

Article

Quantifying Visual Complexity in Generative AI-Designed User Interfaces: Information-Theoretic and Structural Associations with Perceived Cognitive Load and Task Performance

Necati Vardar ^{1,*}  and Çağrı Gümüş ² 

¹ Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Sciences, KTO Karatay University, Konya 42020, Türkiye

² Department of Graphic Design, Faculty of Fine Arts and Design, KTO Karatay University, Konya 42020, Türkiye; cagri.gumus@karatay.edu.tr

* Correspondence: necati.vardar@karatay.edu.tr

Abstract

Generative artificial intelligence is increasingly used to produce user interface designs, yet the usability and task-performance implications of AI-generated interfaces remain insufficiently quantified. This study proposes a reproducible evaluation framework combining computational visual complexity metrics with human-centered interface assessment. Twelve user interfaces were evaluated across four scenarios: a mobile health dashboard, a learning management system dashboard, an e-commerce shopping cart, and a university student information portal. Each scenario included three design conditions: human-designed reference interfaces, raw AI-generated interfaces, and prompt-optimized AI-generated interfaces. Computational metrics included grayscale Shannon entropy, spatial edge density, RGB color entropy, RMS contrast, robust contrast, and white-space ratio. A within-subject user study with 62 participants measured perceived cognitive load, perceived visual complexity, task ease, interface evaluation time, and task accuracy. Friedman tests revealed significant differences among the interface conditions for all five user-centered outcomes, with very large effect sizes. Raw AI-generated interfaces were associated with the highest perceived cognitive load, highest perceived visual complexity, lowest task ease, and longest interface evaluation times. Within the tested stimulus set, prompt-optimized AI-generated interfaces showed more favorable user-centered outcomes than raw AI interfaces but remained statistically distinct from human-designed references in perceived cognitive load, perceived visual complexity, task ease, and interface evaluation time. Task accuracy reached 100% for both human-designed and prompt-optimized AI interfaces, whereas raw AI interfaces achieved 63.10%. Interface-level correlations between computational visual metrics and user outcomes were weak to moderate, suggesting that pixel-level visual complexity measures capture only part of the perceived usability burden. Overall, these findings support the potential value of prompt-based HCI constraints in AI-generated interface design, while robust evaluation should integrate computational metrics with human-centered testing.



Academic Editor: Zi Wang

Received: 14 June 2026

Revised: 28 July 2026

Accepted: 3 August 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: generative artificial intelligence; user interface design; human–computer interaction; visual complexity; cognitive load; Shannon entropy; edge density; prompt optimization; usability evaluation; task performance

1. Introduction

Generative artificial intelligence (GenAI) is increasingly reshaping how digital interfaces are conceptualized, prototyped, and evaluated. In user experience (UX) design and human–computer interaction (HCI), recent studies show that GenAI tools are being adopted for ideation, writing support, design exploration, prototyping, and early-stage concept generation [1–5]. This shift has expanded interface design from a primarily tool-mediated process into a prompt-mediated and co-creative workflow, in which human goals, model outputs, and iterative refinements jointly shape the resulting interface [1–5]. At the same time, established principles of human–AI interaction emphasize that AI-supported systems should remain controllable, understandable, reliable, and aligned with human-centered design goals [6–8].

This transformation creates both opportunities and risks for user interface (UI) design. GenAI can accelerate design iteration, reduce entry barriers for non-expert users, and support rapid exploration of alternative visual layouts [1–5]. However, AI-generated interfaces may also introduce usability problems that are not immediately visible from their surface appearance. Raw AI-generated designs may appear visually polished while still containing inconsistent grouping, weak task hierarchy, unclear affordances, excessive decorative elements, or poor prioritization of task-relevant information [2,3,6–8]. Therefore, the visual plausibility of an AI-generated interface should not be equated with usability, cognitive efficiency, or successful task performance.

A central issue in this context is visual complexity. In HCI research, visual complexity has been associated with perceived aesthetics, first impressions, visual search, usability, visual clutter, and cognitive effort [9–19]. Interfaces with dense visual elements, inconsistent spatial organization, high clutter, or excessive variation in color and contrast may increase the effort required to locate, interpret, and act upon relevant information [13–18]. Conversely, visual simplicity does not automatically guarantee better usability, since overly sparse or weakly structured layouts may reduce salience, impair information hierarchy, or fail to guide users toward task-relevant actions [17–21]. Thus, visual complexity should be understood not only as the quantity of visible information, but also as the organization of that information in relation to human perception and task goals.

Computational metrics provide a reproducible way to quantify certain aspects of visual structure. Previous work has operationalized interface visual complexity using measures related to element density, spatial organization, visual clutter, colorfulness, saliency, and interface aesthetics [9–14,18–22]. Entropy-based measures, edge-based descriptors, contrast-related metrics, and white-space indicators can summarize low-level properties of interface screenshots and enable systematic comparison across design alternatives [9–14,20–22]. Such metrics are valuable because they reduce dependence on purely subjective design judgment and make it possible to evaluate AI-generated interfaces through repeatable computational procedures.

However, computational visual metrics alone may not fully capture how users experience an interface. Human users do not process interfaces merely as distributions of pixels, edges, colors, and contrast values. They also interpret semantic labels, grouping cues, interaction expectations, domain conventions, and the relevance of visible elements to their current task [16,17,20,21,23]. This distinction is particularly important for GenAI-designed interfaces, because an output may be visually coherent at the screenshot level while still imposing additional cognitive effort through unclear information hierarchy, ambiguous affordances, or inefficient task pathways [2,3,6–8]. Consequently, the relationship between computational visual complexity and perceived cognitive load may be partial rather than deterministic.

Cognitive load and usability measurement offer a complementary perspective for evaluating these interfaces. Cognitive load theory emphasizes that task performance depends on the mental effort required to process information, maintain relevant elements in working memory, and execute problem-solving actions [24,25]. In HCI, cognitive workload is commonly studied using subjective ratings, behavioral measures, task completion time, task accuracy, and other performance indicators [23,26]. Usability research similarly stresses that interface quality should be assessed through multiple dimensions, including effectiveness, efficiency, satisfaction, and task success rather than through a single perceptual or aesthetic criterion [27]. Therefore, evaluating AI-generated interfaces requires both computational measurements of visual structure and human-centered measures of perceived effort and task performance. Although prior research has examined GenAI-supported design practices [1–5], human-centered AI principles [6–8], visual complexity and interface aesthetics [9–22], and cognitive workload measurement [23–27], these streams remain only partially integrated into a single empirical framework for evaluating AI-generated user interfaces.

Accordingly, the study addressed the following research questions:

RQ1: How do human-designed reference interfaces, raw AI-generated interfaces, and prompt-optimized AI-generated interfaces differ in their computational visual complexity profiles and user-centered outcomes?

RQ2: Within the tested set of interface stimuli, are prompt-optimized AI-generated interfaces associated with lower perceived cognitive load, lower perceived visual complexity, greater task ease, shorter interface evaluation time, and higher task accuracy than raw AI-generated interfaces?

RQ3: To what extent are information-theoretic and structural image metrics associated with perceived cognitive load, perceived visual complexity, task ease, interface evaluation time, and task accuracy?

To address these questions, the present study proposes a reproducible framework for quantifying visual complexity in generative AI-designed user interfaces. The framework combines computational image-based metrics with a within-subject HCI evaluation. Twelve interfaces were evaluated across four applied scenarios: a mobile health dashboard, a learning management system dashboard, an e-commerce shopping cart, and a university student information portal. Each scenario was represented under three design conditions: a human-designed reference interface, a raw AI-generated interface, and a prompt-optimized AI-generated interface. Computational metrics included grayscale Shannon entropy, spatial edge density, RGB color entropy, RMS contrast, robust contrast, and white-space ratio. Human-centered outcomes included perceived cognitive load, perceived visual complexity, task ease, interface evaluation time, and task accuracy.

This study makes three contributions. First, it provides a controlled empirical comparison of human-designed, raw AI-generated, and prompt-optimized AI-generated interfaces across multiple applied task scenarios. Second, it examines associations between information-theoretic and structural visual metrics and the full set of user-centered outcomes in AI-generated UI designs. Third, it examines the value of integrating computational visual analysis with human-centered usability testing for the robust evaluation of generative AI-designed interfaces. By doing so, the study contributes to the emerging literature on GenAI-supported interface design and provides a practical evaluation pathway for assessing AI-generated interfaces beyond surface-level visual appeal.

2. Methodology

2.1. Research Design

This study employed a hybrid empirical design that combined computational visual analysis with a human-centered user evaluation, as summarized in Figure 1. The computational component quantified the visual complexity of finalized interface screenshots using information-theoretic and structural image metrics. The human-centered component evaluated how users perceived and interacted with the same interfaces in a controlled task-based setting.

