professionals rather than UX or conversion practitioners, most of whom had no professional benchmark against which to judge the report critically.

That last point cuts two ways, and the second way is the more interesting. The sample cannot support a claim that the instrument was validated by usability or conversion specialists, since only 16% worked in technology and computing and none in engineering. It does support a different and arguably more useful claim: the reports were rated highly by senior professionals who do not work in usability, which is direct evidence for the design goal of plain-language findings intelligible to non-specialists. Two further results strengthen the reading. Satisfaction was uncorrelated with the score the instrument assigned ($r = -0.04$) and so was the acceptance questionnaire ($r = 0.01$); respondents neither punished the tool for an unflattering verdict nor rewarded it for a favorable one, which supports treating the reception data as a measure of the instrument rather than of the result. The countervailing reading must be stated too: with satisfaction variance near zero, a null correlation is close to arithmetically inevitable, so the

design cannot distinguish an absence of favorability bias from insufficient variance to detect it.

C. The provenance finding. The single most consequential result is that all 50 evaluations were produced by the deterministic modeled engine and none by the generative AI path. The study therefore did not test the AI-assisted audit pipeline; it tested the deterministic fallback. Describing these results as an evaluation of AI-generated usability and conversion findings would be inaccurate, and the distinction is carried explicitly through Sections V, VII, and IX.

This substantially explains the band concentration. A formula-driven engine operating on a limited set of extracted page signals would be expected to produce tightly clustered mid-range output, so the narrow distribution is plausibly a property of the scoring path rather than of the sites or of the instrument's intended AI assessment. Two competing explanations for the concentration remain, and the study cannot separate them: sample composition, since participants selected established, professionally built sites - the most-audited three are widely cited exemplars of interface design - and engine behavior, in which averaging ten sub-scores compresses composites toward the middle of the range. A single deliberately poor control site would separate them immediately and is the cheapest available validity check.

What the study can still legitimately claim is substantial and should be stated affirmatively rather than treated as salvage. The consent-to-report-to-survey pipeline operated end to end across 50 sessions; the reporting interface and export produced output that 25 senior professionals rated highly; the combined usability-and-conversion report format proved usable and useful to non-specialists; the explanatory rating mechanism produced a graded rather than inflated self-assessment; and the scoring architecture- weighting sensitivity, evidence weighting, and the LIFT construct calibration - was shown to be internally consistent and robust to its contested parameters. Those findings concern the instrument's design, workflow, and reception, all of which are independent of which scoring path populated the numbers.

D. Contribution in context. By producing a uniform, criterion-referenced heuristic and conversion profile

across a heterogeneous set of live sites, the protocol responds directly to the observation that no standardized guidelines have existed for evaluating interface touchpoints at scale [57] and to the finding that standardization is critically underrepresented across the automated and remote UX evaluation literature [15]. To the authors' knowledge, no prior peer-reviewed protocol formally unifies Nielsen's ten heuristics with the LIFT model's six factors. The LIFT calibration in Table XII is the study's principal answer to the construct validity objection that LIFT is a practitioner framework never validated as a measurement model: the framework supplies the organizing structure, while each dimension is anchored in peer-reviewed constructs. This is a legitimate and common approach to instrument design, and it converts an acknowledged weakness into a documented design decision - provided the claim is worded precisely, since citing PERVAL as the construct behind value proposition does not mean PERVAL was administered.

For practitioners, the convergence of usability and conversion weaknesses on conceptually related construct pairs implies that remediation prioritized by the integrated protocol may yield compounding benefits, improving task success and conversion readiness through the same design changes rather than through separate workstreams. The present cross-sectional design evaluates sites at a single point in time and does not test whether remediation produces those improvements.