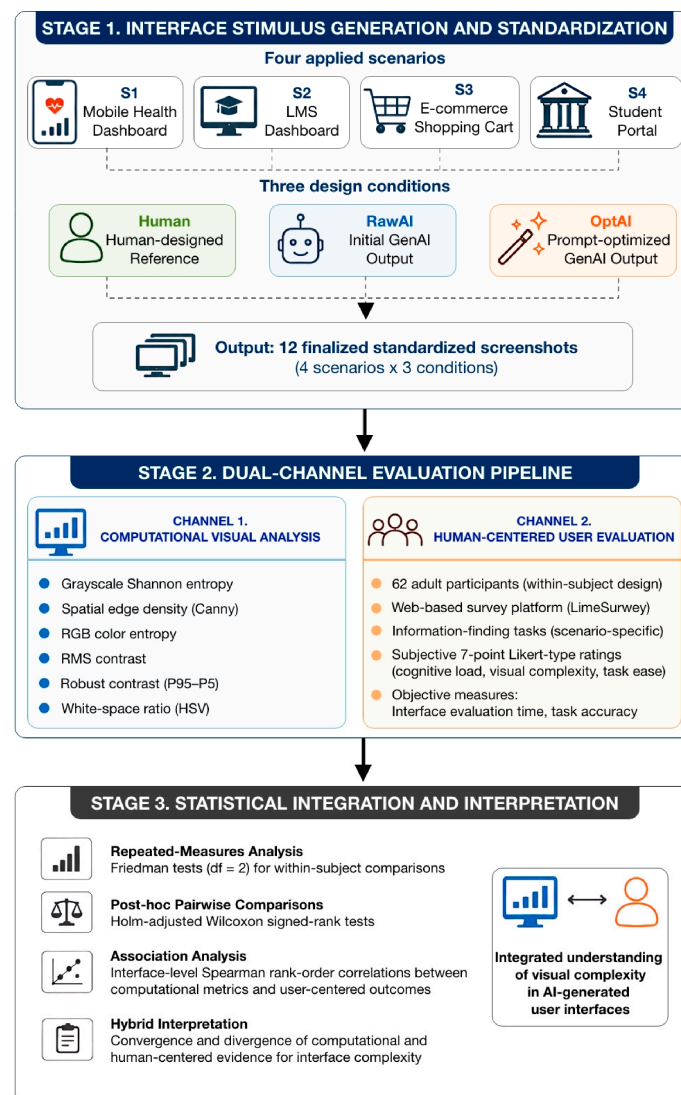


Figure 1. Schematic overview of the proposed hybrid evaluation framework. Twelve interface stimuli were prepared and standardized across four applied scenarios and three design conditions. Each finalized screenshot was evaluated through two complementary channels: computational visual complexity analysis and human-centered user testing. Interface evaluation time was derived from LimeSurvey’s group-level timing variable (groupTime) and represents the total page-level evaluation duration. The resulting image-based metrics and user-centered outcomes were then integrated using repeated-measures and correlation analyses.

The user study followed a within-subject repeated-measures design. Each participant evaluated all twelve interface stimuli, allowing direct comparison of the three design conditions while controlling for individual differences in interface familiarity, digital experience, and response tendencies. The experimental matrix consisted of four applied interface sce-

narios and three design conditions: human-designed reference interfaces, raw AI-generated interfaces, and prompt-optimized AI-generated interfaces.

The stimuli were presented in a fixed mixed order in which scenarios and design conditions were interleaved rather than grouped by condition. This presentation structure was used to reduce condition clustering and to minimize the risk that participants would evaluate multiple interfaces from the same design condition consecutively. Because all participants completed the same set of stimuli, the resulting dataset supported repeated-measures comparisons across design conditions. The study was designed to compare the three interface conditions in terms of perceived cognitive load, perceived visual complexity, task ease, interface evaluation time, and task accuracy.

2.2. Stimuli and Interface Design Scenarios

Twelve interface screenshots were used as experimental stimuli. The stimuli were organized according to a 4×3 matrix: four interface scenarios and three design conditions. The four scenarios represented common digital interaction contexts: a mobile health application dashboard, a learning management system dashboard, an e-commerce shopping cart, and a university student information portal. These scenarios were selected because they involve information search, visual scanning, prioritization of task-relevant content, and decision-making under realistic interface constraints.

For each scenario, three design conditions were prepared. The Human condition consisted of manually designed reference interfaces created using Adobe XD (version 61.0.12.1). The RawAI condition consisted of initial AI-generated interface outputs produced through ChatGPT image generation without systematic prompt refinement beyond the basic scenario request. The OptAI condition consisted of prompt-optimized AI-generated interfaces produced using systematic prompt refinement targeting layout hierarchy, visual alignment, information clarity, spacing, and task salience. In all conditions, the final stimuli were exported as standardized PNG images and used consistently across the computational and user evaluation components.

Mobile-oriented stimuli were prepared at 1080×1920 px, whereas dashboard-oriented desktop/tablet stimuli were prepared at 1440×1024 px. No additional cropping, resizing, or post-hoc visual enhancement was applied after final stimulus standardization. The final stimulus mapping is shown in Table 1. Additional details regarding the scenario structure, stimulus generation focus, and task protocol for each interface are provided in Supplementary Table S1.

Table 1. Experimental stimulus matrix.

Interface	Scenario	Scenario Description	Condition	Image Resolution
1	S1	Mobile health application dashboard	Human	1080×1920 px
2	S2	Learning management system dashboard	RawAI	1440×1024 px
3	S3	E-commerce shopping cart	OptAI	1080×1920 px
4	S4	University student information portal	Human	1440×1024 px
5	S1	Mobile health application dashboard	RawAI	1080×1920 px
6	S2	Learning management system dashboard	OptAI	1440×1024 px
7	S3	E-commerce shopping cart	Human	1080×1920 px
8	S4	University student information portal	RawAI	1440×1024 px
9	S1	Mobile health application dashboard	OptAI	1080×1920 px
10	S2	Learning management system dashboard	Human	1440×1024 px
11	S3	E-commerce shopping cart	RawAI	1080×1920 px
12	S4	University student information portal	OptAI	1440×1024 px

Stimulus Generation, Prompting, and Standardization

The AI-generated interface stimuli were created using ChatGPT Images 2.0 (OpenAI, San Francisco, CA, USA), accessed in May 2026 through the ChatGPT web interface. A more specific backend model or build identifier was not displayed to the user through the ChatGPT interface and therefore could not be documented. The same image-generation system was used for all RawAI and OptAI stimuli to maintain procedural consistency across the experimental conditions.

For each of the four interface scenarios, separate prompts were prepared for the RawAI and OptAI conditions. The RawAI prompts specified the relevant interface context, required functional elements, target display format, and English-language presentation but did not include systematic instructions concerning usability, information architecture, or visual organization. The OptAI prompts retained the same scenario-level functional requirements while incorporating explicit human–computer interaction and interface-design constraints. These constraints addressed visual hierarchy, information grouping, alignment, spacing, typography, readability, contrast, white-space use, task-relevant salience, affordance clarity, and the reduction of unnecessary visual clutter. The complete verbatim prompts used to generate the RawAI and OptAI stimuli for all four scenarios are provided in Supplementary Table S4.

The generated outputs used in the experiment were retained in their original generated form. No semantic, textual, typographic, compositional, or content-level post-editing was performed. Interface labels, numerical values, icons, colors, element positions, spacing, visual hierarchy, and task-relevant information were not manually corrected, replaced, repositioned, or otherwise modified after generation. No additional visual enhancement, retouching, or reconstruction was applied. The final stimuli were saved in PNG format and standardized according to the target interface orientation: mobile-oriented stimuli were presented at 1080×1920 pixels, whereas desktop- or dashboard-oriented stimuli were presented at 1440×1024 pixels. The same finalized images were used both for the computational visual complexity analyses and for the human-centered interface evaluation, ensuring that all measurements were based on identical stimulus files.

2.3. Computational Visual Complexity Metrics

Computational visual complexity was quantified from the finalized interface screenshots using six image-based metrics. These metrics were selected to capture complementary aspects of low-level visual structure: information variability, spatial structure, color distribution, contrast, and empty visual space. All metrics were calculated using the final standardized PNG images. The operational definitions and parameter settings of these metrics are summarized in Table 2.

Grayscale Shannon entropy was used as an information-theoretic measure of pixel intensity variability. Each image was converted to grayscale, and entropy was computed from the normalized grayscale intensity histogram according to Shannon's entropy formulation [28]. Higher entropy values indicate greater variability in pixel intensity distribution. Spatial edge density was used as a structural measure of interface complexity. Edges were detected using the Canny edge detection algorithm [29]. To ensure reproducibility, the low and high Canny thresholds were fixed at 100 and 200, respectively. Edge density was calculated as the ratio of detected edge pixels to the total number of pixels in the image. RGB color entropy was used to quantify variability in the color distribution. The RGB color space was discretized into 8 bins per channel, resulting in 512 possible color bins. Entropy was then calculated from the distribution of pixels across these bins. This metric captures the diversity and irregularity of color usage within the interface.

Table 2. Computational visual complexity metrics.