One interpretive nuance deserves preserving. The framing of help and documentation as an opportunity rather than a defect is genuinely warranted: a low score may indicate a real gap, or a site that legitimately requires no documentation. The instrument identifies where documentation is thin, not whether thin documentation is a problem for that particular site. A related disclosure belongs alongside it: H10 is both the lowest-scoring heuristic and the one whose weight in the Nielsen-1994 scheme rests on an author-determined floor, since H9 and H10 fall outside the top-ten factors of the source analysis [4]. The heuristic's mean is independent of its weight, and the sensitivity analysis shows the composite is unaffected, but the overlap is stated here rather than left for a reviewer to surface.

Limitations

Nine limitations bound the interpretation of these findings; they are conditions on scope rather than deficiencies to be explained away. Table XIX maps each to its mitigation and the study that would resolve it.

1) Scoring provenance. All 50 evaluations came from the deterministic modeled path; the generative AI pipeline

the architecture was built around contributed none. No claim about AI-assisted evaluation performance, reliability, or finding quality is supported by this dataset.

2) Cross-sectional design. Each website was assessed at a single point in time, and no interface was remediated and re-evaluated. The study establishes the protocol's diagnostic output and documents baseline patterns; it does not test the causal proposition embedded in the program hypothesis. The results should not be read as evidence that remediation causes the predicted improvement, only as evidence of where remediation would be prioritized.

3) Single automated evaluator. This secures procedural consistency and controls the evaluator effect by construction [24], but at the cost of criterion validity: a deterministic instrument reliably reproduces whatever it scores, whether or not those scores agree with expert human judgment. The study establishes reliability but not agreement with the multi-evaluator tradition from which the heuristics derive [16].

4) Non-independent observations. The 50 evaluations are not 50 independent observations. Six sites supplied 29 of them and one site supplied 15, so every aggregate statistic reported here is weighted toward a handful of repeatedly audited sites and the effective independent sample is nearer 27 sites than 50 evaluations. The repeat audits also partly explain both the narrow distribution and the early convergence of the cumulative mean, which follow arithmetically when nearly a third of observations are re-scores of one site.

5) Restricted discrimination. With 100% of composites in one of five bands and an observed range of 17 points on a 100-point scale, the effective measurement range is a small fraction of the nominal one. Sample composition and instrument compression are both plausible causes, and this design cannot distinguish them.

6) Sampling and generalization. The site set was participant-selected, and the respondent sample was recruited from the researcher's professional network: highly educated (80% bachelor's or above), geographically concentrated (64% Southeast, none from the Midwest or West), occupationally skewed toward education and health rather than technology, and professionally senior. That profile defines the population to which the reception results apply. Collection also ended when the recruitment channel was exhausted rather than at a predetermined target; the convergence evidence in Section VII-L is offered as post hoc justification for sufficiency, not as a planned stopping rule.

7) Ceiling effects and non-response in the reception data. CSAT of 100% and a 0.32-point spread across ten questionnaire constructs sit at the instrument's ceiling, and the 50% response rate means non-response bias cannot be ruled out. Non-anonymous collection raises the possibility of social-desirability effects.

8) Self-referential quality metric. The 0-5 explanatory rating is generated by the same logic that produced the finding, so it is not independent of it and systematic self-favoring bias cannot be excluded. The rating certifies internal coherence, not correctness: a finding rated 5 may still be wrong. Reporting 58% of findings as well-explained is a statement about explanatory completeness, not accuracy.

9) Author-defined parameters and absent behavioral criterion. The heuristic weights, LIFT factor weights, and

criterion bands were author-defined; although the sensitivity analyses show the composites are essentially invariant to the weighting, the weights remain assumptions rather than empirically derived parameters, and no empirical benchmark establishes that 65 or 78 constitutes good conversion quality. The Conversion Quality index is a model-derived, criterion-referenced estimate rather than a measured outcome: no analytics, transaction, or A/B data were collected, so no figure here has been validated against behavioral conversion. Finally, the satisfaction item measured perceived usefulness of the evaluation output, not end-user satisfaction with the audited websites, and should not be read as the latter.