Metric	Operational Definition	Parameter/Computation
Grayscale Shannon Entropy	Information-theoretic variability of grayscale pixel intensities	Computed from grayscale intensity histogram as $H = - \sum_i p_i \log_2 (p_i)$
Spatial Edge Density	Structural density of edge pixels in the interface image	Canny edge detection; low threshold = 100, high threshold = 200; edge pixels/total pixels
RGB Color Entropy	Variability of color distribution in RGB space	8 bins per RGB channel; 512 possible color bins
RMS Contrast	Standard deviation of normalized grayscale luminance	Computed from normalized grayscale image as RMS contrast = standard deviation of grayscale luminance
Robust Contrast	Robust luminance spread	Computed as P95 – P5, where P95 and P5 represent the 95th and 5th percentile grayscale luminance values
White-Space Ratio	Estimated proportion of visually empty or light background regions	HSV thresholding; saturation < 35 and value > 235; white-space ratio = white-space-like pixels/total pixels

Two contrast metrics were also computed. RMS contrast was calculated as the standard deviation of normalized grayscale luminance values. Robust contrast was calculated as the difference between the 95th and 5th percentiles of grayscale luminance, reducing the influence of extreme outlier pixels. Finally, white-space ratio was estimated in the HSV color space using a threshold-based procedure. Pixels with saturation below 35 and value above 235 were classified as white-space-like regions, and the white-space ratio was calculated as the proportion of such pixels relative to the total image area.

2.4. Participants

Participants were recruited using a non-probability convenience sampling strategy. The survey invitation was distributed through institutional academic networks and online communication channels to adult volunteers who regularly used digital interfaces. Recruitment materials briefly described the study as an interface evaluation task and invited participants to complete a web-based survey anonymously. Participation was voluntary, and no personally identifying information was collected. The sampling strategy was considered appropriate for the exploratory within-subject design of the study because the primary objective was to compare user responses across controlled interface conditions rather than to estimate population-level prevalence parameters. Eligibility required participants to be 18 years of age or older and to complete all twelve interface pages with sufficient evaluable data for the repeated-measures analyses. The final analytic sample comprised 62 adult participants who met these criteria.

Participants represented a digitally experienced sample. All participants reported frequent daily interaction with digital interfaces. The sample included participants from engineering and technology, design and visual communication, health sciences, and social sciences backgrounds. Table 3 summarizes the participant characteristics included in the final dataset.

Table 3. Participant characteristics.

Variable	Category	<i>n</i>	%
Age	18–24	24	38.7
Age	25–34	24	38.7
Age	35–44	14	22.6
Gender	Female	37	59.7
Gender	Male	25	40.3
Field	Engineering/Technology	20	32.3
Field	Design/Visual communication/Graphic Design	14	22.6
Field	Health Sciences	9	14.5
Field	Social Sciences	19	30.6
Daily digital interface use	Almost continuously	62	100.0
UI/design experience	No	39	62.9
UI/design experience	Yes, beginner level	23	37.1
AI tool experience	No	30	48.4
AI tool experience	Yes, used a few times	32	51.6
Device used for participation	Desktop Computer	18	29.0
Device used for participation	Laptop Computer	16	25.8
Device used for participation	Tablet	28	45.2

The study was conducted following informed consent procedures. Participants were informed about the purpose of the study, the voluntary nature of participation, and the anonymous processing of their responses. The study protocol was reviewed and approved by the Human Research Ethics Committee of KTO Karatay University, Konya, Türkiye, on 30 April 2026 (Meeting No. 04, Decision No. 34).

2.5. Experimental Procedure and Tasks

The user evaluation was conducted through LimeSurvey, a web-based survey platform. The survey structure consisted of five main stages: informed consent, demographic questions, task instructions, interface evaluation tasks, and final general evaluation items. After providing informed consent, participants completed demographic questions related to age, gender, field of study or expertise, daily digital interface use, prior UI/design experience, AI tool experience, and device type. Prior UI/design experience and AI tool experience were collected to characterize the sample and to identify possible experience-related sources of response bias. These variables were not used as exclusion criteria; rather, they were recorded to improve transparency regarding participants' background familiarity with interface design and generative AI tools. Participants were then instructed to evaluate each interface by completing short information-finding tasks and rating their subjective experience. Each of the twelve interface pages presented one interface screenshot and two task questions requiring participants to locate specific information or identify a task-relevant action within the interface.

Task equivalence was controlled at the scenario level. Within each scenario, the Human, RawAI, and OptAI interfaces were evaluated using the same two functional information-seeking goals, the same number of required responses, and the same response format. For example, all three mobile health interfaces required participants to identify the displayed daily step count and heart-rate value, whereas all three shopping-cart interfaces required participants to identify the displayed total price and locate the checkout action. The exact numerical values, labels, or dates shown in the interfaces were stimulus-specific;

therefore, correct responses were defined separately for each stimulus. Equivalence referred to the semantic category and functional goal of the task rather than to identical answer values or identical visual presentation. Semantic task demand was held constant within each scenario, whereas visual search difficulty and target salience were allowed to vary as characteristics of the interface design condition.

After completing the task responses, participants rated the interface using 7-point Likert-type items measuring perceived cognitive load, perceived visual complexity, and task ease. The Likert scale ranged from 1 to 7, with higher values indicating greater perceived cognitive load, greater perceived visual complexity, or greater task ease, respectively.

Interface evaluation time was recorded automatically through LimeSurvey's group-level timing variable (groupTime). For each interface page, the corresponding groupTime value was extracted and converted into seconds with decimal precision. The timer began when the interface page was presented to the participant and ended when the participant submitted the page and proceeded to the next survey page. Consequently, the recorded duration included visual inspection of the interface, completion of the two information-finding questions, completion of the three subjective rating items, and any pause occurring before page submission. The measure is therefore referred to as interface evaluation time rather than isolated task-completion time and should be interpreted as a page-level behavioral duration measure reflecting the overall interface evaluation process. Because timing was captured automatically in the background, it did not rely on participant self-report. The interfaces belonging to the same scenario were presented at the same target resolution, which reduced potential resolution-related differences across design conditions. However, LimeSurvey did not provide a separate image-loading time variable; therefore, possible loading latency could not be isolated from the total interface evaluation time.

Task accuracy was calculated from the two task responses associated with each interface. Each response was coded as correct or incorrect according to a predefined answer key. Blank or unanswered task responses were coded as incorrect. The stimulus-specific answer key specified the exact numerical response, label, date, or interface action accepted for each of the twenty-four task questions. Numerical formatting variations, including the use or omission of thousands separators and currency symbols, were accepted when the underlying value was correct. Text responses were evaluated case-insensitively, and semantically equivalent labels were accepted where applicable. Interface-level task accuracy was computed as the average of the two binary task scores, resulting in possible values of 0, 0.5, or 1 for each participant-interface observation. Condition-level task accuracy was then calculated by averaging interface-level task accuracy values across interfaces within each design condition. The complete stimulus-specific answer key and accuracy coding criteria are provided in Supplementary Table S3.

2.6. Dependent Measures and Statistical Analysis

The study included both subjective and objective dependent measures. The subjective measures were perceived cognitive load, perceived visual complexity, and task ease, each measured using 7-point Likert-type items. The objective measures were interface evaluation time and task accuracy. Interface evaluation time was derived from LimeSurvey groupTime variables and expressed in seconds. Interface evaluation times were screened separately for each of the twelve interfaces using Tukey's $1.5 \times$ interquartile range criterion to identify unusually short or long observations. Task accuracy was derived from binary coding of the two task responses per interface and expressed as both a proportion and a percentage. Because the subjective outcomes were measured using Likert-type items and distributional assumptions were not imposed for the repeated-measures comparisons, non-parametric methods were used consistently across all outcomes [30]. For

each outcome, participant-level condition means were calculated for the Human, RawAI, and OptAI conditions. To strengthen the interpretability of the main descriptive findings, 95% confidence intervals (CIs) were also calculated for participant-level condition means. Two-sided 95% confidence intervals for participant-level condition means were calculated using the t distribution as $M \pm t_{0.975, n-1} (SD/\sqrt{n})$. These CIs were reported alongside means and standard deviations for the main user-centered outcomes, including perceived cognitive load, perceived visual complexity, task ease, interface evaluation time, and task accuracy. Differences among the three interface conditions were analyzed using Friedman tests, which provide a non-parametric alternative to repeated-measures ANOVA for related samples [31]. Kendall's W was reported as the effect size for the Friedman tests. Pairwise post-hoc comparisons were then conducted using Wilcoxon signed-rank tests [32]. Holm adjustment was applied to control the family-wise error rate across multiple post-hoc comparisons [33]. Effect sizes for the Wilcoxon signed-rank comparisons were reported as r values calculated as $r = |Z|/\sqrt{N}$, where N represented the number of paired observations. Associations between computational visual complexity metrics and user-centered outcomes were examined using Spearman rank-order correlations [34]. Spearman correlations were selected because they do not require normally distributed variables and are appropriate for monotonic associations between visual metrics and perceptual or behavioral outcomes. Correlation analyses were conducted at the interface level by linking each interface's computational metrics with its corresponding aggregated user-centered outcomes.

All analyses were conducted using Python 3.12.13 (Python Software Foundation, Wilmington, DE, USA). Data processing included survey export cleaning, interface-condition mapping, task accuracy coding, descriptive summaries, repeated-measures tests, post-hoc comparisons, and correlation analyses.

Statistical significance was evaluated using an alpha level of 0.05. Exact p -values were reported where appropriate; p -values below 0.001 were reported as $p < 0.001$.