TABLE XIX. Limitations, Mitigations Applied, and Resolving Study

Limitation	Mitigation applied here	Study that resolves it
All records from the modeled path	Provenance logged and disclosed; all affected readings revised in Table XVI	Re-run a matched subset through the AI path and report modeled-versus-AI agreement
No criterion validity	Construct calibration documented (Table XII); profile checked against published patterns	Three-to-five-expert human panel applying the same protocol [16]
No behavioral conversion measure	All conversion output reported as relative and criterion-referenced	Pre-post remediation trial with controlled online experimentation [73]
Author-defined weights and bands	Three-scheme sensitivity analysis on both composites	Weight estimation against measured outcomes; empirical band calibration
Band concentration	Distribution reported in full; competing explanations stated	Include known poor control sites to test the instrument's low end
Non-independent observations	Repeat-audit structure quantified (Table XVII, Fig. 14)	Report site-level alongside evaluation-level statistics; one evaluation per site
Reception ceiling and non-response	Full distributions reported; independence from assigned score tested	Anonymous administration; independent recruitment; a discriminating instrument
Self-rated finding quality	Reported as explanatory completeness, not accuracy	Human reviewers rate a sample of findings on the same 0-5 scale; report agreement
Non-probability sample	No inferential testing conducted	Larger probability-based sample across more industries

Implications

A. Theoretical. The central theoretical contribution is the reframing of usability and conversion not as parallel, independently evaluated dimensions but as nested determinants of the same user behavior. The moderate

association between the composites, together with the conceptually interpretable alignment of the weakest and strongest dimensions on both sides, supports the position that interface obstruction and conversion suppression are frequently co-located properties of the same design decisions, and that an integrated diagnostic lens captures

relationships a single-framework evaluation would miss. This bridges two literatures - usability inspection in the tradition of Nielsen and Molich [16] and conversion optimization in the tradition of Goward [26] - that have historically operated in isolation. The magnitude of the association matters as much as its direction: at $r_2 \approx 0.16$, the two constructs are neither independent nor interchangeable, which is precisely the condition under which a combined instrument earns its cost.

B. Methodological. By operationalizing both frameworks within one deterministic, criterion-referenced instrument, the study offers a reproducible protocol addressing the standardization gap documented in the literature [15], [57]. The criterion-referenced scorecard - anchoring each LIFT factor to a validated construct and interpreting scores against defined bands - provides a transparent, auditable alternative to benchmarking against proprietary or unstable industry conversion averages. Two procedural contributions generalize beyond this instrument: reporting a weighting sensitivity analysis as a standard accompaniment to any composite usability index, and logging and disclosing scoring provenance in any evaluation pipeline that can fall back from a generative to a deterministic path.

C. Practical. The co-variation of usability and conversion weaknesses on conceptually related construct pairs implies that remediation prioritized by the integrated protocol may improve task success and conversion readiness through the same design changes rather than through separate workstreams. The protocol's single-pass, automated nature lowers the cost and expertise barrier to obtaining a structured cross-framework diagnosis, making systematic evaluation accessible to teams without a dedicated usability or conversion specialist, an accessibility claim the reception data support directly, since the respondents who rated the reports highly were largely not specialists. Practitioners should read the profile rather than the composite: the actionable output of this instrument is the ranked dimension table, not the headline score.

Future Research

Five directions follow from the limitations, ordered by the cost of the work and the size of the inferential gap it closes.

1) **AI-path validation.** The immediate next step is to re-run a matched subset of the present sites through the generative AI path and report agreement with the modeled scores, together with the run-to-run variance that non-deterministic generation introduces. This is the cheapest study available, and it converts the present

dataset from a test of the fallback engine into a modeled-versus-AI comparison.