3. Results

3.1. Computational Analysis of Interface Stimuli

Computational visual metrics were calculated for all twelve finalized interface stimuli. Table 4 presents the interface-level values for grayscale Shannon entropy, spatial edge density, RGB color entropy, RMS contrast, robust contrast, and white-space ratio. At the condition level, mean grayscale Shannon entropy was 3.918 for Human, 4.091 for RawAI, and 3.315 for OptAI interfaces. Mean spatial edge density was 0.041 for Human, 0.032 for RawAI, and 0.027 for OptAI interfaces. Mean RGB color entropy was 1.151 for Human, 1.177 for RawAI, and 0.963 for OptAI interfaces. Mean RMS contrast was 0.205 for Human, 0.230 for RawAI, and 0.185 for OptAI interfaces. Mean robust contrast was 0.522 for Human, 0.631 for RawAI, and 0.512 for OptAI interfaces. Mean white-space ratio was 0.835 for Human, 0.840 for RawAI, and 0.882 for OptAI interfaces.

The standardized computational visual metric profiles of Human, RawAI, and OptAI interfaces are provided in Supplementary Figure S1.

3.2. Human-Centered Evaluation and Task Performance

A total of 62 participants completed the within-subject evaluation. Each participant evaluated all twelve interfaces, resulting in 744 participant-interface observations and 248 observations per design condition. Descriptive statistics for the user-centered outcomes are presented in Table 5. Figure 2 shows participant-level condition means for perceived cognitive load, perceived visual complexity, task ease, and interface evaluation time.

Table 4. Interface-level computational visual complexity metrics.

Interface	Scenario	Condition	Shannon Entropy	Edge Density	RGB Color Entropy	RMS Contrast	Robust Contrast	White-Space Ratio
1	S1	Human	4.183	0.031	1.392	0.256	0.612	0.733
2	S2	RawAI	4.005	0.042	1.447	0.303	0.871	0.784
3	S3	OptAI	3.507	0.019	1.012	0.199	0.753	0.875
4	S4	Human	3.983	0.053	1.195	0.185	0.420	0.858
5	S1	RawAI	3.910	0.028	0.705	0.120	0.208	0.929
6	S2	OptAI	2.904	0.032	0.688	0.113	0.212	0.932
7	S3	Human	3.759	0.035	1.069	0.221	0.784	0.844
8	S4	RawAI	4.358	0.034	1.512	0.327	0.882	0.764
9	S1	OptAI	3.211	0.025	0.763	0.117	0.224	0.931
10	S2	Human	3.748	0.044	0.949	0.158	0.271	0.907
11	S3	RawAI	4.092	0.025	1.045	0.167	0.565	0.884
12	S4	OptAI	3.638	0.031	1.391	0.310	0.859	0.791

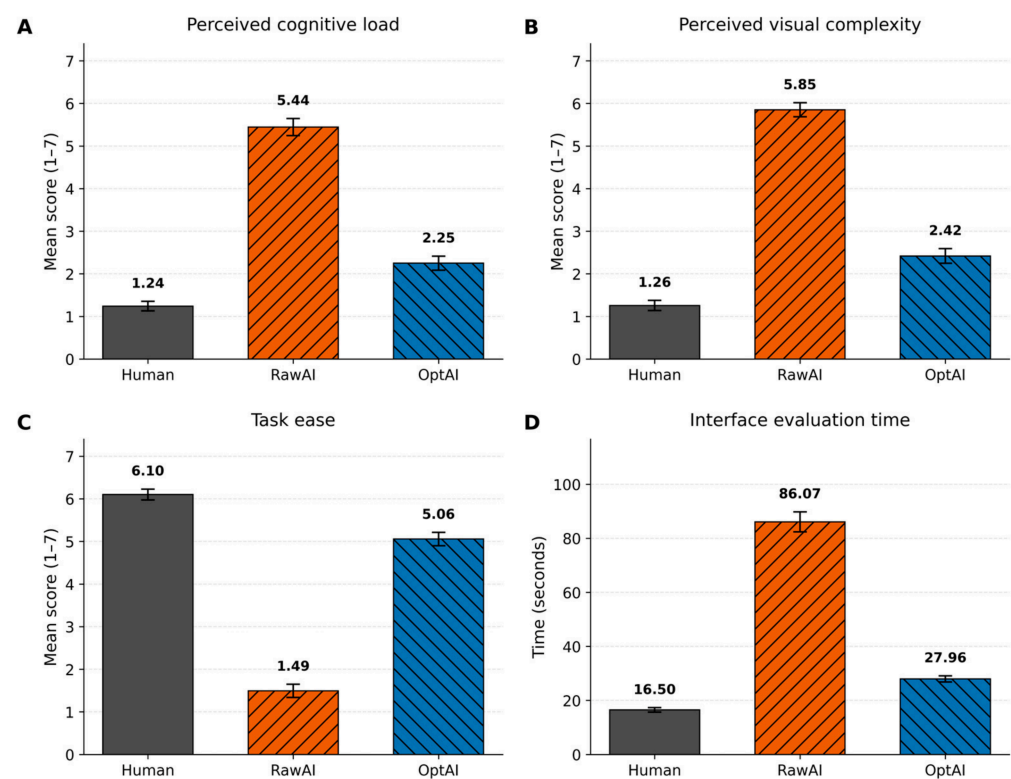


Figure 2. User-centered outcomes by interface condition: (A) perceived cognitive load; (B) perceived visual complexity; (C) task ease; and (D) interface evaluation time. Bars represent participant-level condition means, and error bars indicate 95% confidence intervals. Higher scores indicate greater perceived cognitive load, greater perceived visual complexity, and greater task ease, respectively. Interface evaluation time is reported in seconds.

To provide more precise estimates of the main user-centered outcomes, 95% confidence intervals were added to the descriptive results. The intervals supported a clear descriptive separation among the three design conditions for perceived cognitive load, perceived visual complexity, task ease, and interface evaluation time. RawAI interfaces showed substantially higher perceived cognitive load, perceived visual complexity, and interface evaluation time than both Human and OptAI interfaces, whereas Human interfaces showed the highest task ease and the shortest interface evaluation time. Task accuracy was identical for Human and OptAI interfaces and lower for RawAI interfaces.

Table 5. Descriptive statistics for user-centered outcomes by interface condition.

Condition	<i>n</i>	Cognitive Load M ± SD [95% CI]	Visual Complexity M ± SD [95% CI]	Task Ease M ± SD [95% CI]	Interface Evaluation Time M ± SD [95% CI], s	Task Accuracy M ± SD [95% CI]	Task Accuracy (%)
Human	62	1.242 ± 0.169 [1.199–1.285]	1.258 ± 0.244 [1.196–1.320]	6.101 ± 0.258 [6.035–6.166]	16.495 ± 1.671 [16.071–16.920]	1.000 ± 0.000 [1.000–1.000]	100.00
RawAI	62	5.444 ± 0.411 [5.339–5.548]	5.851 ± 0.266 [5.783–5.918]	1.492 ± 0.307 [1.414–1.570]	86.068 ± 7.654 [84.124–88.011]	0.631 ± 0.142 [0.595–0.667]	63.10
OptAI	62	2.250 ± 0.336 [2.165–2.335]	2.419 ± 0.306 [2.342–2.497]	5.056 ± 0.305 [4.979–5.134]	2.961 ± 2.003 [27.453–28.470]	1.000 ± 0.000 [1.000–1.000]	100.00

Note: *n* = number of participants. CI = confidence interval. Cognitive load and visual complexity denote participants' perceived ratings. Higher scores indicate greater perceived cognitive load, greater perceived visual complexity, and greater task ease, respectively. Interface evaluation time is reported in seconds. Task accuracy is reported both as a proportion and as a percentage. Means, standard deviations, and two-sided 95% confidence intervals were calculated across participant-level condition means (*n* = 62), with the confidence intervals estimated using the *t* distribution.

3.2.1. Perceived Cognitive Load, Perceived Visual Complexity, and Task Ease

Friedman tests were conducted to compare the three interface conditions for perceived cognitive load, perceived visual complexity, and task ease. Significant differences were observed for perceived cognitive load, $\chi^2(2) = 124.00$, $p < 0.001$, Kendall's *W* = 1.000; perceived visual complexity, $\chi^2(2) = 124.00$, $p < 0.001$, Kendall's *W* = 1.000; and task ease, $\chi^2(2) = 124.00$, $p < 0.001$, Kendall's *W* = 1.000. The Kendall's *W* value of 1.000 should be interpreted in relation to the participant-level condition means used in the repeated-measures analysis. Across participants, the three interface conditions showed a fully consistent ordinal pattern: RawAI was associated with higher perceived cognitive load and perceived visual complexity than OptAI and Human interfaces, whereas Human interfaces produced the highest task ease, followed by OptAI and RawAI. This pattern reflects the large separation among the three design conditions rather than identical raw responses at the individual item level. Holm-adjusted Wilcoxon signed-rank post-hoc comparisons are presented in Table 6. All pairwise comparisons were significant for perceived cognitive load, perceived visual complexity, and task ease.

Table 6. Pairwise post-hoc comparisons for subjective outcomes.

Outcome	Comparison	Mean A	Mean B	Difference B – A	Wilcoxon W	Holm-Adjusted <i>p</i>	Effect Size <i>r</i>
Cognitive load	Human vs. RawAI	1.242	5.444	4.202	0.000	<0.001	0.869
Cognitive load	Human vs. OptAI	1.242	2.250	1.008	0.000	<0.001	0.869
Cognitive load	RawAI vs. OptAI	5.444	2.250	−3.194	0.000	<0.001	0.869
Visual complexity	Human vs. RawAI	1.258	5.851	4.593	0.000	<0.001	0.869
Visual complexity	Human vs. OptAI	1.258	2.419	1.161	0.000	<0.001	0.869
Visual complexity	RawAI vs. OptAI	5.851	2.419	−3.431	0.000	<0.001	0.869
Task ease	Human vs. RawAI	6.101	1.492	−4.609	0.000	<0.001	0.869
Task ease	Human vs. OptAI	6.101	5.056	−1.044	0.000	<0.001	0.869
Task ease	RawAI vs. OptAI	1.492	5.056	3.565	0.000	<0.001	0.869

Note: Cognitive load and visual complexity denote participants' perceived ratings. All three subjective outcomes were measured using 7-point Likert-type items. Higher values indicate greater perceived cognitive load, greater perceived visual complexity, and greater task ease, respectively. Holm-adjusted *p*-values are reported for the Wilcoxon signed-rank comparisons.

3.2.2. Objective Outcomes: Interface Evaluation Time and Task Accuracy

Interface evaluation time differed significantly among the three interface conditions, $\chi^2(2) = 124.00$, $p < 0.001$, Kendall's *W* = 1.000. Mean interface evaluation time was 16.495 s for Human interfaces, 86.068 s for RawAI interfaces, and 27.961 s for OptAI interfaces.

Holm-adjusted Wilcoxon signed-rank tests showed significant differences for Human vs. RawAI, Human vs. OptAI, and RawAI vs. OptAI comparisons, all $p < 0.001$.

Interface-specific screening using Tukey's $1.5 \times$ interquartile range criterion did not identify any unusually short or long interface evaluation times among the 744 observations. Therefore, all 744 timing observations were retained in the primary analysis.

Task accuracy also differed significantly among the three interface conditions, $\chi^2(2) = 120.00$, $p < 0.001$, Kendall's $W = 0.968$. Mean task accuracy was 100.00% for Human interfaces, 63.10% for RawAI interfaces, and 100.00% for OptAI interfaces. These condition-level task accuracy results are illustrated in Figure 3. Pairwise comparisons showed significant differences between Human and RawAI interfaces and between RawAI and OptAI interfaces. The pairwise post-hoc comparisons for interface evaluation time and task accuracy are presented in Table 7. Interface-level descriptive values for all user-centered outcomes are shown in Table 8.

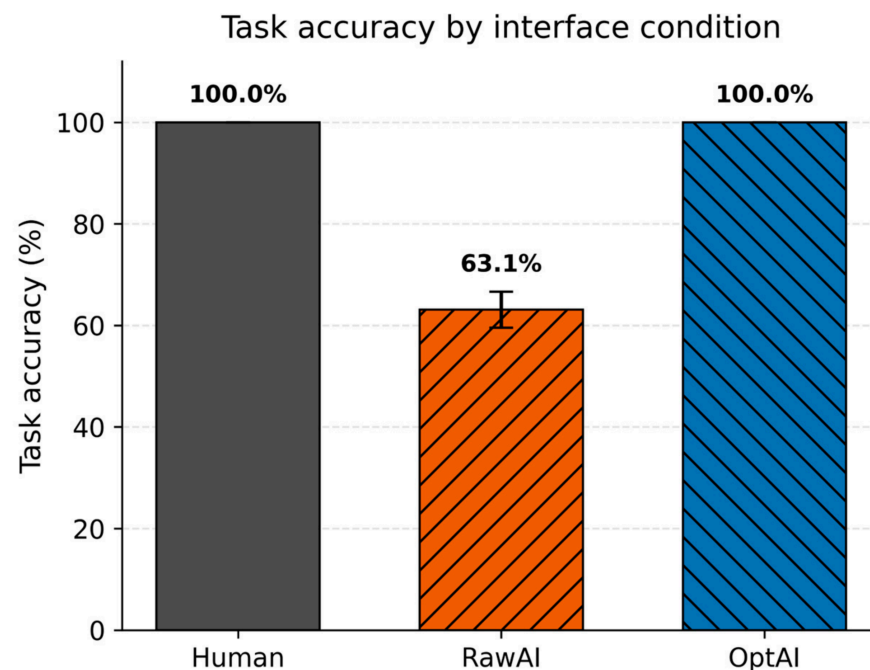


Figure 3. Task accuracy by interface condition. Bars represent participant-level condition means calculated from two binary task responses per interface, and error bars indicate 95% confidence intervals. Human-designed and prompt-optimized AI interfaces achieved 100% task accuracy, compared with 63.10% for raw AI-generated interfaces.

Table 7. Pairwise post-hoc comparisons for objective outcomes.

Outcome	Comparison	Mean A	Mean B	Difference B – A	Wilcoxon W	Holm-Adjusted p	Effect Size r
Interface evaluation time	Human vs. RawAI	16.495	86.068	69.572	0.000	<0.001	0.869
Interface evaluation time	Human vs. OptAI	16.495	27.961	11.466	0.000	<0.001	0.869
Interface evaluation time	RawAI vs. OptAI	86.068	27.961	−58.106	0.000	<0.001	0.869
Task accuracy	Human vs. RawAI	1.000	0.631	−0.369	0.000	<0.001	0.870
Task accuracy	Human vs. OptAI	1.000	1.000	0.000	0.000	1.000	0.000
Task accuracy	RawAI vs. OptAI	0.631	1.000	0.369	0.000	<0.001	0.870

Note: Interface evaluation time is reported in seconds, whereas task accuracy is reported as a proportion. Holm-adjusted p -values are reported for the Wilcoxon signed-rank comparisons.

Table 8. Interface-level user-centered outcomes.

Interface	Scenario	Condition	Cognitive Load	Visual Complexity	Task Ease	Interface Evaluation Time, s	Task Accuracy
1	S1	Human	1.194	1.387	5.984	17.913	1.000
2	S2	RawAI	5.242	5.661	1.613	85.419	0.556
3	S3	OptAI	2.290	2.403	4.984	29.844	1.000
4	S4	Human	1.306	1.226	6.129	16.845	1.000
5	S1	RawAI	5.387	5.79	1.500	83.023	0.645
6	S2	OptAI	2.306	2.581	5.000	28.024	1.000
7	S3	Human	1.242	1.194	6.210	15.376	1.000
8	S4	RawAI	5.548	5.935	1.500	88.998	0.323
9	S1	OptAI	2.145	2.242	5.097	28.062	1.000
10	S2	Human	1.226	1.226	6.081	15.846	1.000
11	S3	RawAI	5.597	6.016	1.355	86.831	1.000
12	S4	OptAI	2.258	2.452	5.145	25.916	1.000

Note: Cognitive load and visual complexity refer to participants' perceived ratings measured on 7-point Likert-type scales. Higher values indicate greater perceived cognitive load, greater perceived visual complexity, and greater task ease, respectively. Interface evaluation time represents the mean page-level evaluation duration and is reported in seconds. Task accuracy is reported as a proportion.

3.3. Associations Between Computational Metrics and User Outcomes

Spearman rank-order correlations were calculated between the six computational visual metrics and the aggregated user-centered outcomes at the interface level. The resulting correlation coefficients are shown in Table 9 and visualized in Figure 4. The corresponding *p*-value matrix is provided in Supplementary Table S2.

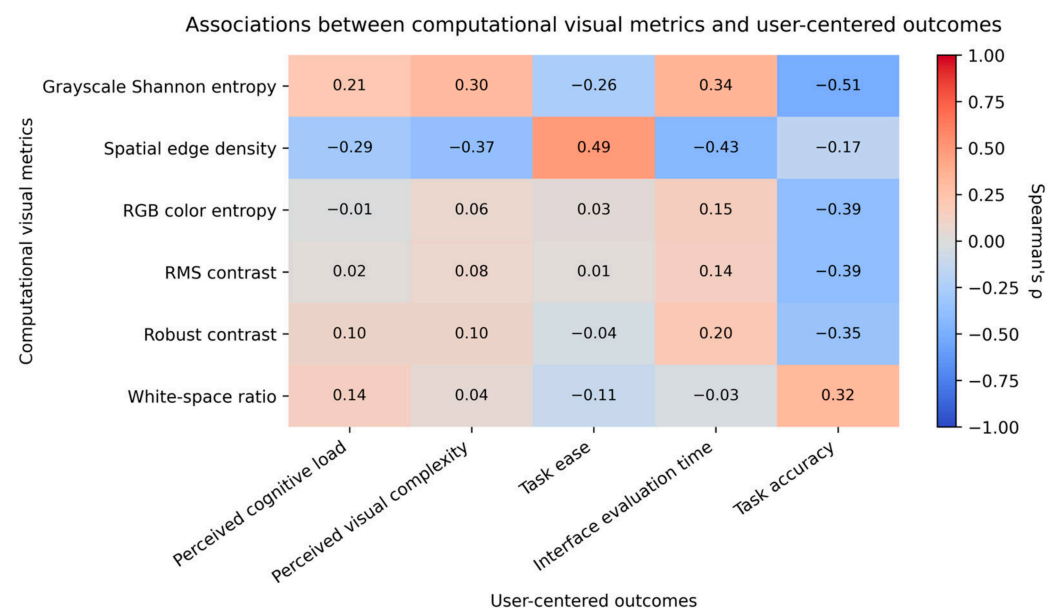


Figure 4. Spearman correlation heatmap between computational visual complexity metrics and user-centered outcomes. Correlations were calculated at the interface level by linking each interface's computational metrics with aggregated user-centered outcomes.