2) **Low-end and control-site testing.** Including deliberately poor sites and known-bad controls would separate the two competing explanations for band concentration, test the instrument's behavior at the low end of the scale, and establish whether the observed 17-point range reflects the sample or the scoring model.

3) **Human-panel criterion validation.** Administering the same unified protocol to a panel of three to five expert evaluators [16] would triangulate the automated scores against human inter-rater judgment and establish the criterion validity that a single deterministic evaluator cannot. The same panel could rate a sample of findings on the 0-5 explanatory scale, converting that metric from self-report into a validated one.

4) **Convergent-validity study.** Correlating the instrument's factor scores with administered, validated instruments (e.g., perceived value, perceived ease of use, and online trust scales) on samples of real users would test whether interface-level scores track the user-perception constructs to which they are theoretically anchored, moving the framework from construct-grounded toward construct-calibrated.

5) **Longitudinal remediation trial.** Finally, the causal hypothesis requires remediating the prioritized barriers the protocol identifies and measuring realized conversion through controlled online experimentation [73], on a larger probability-based sample spanning more industries and with element-level mapping that records each heuristic and conversion flag at the specific interface element it concerns. That design would both test the program hypothesis and quantify the element-level co-occurrence the present study could observe only at the construct level.

Conclusion

This study examined whether two long-separated traditions of interface evaluation (e.g., Nielsen's usability heuristics [4], [16] and the LIFT conversion model [26]) could be integrated into a single standardized, criterion-referenced protocol, and what such integration would reveal. Applied through an automated instrument to 50 evaluations across 27 live websites, the protocol produced usability scores in the good range ($M = 72.24$, $SD = 3.51$) and conversion quality likewise ($M = 72.0$, $SD = 5.90$), both invariant to the weighting schemes used to aggregate them. Diagnostic value appeared in the profile rather than the composite: the ten heuristics spanned 19 points and the six conversion factors 13, yielding a specific and literature-consistent prioritization (e.g., help

and documentation, visibility of system status, clarity, and trust signaling) where the composite alone yielded a single undifferentiated verdict. The two composites were related but not redundant ($r = 0.40$), which is the condition that justifies assessing both in one pass rather than either alone.

The contribution is methodological and theoretical rather than causal. The study offers a reproducible, construct-grounded instrument addressing the standardization gap documented in the literature [15], [57], and it reframes usability and conversion as nested determinants of the same user behavior rather than parallel concerns evaluated in isolation. Two facts bound every claim: all 50 records were produced by the deterministic modeled scoring path rather than the generative AI pipeline, so nothing here evaluates AI-assisted inspection; and six repeat-audited sites supplied 29 of the 50 evaluations, so the effective independent sample is nearer 27 sites. Within those bounds, the work establishes the protocol's internal consistency and diagnostic logic, but not its criterion validity or its capacity to produce the conversion gains a remediation study would test. It provides a usable instrument and a theoretical foundation on which AI-path validation, human-panel triangulation, convergent-validity work, and a longitudinal remediation trial can build toward a validated unified standard for evaluating the usability and conversion performance of digital interfaces.