Across the correlation matrix, absolute Spearman coefficients ranged from 0.007 to 0.514. None of the interface-level correlations reached statistical significance ($p < 0.05$). The lowest *p*-value was observed for the association between grayscale Shannon entropy and task accuracy (Spearman's $\rho = -0.514$, $p = 0.087$).

Table 9. Spearman correlations between computational visual metrics and user-centered outcomes.

Computational Metric	Cognitive Load	Visual Complexity	Task Ease	Interface Evaluation Time	Task Accuracy
Grayscale Shannon entropy	0.210	0.298	−0.263	0.336	−0.514
Spatial edge density	−0.294	−0.368	0.490	−0.427	−0.174
RGB color entropy	−0.007	0.063	0.028	0.154	−0.395
RMS contrast	0.021	0.077	0.011	0.140	−0.386
Robust contrast	0.105	0.102	−0.039	0.203	−0.349
White-space ratio	0.140	0.039	−0.112	−0.035	0.321

Note: Values are Spearman rank-order correlation coefficients calculated across the twelve interface stimuli ($n = 12$). Cognitive load and visual complexity denote aggregated perceived ratings, whereas interface evaluation time denotes aggregated page-level evaluation duration. None of the correlations reached $p < 0.05$; corresponding p -values are provided in Supplementary Table S2.

4. Discussion

This study examined whether computational visual complexity metrics and human-centered usability outcomes can jointly characterize the quality of generative AI-designed user interfaces. The findings show a consistent pattern across subjective and objective measures. Within the tested stimulus set, RawAI interfaces were associated with the highest perceived cognitive load, highest perceived visual complexity, lowest task ease, longest interface evaluation times, and the lowest task accuracy. In contrast, prompt-optimized AI interfaces showed more favorable user-centered outcomes than raw AI outputs. Notably, task accuracy reached 100% for both human-designed and prompt-optimized AI interfaces, whereas raw AI-generated interfaces achieved 63.10%. However, prompt-optimized AI interfaces remained statistically distinct from human-designed references in perceived cognitive load, perceived visual complexity, task ease, and interface evaluation time. This pattern is consistent with the potential usability benefits of incorporating explicit HCI constraints into generation prompts; however, the present design does not establish prompt wording as the sole causal factor, and the observed differences indicate that usability-related differences relative to human-designed references remained.

4.1. Observed Usability Differences Between RawAI and Prompt-Optimized Interfaces

The first major finding concerns the difference between raw and prompt-optimized AI-generated interfaces. Raw AI-generated interfaces appeared visually complete as static interface screenshots, yet they were associated with substantially less favorable user-centered outcomes than the other two conditions. This pattern suggests that surface-level visual plausibility should not be treated as equivalent to usability. A raw AI-generated interface may contain recognizable UI elements and visually polished components while still failing to support efficient information search, task-relevant attention, or accurate action selection.

This finding is consistent with human–AI interaction principles emphasizing that AI-supported systems should remain understandable, controllable, and aligned with user goals [6]. It is also consistent with recent design principles for generative AI applications, which argue that generated outputs require human oversight, contextual adaptation, and iterative refinement rather than direct uncritical adoption [2]. In the present study, raw AI-generated interfaces may have imposed additional user effort because relevant information was not always sufficiently prioritized, grouped, or visually emphasized. From the perspective of cognitive load theory, such problems can be interpreted as sources of extraneous cognitive load: effort generated not by the task itself, but by the way information is presented to the user [24,25].

Within the tested stimulus set, OptAI interfaces were associated with a lower usability burden than RawAI interfaces. The OptAI interfaces were generated through systematic prompt refinement targeting layout hierarchy, visual alignment, information clarity, spacing, and task salience. The observed differences in perceived cognitive load, perceived visual complexity, task ease, interface evaluation time, and task accuracy are consistent with the potential value of prompt-based HCI constraints as a usability-oriented design strategy. However, the present design does not isolate prompt wording as the sole causal source of these differences.

However, the results also show that prompt optimization did not make AI-generated interfaces indistinguishable from human-designed references. Although OptAI reached Human-level task accuracy, OptAI interfaces were still associated with longer interface evaluation times and were rated as more cognitively demanding and visually complex than Human interfaces. This distinction suggests that prompt-based HCI constraints may help narrow, but do not eliminate, the observed usability gap; human-centered refinement therefore remains necessary.

4.2. The Semantic Gap Between Pixel-Level Complexity and Human Interpretation

The second major finding concerns the relationship between computational visual metrics and human-centered outcomes. The interface-level Spearman correlations between visual metrics and user outcomes were weak to moderate, and none reached statistical significance. This result does not mean that computational metrics are uninformative. Rather, it indicates that low-level visual descriptors capture only part of the user experience associated with interface complexity.

Computational metrics such as Shannon entropy, edge density, color entropy, contrast, and white-space ratio describe the visual structure of an interface at the screenshot level. They quantify distributions of pixels, edges, luminance values, color variability, and light background regions. These features are useful for measuring reproducible properties of interface images and are consistent with prior work on visual clutter, interface complexity, and perceptual aesthetics [13–18]. However, human users do not experience an interface only as a collection of pixels or edges. They interpret labels, visual hierarchy, grouping cues, affordances, task relevance, and domain expectations. These higher-level interpretive processes may not be fully represented by screenshot-level metrics.

This divergence reflects a broader semantic gap between low-level computational features and high-level human interpretation [35]. A visually simple interface at the pixel level may still be difficult to use if the semantic organization is unclear. Conversely, an interface with relatively high edge density or visual information may remain usable if the layout supports predictable grouping, clear labeling, and task-relevant salience. In the present study, this helps explain why computational metrics showed only partial alignment with perceived cognitive load, perceived visual complexity, task ease, interface evaluation time, and task accuracy.

The interface-level results also provide an informative exception to an interpretation of usability based solely on task accuracy. Interface 11 (S3, RawAI) achieved a task accuracy of 1.000, yet it was also associated with high perceived cognitive load, high perceived visual complexity, low task ease, and a long interface evaluation time. This pattern suggests that users may occasionally identify a specific target or primary action correctly in a raw AI-generated interface, while still experiencing substantial processing burden due to poor surrounding structure, weak hierarchy, or inefficient visual organization. Therefore, task accuracy alone should not be interpreted as sufficient evidence of usability; it should be evaluated together with perceived cognitive load, perceived visual complexity, task ease, and interface evaluation time.

The results therefore suggest that visual complexity in AI-generated interfaces should be understood as a multi-layered construct. Pixel-level complexity describes one layer of the interface. Semantic organization, information hierarchy, and task pathway clarity describe another. The weak-to-moderate correlations observed in this study indicate that computational visual metrics can contribute to interface evaluation, but they should not be used as standalone proxies for perceived cognitive load or usability-related outcomes. For generative AI-designed interfaces, this distinction is especially important because AI systems may generate visually coherent screenshots that are semantically inefficient for users.

4.3. Theoretical and Practical Implications for Human-Centered Generative UI Design

Theoretically, this study contributes to the integration of visual complexity research, cognitive workload measurement, and human-centered AI evaluation. Prior work has examined visual complexity and interface aesthetics in human-designed systems [9–19], while other work has focused on human–AI interaction and GenAI-supported design workflows [1–8]. The present study connects these areas by showing that the evaluation of generative AI-designed interfaces requires both computational and human-centered evidence. Visual metrics provide reproducible descriptors of interface structure, whereas user testing reveals how those structures are interpreted during task performance.

This finding extends human-centered AI design principles by emphasizing that generative UI systems should not be evaluated only by the apparent visual quality of their outputs. A generated interface can look plausible while still being associated with greater perceived cognitive load or lower task accuracy. Therefore, the evaluation of AI-generated UI designs should include task-based measures of effectiveness, efficiency, and perceived effort, in addition to computational measures of visual structure. This aligns with usability measurement frameworks that treat interface quality as multidimensional rather than reducible to a single aesthetic or perceptual score [27].

Practically, the results provide several implications for designers and developers. For UI/UX practitioners, raw generative AI outputs should be treated as early design drafts rather than ready-to-use interfaces. Prompt instructions should explicitly include requirements related to information hierarchy, grouping, spacing, task salience, readability, affordance clarity, and user goal alignment. The observed differences between RawAI and OptAI are consistent with the possibility that such prompt-level constraints may contribute to more favorable user outcomes.

For developers of generative UI tools, the findings suggest that future systems may benefit from embedding human-centered design constraints into generation and post-generation evaluation pipelines. Automated metrics may be useful as an initial screening layer, but they should be combined with semantic checks, task-oriented evaluation, accessibility criteria, and human feedback. Generative UI systems that optimize only visual style may produce attractive but inefficient interfaces. In contrast, systems that incorporate usability constraints and human-centered evaluation loops may better support reliable AI-assisted interface design.