References

- [1] S. Genne, "Optimizing user experience in high-traffic financial web applications using analytics," *Int. J. Eng. Extended Technol. Res.*, vol. 5, no. 5, pp. 7231–7241, 2023.
- [2] F. Nawir and S. A. Hendrawan, "The impact of website usability and mobile optimization on customer satisfaction and sales conversion rates in e-commerce businesses in Indonesia," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 2, no. 1, pp. 15–30, 2024.
- [3] M. Maguire, "A review of usability guidelines for e-commerce website design," in *Proc. Int. Conf. Human-Computer Interaction*, Cham, Switzerland: Springer, 2023, pp. 30–43.
- [4] J. Nielsen, "Enhancing the explanatory power of usability heuristics," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, ACM, 1994, pp. 152–158, doi: 10.1145/191666.191729.
- [5] J. Nielsen, "10 usability heuristics for user interface design," Nielsen Norman Group, Nov. 15, 2020. [Online]. Available: <https://www.nngroup.com/articles/ten-usability-heuristics/>
- [6] R. Langevin, R. J. Lordon, T. Avrahami, B. R. Cowan, T. Hirsch, and G. Hsieh, "Heuristic evaluation of conversational agents," in *Proc. 2021 CHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–15.
- [7] T. Kumar, X. Tu, and M. Zallio, "How generative AI is reshaping UI/UX design workflows: A systematic review," in *AHFE Int. Conf. Human Factors Des., Eng. Comput.*, AHFE International, 2025, pp. 2438–2452.
- [8] H. Cho et al., "Assessing the usability of a clinical decision support system: Heuristic evaluation," *JMIR Human Factors*, vol. 9, no. 2, p. e31758, 2022.
- [9] S. Ntoa, G. Margetis, M. Antona, and C. Stephanidis, "User experience evaluation in intelligent environments: A comprehensive framework," *Technologies*, vol. 9, no. 2, p. 41, 2021.
- [10] J. Nielsen, "Finding usability problems through heuristic evaluation," in *Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI '92)*, New York, NY, USA: ACM, 1992, pp. 373–380.
- [11] M. Abdul Aziz, M. H. Sani, H. Baharin, S. Zambri, N. I. Jiman, and I. N. Mohd Nasir, "Heuristic evaluation of e-commerce marketplace," *Malaysian J. Comput.*, vol. 7, no. 2, pp. 1178–1187, 2022.
- [12] S. Seethaladevi and M. Felisiya, "A study on the impact of website user experience on conversion rates," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 15, no. 4, 2026.
- [13] H. Farhana, H. Suryana, S. Rizki Aulia, and N. Lizza Tiara, "Website prototype design for macaron sales using the design thinking method," *Digital Frontier: J. Comput. Sci. Innov.*, vol. 1, no. 2, pp. 31–39, 2025.
- [14] J. Wang, C. Ranscombe, and B. Eisenbart, "Prototyping in smart product design: Investigating prototyping tools to support communication in the early stage smart product development," *Int. J. Des. Creativity Innov.*, vol. 11, no. 3, pp. 159–180, 2023.
- [15] J. Liu, "AI in automated and remote UX evaluation: A systematic review (2014–2024)," *Adv. Human-Computer Interaction*, vol. 2025, art. 7442179, 2025, doi: 10.1155/ahci/7442179.
- [16] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, ACM, 1990, pp. 249–256, doi: 10.1145/97243.97281.

- [17] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in *Proc. INTERACT '93 and CHI '93*, ACM, 1993, pp. 206–213, doi: 10.1145/169059.169166.
- [18] J. Nielsen, "Why you only need to test with 5 users," Nielsen Norman Group, Mar. 18, 2000. [Online]. Available: <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/>
- [19] J. Nielsen, "How many test users in a usability study?," Nielsen Norman Group, Jun. 3, 2012. [Online]. Available: <https://www.nngroup.com/articles/how-many-test-users/>
- [20] K. Moran and K. Gordon, "How to conduct a heuristic evaluation," Nielsen Norman Group, Jun. 25, 2023. [Online]. Available: <https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/>
- [21] J. Brooke, "SUS: A 'quick and dirty' usability scale," in *Usability Evaluation in Industry*, P. W. Jordan et al., Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.
- [22] R. Jeffries, J. R. Miller, C. Wharton, and K. Uyeda, "User interface evaluation in the real world: A comparison of four techniques," in *Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI '91)*, ACM, 1991, pp. 119–124, doi: 10.1145/108844.108862.
- [23] K. R. Lynch, "Heuristic evaluation and usability testing of Ohio Area Agency on Aging websites for older adults," M.S. thesis, Ohio Univ., Athens, OH, USA, 2011.
- [24] M. Hertzum and N. E. Jacobsen, "The evaluator effect: A chilling fact about usability evaluation methods," *Int. J. Human-Computer Interaction*, vol. 13, no. 4, pp. 421–443, 2001, doi: 10.1207/S15327590IJHCI304_05.
- [25] Conversion, "The LIFT Model," Conversion.com, 2023. [Online]. Available: <https://conversion.com/framework/the-lift-model/>
- [26] C. Goward, *You Should Test That! Conversion Optimization for More Leads, Sales, and Profit*. Hoboken, NJ, USA: Sybex/John Wiley & Sons, 2012.
- [27] J. C. Sweeney and G. N. Soutar, "Consumer perceived value: The development of a multiple item scale," *J. Retailing*, vol. 77, no. 2, pp. 203–220, 2001, doi: 10.1016/S0022-4359(01)00041-0.
- [28] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quart.*, vol. 13, no. 3, pp. 319–340, 1989, doi: 10.2307/249008.
- [29] M. Blut, D. Chaney, R. Lunardo, R. Mencarelli, and D. Grewal, "Customer perceived value: A comprehensive meta-analysis," *J. Service Res.*, vol. 27, no. 4, pp. 501–524, 2024, doi: 10.1177/10946705231222295.
- [30] P. Pirolli and S. Card, "Information foraging," *Psychol. Rev.*, vol. 106, no. 4, pp. 643–675, 1999, doi: 10.1037/0033-295X.106.4.643.
- [31] D. Di Fatta, D. Patton, and G. Viglia, "The determinants of conversion rates in SME e-commerce websites," *J. Retailing Consumer Services*, vol. 41, pp. 161–168, 2018, doi: 10.1016/j.jretconser.2017.12.008.
- [32] R. Nadlifatin et al., "Determinant of online purchase intention and corporate website contextual relevance," *Cogent Bus. Manage.*, vol. 12, no. 1, art. 2438860, 2025, doi: 10.1080/23311975.2024.2438860.
- [33] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory*. New York, NY, USA: Springer, 2011, doi: 10.1007/978-1-4419-8126-4.
- [34] W. C. McDowell, R. C. Wilson, and C. O. Kile Jr., "An examination of retail website design and conversion rate," *J. Bus. Res.*, vol. 69, no. 11, pp. 4837–4842, 2016, doi: 10.1016/j.jbusres.2016.04.040.
- [35] C. Flavián, M. Guinaliú, and R. Gurrea, "The role played by perceived usability, satisfaction and consumer trust on website loyalty," *Inf. Manage.*, vol. 43, no. 1, pp. 1–14, 2006, doi: 10.1016/j.im.2005.01.002.
- [36] S. A. Qalati, E. G. Vela, W. Li, S. A. Dakhan, T. T. Hong Thuy, and S. H. Merani, "Effects of perceived service quality, website quality, and reputation on purchase intention," *Cogent Bus. Manage.*, vol. 8, no. 1, art. 1869363, 2021, doi: 10.1080/23311975.2020.1869363.
- [37] L. Deng and M. S. Poole, "Affect in web interfaces: A study of the impacts of web page visual complexity and order," *MIS Quart.*, vol. 34, no. 4, pp. 711–730, 2010, doi: 10.2307/25750702.
- [38] Y. Shen, "Distractions in digital reading: A meta-analysis of attentional interference effects," *Frontiers Psychol.*, vol. 16, art. 1671214, 2025, doi: 10.3389/fpsyg.2025.1671214.
- [39] R. B. Cialdini, *Influence: Science and Practice*, 5th ed. Boston, MA, USA: Pearson, 2009.
- [40] B. J. Fogg, "A behavior model for persuasive design," in *Proc. 4th Int. Conf. Persuasive Technol.*, ACM, 2009, art. 40, doi: 10.1145/1541948.1541999.
- [41] C. Fornell, M. D. Johnson, E. W. Anderson, J. Cha, and B. E. Bryant, "The American Customer Satisfaction Index: Nature, purpose, and findings," *J.*

Marketing, vol. 60, no. 4, pp. 7–18, 1996, doi: 10.1177/002224299606000403.