From a practical workflow perspective, the proposed framework can be integrated into AI-assisted UI design in three stages. First, designers can generate multiple candidate interfaces under both unconstrained and prompt-optimized conditions, explicitly specifying task hierarchy, information grouping, spacing, contrast, readability, and action salience in the optimized prompts. Second, candidate outputs can be screened using computational indicators such as entropy, edge density, contrast, and white-space ratio to identify potentially overloaded or poorly structured layouts before user testing. Third, shortlisted designs should be evaluated through brief task-based user tests measuring perceived cognitive

load, perceived visual complexity, task ease, interface evaluation time, and task accuracy. This workflow enables design teams to use computational metrics as an early diagnostic layer while preserving human-centered testing as the final decision criterion.

For product teams and practitioners, the findings suggest that prompt-based HCI constraints may serve as a potentially useful usability-oriented design strategy rather than merely a content-generation technique. Prompts that explicitly encode HCI principles may reduce visual clutter, improve task salience, and support more efficient information search. However, computational scores should not be interpreted as automatic indicators of usability. Instead, they should be used to flag interfaces for further inspection, guide prompt revision, and prioritize design alternatives for human evaluation.

4.4. Limitations and Future Research

This study has several limitations. First, the evaluation was conducted using static interface screenshots rather than fully interactive prototypes. Static screenshots allow controlled comparison of visual structure, but they cannot capture interaction dynamics such as scrolling, clicking, hover states, navigation flow, feedback, or error recovery. Future studies should examine whether the same patterns hold in interactive prototypes or functional applications.

Second, the stimuli were presented using a fixed mixed order. This structure interleaved scenarios and design conditions to reduce condition clustering, but it did not provide full counterbalancing or Latin-square randomization. Future research should use counterbalanced presentation orders to further control for learning, fatigue, and order effects. The very high concordance observed in several repeated-measures outcomes may reflect the strong separation among the selected experimental conditions. Future studies should test larger and more heterogeneous stimulus sets to examine whether the same ordinal pattern holds across more diverse AI-generated interface outputs.

Third, the study focused on four applied interface scenarios: a mobile health dashboard, a learning management system dashboard, an e-commerce shopping cart, and a university student information portal. These scenarios represent common information-seeking and decision-oriented interface contexts, but they do not cover all possible UI domains. Future work should include additional interface types, such as productivity tools, financial dashboards, public service platforms, accessibility-focused interfaces, and conversational or multimodal interfaces.

Fourth, although the semantic task categories, number of questions, and response formats were held constant within each scenario, the exact values, labels, and surrounding content differed across the three design conditions. Consequently, the tasks were functionally equivalent rather than item-identical. These stimulus-specific differences may have contributed to variation in task accuracy or interface evaluation time independently of visual organization. The observed condition differences should therefore be interpreted as reflecting the complete interface stimuli, including their visual and semantic properties, rather than visual complexity or prompt condition alone. Future studies should use multiple stimulus replications with identical target values and labels to isolate the contribution of visual organization more precisely.

Fifth, each AI condition was represented by one selected output per scenario generated using a single image-generation system. The study therefore does not isolate prompt wording as the sole causal source of the observed differences. Stochastic model variation, output-selection decisions, stimulus-specific semantic content, and differences in target salience may also have contributed. The findings should consequently be interpreted as evidence from the tested set of stimuli rather than as a universal estimate of prompt-optimization effectiveness. Future research should generate multiple independent outputs

per prompt condition, use controlled seeds where available, and replicate the comparison across different generative models.

Sixth, the participant sample consisted of digitally experienced users. This increases the relevance of the findings for users familiar with digital interfaces, but it may limit generalizability to populations with lower digital literacy, older users, or users with accessibility needs. Future studies should test more diverse samples and examine whether AI-generated interface complexity affects different user groups in distinct ways.

Seventh, a related limitation concerns participants' prior familiarity with UI/design practices and AI tools. Although the sample was digitally experienced, participants differed in their background exposure to interface design and generative AI tools. Such differences may influence how participants visually inspect interface layouts, interpret design quality, or tolerate AI-generated inconsistencies. Participants with prior UI/design experience may be more sensitive to hierarchy, spacing, alignment, and affordance problems, whereas participants familiar with AI tools may evaluate AI-generated outputs with different expectations. Although these variables were collected to characterize the sample, the present study was not primarily designed to model expertise-based moderation effects. Future studies should recruit larger and more stratified samples to examine whether UI/design expertise and AI familiarity moderate the relationship between computational visual complexity and user-centered outcomes.

Eighth, the correlation analyses were conducted at the interface level using twelve stimuli. This was appropriate for linking screenshot-level computational metrics with aggregated user-centered outcomes, but it limits statistical power for detecting associations. Future research should include larger stimulus sets to obtain more stable estimates of the relationship between computational visual metrics and human-centered outcomes.

Ninth, the groupTime variable captured total page-level interface evaluation duration rather than isolated task execution time. It may therefore include image-loading latency, reading and rating time, and brief participant pauses. Although the within-subject design and scenario-specific resolution standardization reduced systematic device-related differences across conditions, the measure should not be interpreted as a pure reaction-time or task-completion measure. Future studies should use experimental platforms that separately record stimulus-onset time, first-response latency, task-response time, and post-task rating time.

Tenth, perceived cognitive load was assessed using a single-item subjective rating, whereas task accuracy and interface evaluation time were treated as complementary behavioral outcomes. Although single-item ratings are practical in repeated interface evaluations, they do not provide the multidimensional coverage offered by validated workload instruments. Future studies could include multi-item workload instruments, eye-tracking, pupillometry, EEG, fNIRS, or other neuroergonomic measures to obtain more direct and comprehensive evidence of attentional allocation and cognitive workload [23,26]. Such multimodal evaluation would help clarify how low-level visual structure, semantic organization, and task demands jointly shape user experience in generative AI-designed interfaces.

Overall, the findings indicate that, within the tested stimulus set, prompt-optimized interfaces were associated with more favorable usability-related outcomes than raw AI-generated interfaces. However, robust evaluation requires more than visual inspection or computational metrics alone. A hybrid framework combining image-based complexity analysis with human-centered usability testing offers a more reliable pathway for assessing generative AI-designed user interfaces.

5. Conclusions

This study proposed and evaluated a hybrid framework for assessing generative AI-designed user interfaces by combining computational visual complexity metrics with human-centered usability testing. Rather than treating AI-generated interface images as purely visual artifacts, the study examined them as task-oriented interaction surfaces that must support perception, interpretation, decision-making, and accurate user action.

Within the tested stimulus set, RawAI interfaces were associated with the highest perceived cognitive load, highest perceived visual complexity, longest interface evaluation times, and lowest task accuracy. Prompt-optimized AI interfaces showed more favorable user-centered outcomes than RawAI interfaces and reached 100% task accuracy, matching the human-designed reference interfaces on this measure. However, they remained distinct from human-designed references in perceived cognitive load, perceived visual complexity, task ease, and interface evaluation time. These findings suggest that prompt-based HCI constraints may help narrow, but do not eliminate, the observed usability gap. They should not be interpreted as establishing prompt wording as the sole causal factor or as a substitute for human-centered design judgment.

A second major conclusion is that pixel-level computational metrics alone are insufficient to explain the full cognitive and usability burden of AI-generated interfaces. The observed weak-to-moderate, non-significant associations between visual metrics and user outcomes suggest that screenshot-level properties such as entropy, edge density, contrast, color distribution, and white-space ratio capture only part of interface complexity. Human users process not only pixels and visual density, but also semantic labels, information hierarchy, grouping, affordances, and task relevance. Therefore, the quality of AI-generated interfaces should be evaluated through both computational descriptors and empirical user-centered outcomes.

Generative AI can readily produce interface-like images; the more important question is whether these generated interfaces are cognitively efficient, semantically coherent, and task-supportive for human users. Future generative UI systems should therefore move beyond visual plausibility and incorporate constraints related to cognitive load, information architecture, task salience, accessibility, and usability performance.

In applied UI/UX practice, the framework may be used as a staged evaluation procedure: generate candidate AI interfaces, screen them with reproducible computational visual metrics, refine prompts according to human-centered design principles, and validate the shortlisted outputs through task-based user testing. This approach can help designers and product teams move from visually plausible AI-generated screenshots toward interfaces that are better aligned with human-centered usability requirements, support task completion, and are suitable for human–AI co-creative workflows.

In conclusion, robust evaluation of generative AI-designed user interfaces requires an integrated approach. Computational visual analysis provides reproducible measurements of interface structure, while human-centered testing reveals how users experience and perform with those interfaces. By bringing these two perspectives together, this study demonstrates the complementary roles of computational screening and human-centered validation in the evaluation and improvement of AI-generated interfaces.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/electronics15153458/s1>. Figure S1: Standardized computational visual metric profiles by interface condition; Figures S2–S5: Complete Human, RawAI, and OptAI stimulus sets for Scenarios S1–S4; Table S1: Scenario characteristics, stimulus-generation conditions, and task targets for the twelve experimental interfaces; Table S2: Two-sided *p* values for interface-level Spearman rank-order correlations between computational visual metrics and aggregated user-centered outcomes; Table S3: Stimulus-specific task answer key and binary accuracy-

coding criteria; Table S4: RawAI and OptAI prompt specifications and standardized stimulus-generation details.