[42] R. L. Oliver, *Satisfaction: A Behavioral Perspective on the Consumer*, 2nd ed. New York, NY, USA: Routledge, 2010.

[43] J. P. Wanous, A. E. Reichers, and M. J. Hudy, "Overall job satisfaction: How good are single-item measures?," *J. Appl. Psychol.*, vol. 82, no. 2, pp. 247–252, 1997, doi: 10.1037/0021-9010.82.2.247.

[44] American Customer Satisfaction Index, "U.S. overall customer satisfaction," ACSI, 2026. [Online]. Available: <https://theacsi.org/>

[45] A. Kirkpatrick, J. O'Connor, A. Campbell, and M. Cooper, Eds., *Web Content Accessibility Guidelines (WCAG) 2.1*. W3C, 2025. [Online]. Available: <https://www.w3.org/TR/WCAG21/>

[46] S. Schmutz, A. Sonderegger, and J. Sauer, "Implementing recommendations from web accessibility guidelines: A comparative study of nondisabled users and users with visual impairments," *Human Factors*, vol. 59, no. 6, pp. 956–972, 2017, doi: 10.1177/0018720817708397.

[47] W3C, "WCAG 2.1: Success Criterion 1.4.3 Contrast (Minimum)," 2025. [Online]. Available: <https://www.w3.org/TR/WCAG21/>

[48] W3C, "WCAG 2.1: Success Criterion 1.4.11 Non-text Contrast," 2018. [Online]. Available: <https://www.w3.org/TR/WCAG21/>

[49] S. Zuffi, C. Brambilla, G. Beretta, and P. Scala, "Human-computer interaction: Legibility and contrast," in *Proc. 14th Int. Conf. Image Anal. Process.*, IEEE, 2007, pp. 629–634, doi: 10.1109/ICIAP.2007.76.

[50] A. Bangor, "Electronic text readability issues for the visually impaired," in *Proc. Human Factors Ergonomics Soc. Annu. Meeting*, vol. 43, no. 23, SAGE, 1999, pp. 1372–1375, doi: 10.1177/154193129904302323.

[51] K. S. Papadopoulos and D. B. Goudiras, "Accessibility assistance for visually-impaired people in digital texts," *Brit. J. Visual Impairment*, vol. 23, no. 2, pp. 75–83, 2005, doi: 10.1177/0264619605054779.

[52] F. E. Sandnes, "Understanding WCAG 2.0 colour contrast requirements through 3D colour space visualisation," in *Proc. Universal Design*, 2016, pp. 366–375.

[53] J. Birch, "Worldwide prevalence of red-green color deficiency," *J. Opt. Soc. Amer. A*, vol. 29, no. 3, pp. 313–320, 2012, doi: 10.1364/JOSAA.29.000313.

[54] J. R. Nuñez, C. R. Anderton, and R. S. Renslow, "Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data," *PLOS ONE*, vol. 13, no. 7, p. e0199239, 2018, doi: 10.1371/journal.pone.0199239.

[55] W3C, *Web Content Accessibility Guidelines (WCAG) 2.2*, 2024. [Online]. Available: <https://www.w3.org/TR/WCAG22/>

[56] R. Korniiuchuk and M. Boryczka, "Conversion rate prediction based on text readability analysis of landing pages," *Entropy*, vol. 23, no. 11, art. 1388, 2021, doi: 10.3390/e2311388.

[57] R. Zimmermann and A. Auinger, "Developing a conversion rate optimization framework for digital retailers—Case study," *J. Marketing Anal.*, vol. 11, no. 2, pp. 233–243, 2023, doi: 10.1057/s41270-022-00161-y.

[58] J. Schotz, S. Ford, M. Gruber, K. Schmidt, S. Soares, and M. Mouloua, "Convergent usability evaluation of the HFES website," *Proc. Human Factors Ergonomics Soc. Annu. Meeting*, vol. 66, no. 1, pp. 1335–1339, 2022.

[59] S. Duvaud, C. Gavella, F. Lisacek, H. Stockinger, V. Ioannidis, and C. Durinx, "Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users," *Nucleic Acids Res.*, vol. 49, pp. W216–W227, 2021.

[60] M. Farzandipour, E. Nabovati, and M. Sadeqi Jabali, "Comparison of usability evaluation methods for a health information system," *BMC Med. Informat. Decis. Making*, vol. 22, art. 157, 2022.

[61] J. Iminibi, K. N. Elliot, A. B. Moko, and B. Okardi, "A systematic review of heuristic evaluation vs think-aloud protocols in detecting usability problems," *Int. J. Adv. Res. Comput. Sci.*, vol. 17, no. 1, pp. 12–14, 2026.

[62] A. Karani, H. Thanki, and S. Achuthan, "Impact of university website usability on satisfaction: A structural equation modelling approach," *Manage. Labour Stud.*, vol. 46, no. 2, pp. 119–138, 2021.

[63] J. Li, K. Yu, X. Bao, X. Liu, and J. Yao, "Patterns of eHealth website user engagement based on cross-site clickstream data," *J. Med. Internet Res.*, vol. 23, no. 8, p. e29299, 2021.

[64] Z. Galavi, S. Norouzi, and R. Khajouei, "Heuristics used for evaluating the usability of mobile health

applications: A systematic literature review," *Digital Health*, vol. 10, 2024, doi: 10.1177/20552076241253559.

[65] Z. Galavi, M. Montazeri, and R. Khajouei, "Which criteria are important in usability evaluation of mHealth applications: An umbrella review," *BMC Med. Inform. Decis. Making*, vol. 24, art. 365, 2024, doi: 10.1186/s12911-024-02738-2.

[66] D. M. Rothschild, C. Needell, J. Veverka, and E. Yom-Tov, "Search conversion journeys and the missed opportunity of associated keywords," *Frontiers Commun.*, vol. 10, 2025, doi: 10.3389/fcomm.2025.1560141.

[67] J. Guo, W. Zhang, and T. Xia, "Impact of shopping website design on customer satisfaction and loyalty," *Sustainability*, vol. 15, no. 8, art. 6347, 2023.

[68] D. Novita and Hanifah, "Optimizing user interface and user experience on e-commerce platforms to improve customer retention," *Jurnal Teknologi dan Open Source*, vol. 8, no. 2, pp. 837–842, 2025.

[69] P. W. Jordan and B. Thomas, "Ecological validity in laboratory based usability evaluations," *Proc. Human Factors Ergonomics Soc. Annu. Meeting*, vol. 38, no. 17, pp. 1009–1013, 1994, doi: 10.1177/154193129403801712.

[70] M. Q. Patton, *Qualitative Research & Evaluation Methods*, 4th ed. Thousand Oaks, CA, USA: SAGE, 2015.

[71] I. Etikan, S. A. Musa, and R. S. Alkassim, "Comparison of convenience sampling and purposive sampling," *Amer. J. Theoretical Appl. Statist.*, vol. 5, no. 1, pp. 1–4, 2016, doi: 10.11648/j.ajtas.20160501.11.

[72] H. Meslem and A. Abbaci, "A practical approach for optimizing the conversion rate of a landing page's visitors," *Marketing Sci. Inspirations*, vol. 17, no. 4, pp. 40–53, 2022, doi: 10.46286/msi.2022.17.4.4.

[73] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: Survey and practical guide," *Data Mining Knowl. Discovery*, vol. 18, no. 1, pp. 140–181, 2009, doi: 10.1007/s10618-008-0114-1.