Author Contributions: Conceptualization, N.V. and Ç.G.; methodology, N.V.; software, N.V.; validation, N.V. and Ç.G.; formal analysis, N.V.; investigation, N.V. and Ç.G.; resources, N.V. and Ç.G.; data curation, N.V.; writing—original draft preparation, N.V.; writing—review and editing, N.V. and Ç.G.; visualization, N.V.; supervision, N.V.; project administration, N.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Human Research Ethics Committee of KTO Karatay University, Konya, Türkiye, on 30 April 2026 (Meeting No. 04, Decision No. 34).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. Raw participant-level survey data are not publicly available due to privacy and ethical considerations related to human-subject responses. De-identified and aggregated datasets, computational visual metric outputs, statistical analysis tables, and figure outputs can be made available upon reasonable request.

Acknowledgments: The authors would like to thank all participants who voluntarily took part in the user evaluation study. The authors also acknowledge KTO Karatay University for institutional and academic support. The authors thank the developers and maintainers of the open-source software libraries used for image processing, statistical analysis, and visualization in this study. During the preparation of the study materials, the authors used ChatGPT Images 2.0 (OpenAI), accessed through the ChatGPT web interface in May 2026, to generate the interface stimuli used in the RawAI and OptAI conditions. A more specific backend build identifier was not exposed through the user interface. Grammarly was also used for English language editing and grammar checking. The authors have reviewed and approved the final version of the manuscript and take full responsibility for its content.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
GenAI	Generative Artificial Intelligence
UI	User Interface
UX	User Experience
HCI	Human–Computer Interaction
RawAI	Raw AI-generated interface condition
OptAI	Prompt-optimized AI-generated interface condition
RQ	Research Question
RGB	Red, Green, Blue
HSV	Hue, Saturation, Value
RMS	Root Mean Square
PNG	Portable Network Graphics
SD	Standard Deviation
CI	Confidence Interval
EEG	Electroencephalography
fNIRS	Functional Near-Infrared Spectroscopy

References

- Li, J.; Cao, H.; Lin, L.; Hou, Y.; Zhu, R.; El Ali, A. User Experience Design Professionals' Perceptions of Generative Artificial Intelligence. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024*; Association for Computing Machinery: New York, NY, USA, 2024; Article 381, pp. 1–18. [\[CrossRef\]](#)
- Weisz, J.D.; He, J.; Muller, M.; Hoefer, G.; Miles, R.; Geyer, W. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024*; Association for Computing Machinery: New York, NY, USA, 2024; Article 378, pp. 1–22. [\[CrossRef\]](#)
- Takaffoli, M.; Li, S.; Mäkelä, V. Generative AI in User Experience Design and Research: How Do UX Practitioners, Teams, and Companies Use GenAI in Industry? In *Proceedings of the 2024 ACM Designing Interactive Systems Conference, Copenhagen, Denmark, 1–5 July 2024*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1579–1593. [\[CrossRef\]](#)
- Muller, M.; Kantosalo, A.; Maher, M.L.; Martin, C.P.; Walsh, G. GenAICHI 2024: Generative AI and HCI at CHI 2024. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024*; Association for Computing Machinery: New York, NY, USA, 2024; Article 470, pp. 1–7. [\[CrossRef\]](#)
- Lin, D.C.-E.; Kang, H.B.; Martelaro, N.; Kittur, A.; Chen, Y.-Y.; Hong, M.K. Inkspire: Supporting Design Exploration with Generative AI through Analogical Sketching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025*; Association for Computing Machinery: New York, NY, USA, 2025; Article 427, pp. 1–18. [\[CrossRef\]](#)
- Amershi, S.; Weld, D.; Vorvoreanu, M.; Fournery, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K.; et al. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 4–9 May 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–13. [\[CrossRef\]](#)
- Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–15. [\[CrossRef\]](#)
- Shneiderman, B. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. *ACM Trans. Interact. Intell. Syst.* **2020**, *10*, 26. [\[CrossRef\]](#)
- Zhang, R.; Son, O.-S. A Measurement Model for Visual Complexity in HCI: Focusing on Visual Elements in Mobile GUI Design. *Electronics* **2025**, *14*, 942. [\[CrossRef\]](#)
- Miniukovich, A.; De Angeli, A. Quantification of Interface Visual Complexity. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, Bolzano, Italy, 27–30 May 2014*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 153–160. [\[CrossRef\]](#)
- Miniukovich, A.; Sulpizio, S.; De Angeli, A. Visual Complexity of Graphical User Interfaces. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces, Castiglione della Pescaia, Italy, 29 May–1 June 2018*; Association for Computing Machinery: New York, NY, USA, 2018; Article 20, pp. 1–9. [\[CrossRef\]](#)
- Riegler, A.; Holzmann, C. Measuring Visual User Interface Complexity of Mobile Applications with Metrics. *Interact. Comput.* **2018**, *30*, 207–223. [\[CrossRef\]](#)
- Rosenholtz, R.; Li, Y.; Mansfield, J.; Jin, Z. Feature Congestion: A Measure of Display Clutter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, OR, USA, 2–7 April 2005*; Association for Computing Machinery: New York, NY, USA, 2005; pp. 761–770. [\[CrossRef\]](#)
- Rosenholtz, R.; Li, Y.; Nakano, L. Measuring Visual Clutter. *J. Vis.* **2007**, *7*, 17. [\[CrossRef\]](#) [\[PubMed\]](#)
- Michailidou, E.; Harper, S.; Bechhofer, S. Visual Complexity and Aesthetic Perception of Web Pages. In *Proceedings of the 26th ACM International Conference on Design of Communication, Lisbon, Portugal, 22–24 September 2008*; Association for Computing Machinery: New York, NY, USA, 2008; pp. 215–224. [\[CrossRef\]](#)
- Harper, S.; Michailidou, E.; Stevens, R. Toward a Definition of Visual Complexity as an Implicit Measure of Cognitive Load. *ACM Trans. Appl. Percept.* **2009**, *6*, 10. [\[CrossRef\]](#)
- Tuch, A.N.; Presslauer, E.E.; Stöcklin, M.; Opwis, K.; Bargas-Avila, J.A. The Role of Visual Complexity and Prototypicality Regarding First Impression of Websites: Working towards Understanding Aesthetic Judgments. *Int. J. Hum.-Comput. Stud.* **2012**, *70*, 794–811. [\[CrossRef\]](#)
- Reinecke, K.; Yeh, T.; Miratrix, L.; Mardiko, R.; Zhao, Y.; Liu, J.; Gajos, K.Z. Predicting Users' First Impressions of Website Aesthetics with a Quantification of Perceived Visual Complexity and Colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013*; Association for Computing Machinery: New York, NY, USA, 2013; pp. 2049–2058. [\[CrossRef\]](#)
- Lavie, T.; Tractinsky, N. Assessing Dimensions of Perceived Visual Aesthetics of Web Sites. *Int. J. Hum.-Comput. Stud.* **2004**, *60*, 269–298. [\[CrossRef\]](#)

20. Leiva, L.A.; Xue, Y.; Bansal, A.; Tavakoli, H.R.; Köroğlu, T.; Dayama, N.R.; Oulasvirta, A. Understanding Visual Saliency in Mobile User Interfaces. In *Proceedings of the 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, Oldenburg, Germany, 5–8 October 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–12. [\[CrossRef\]](#)
21. Jiang, Y.; Leiva, L.A.; Tavakoli, H.R.; Houssel, P.R.B.; Kylmälä, J.; Oulasvirta, A. UEyes: Understanding Visual Saliency across User Interface Types. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–21. [\[CrossRef\]](#)
22. Ngo, D.C.L.; Teo, L.S.; Byrne, J.G. Modelling Interface Aesthetics. *Inf. Sci.* **2003**, *152*, 25–46. [\[CrossRef\]](#)
23. Kosch, T.; Karolus, J.; Zagermann, J.; Reiterer, H.; Schmidt, A.; Woźniak, P.W. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* **2023**, *55*, 283. [\[CrossRef\]](#)
24. Sweller, J. Cognitive Load during Problem Solving: Effects on Learning. *Cogn. Sci.* **1988**, *12*, 257–285. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Paas, F.; Tuovinen, J.E.; Tabbers, H.; Van Gerven, P.W.M. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educ. Psychol.* **2003**, *38*, 63–71. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Hart, S.G. NASA-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, San Francisco, CA, USA, 16–20 October 2006*; Sage: Thousand Oaks, CA, USA, 2006; Volume 50, pp. 904–908. [\[CrossRef\]](#)
27. Hornbæk, K. Current Practice in Measuring Usability: Challenges to Usability Studies and Research. *Int. J. Hum.-Comput. Stud.* **2006**, *64*, 79–102. [\[CrossRef\]](#)
28. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [\[CrossRef\]](#)
29. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, PAMI-8, 679–698. [\[CrossRef\]](#)
30. Sullivan, G.M.; Artino, A.R., Jr. Analyzing and Interpreting Data from Likert-Type Scales. *J. Grad. Med. Educ.* **2013**, *5*, 541–542. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [\[CrossRef\]](#)
32. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [\[CrossRef\]](#)
33. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
34. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Smeulders, A.W.M.; Worring, M.; Santini, S.; Gupta, A.; Jain, R. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1349–1380. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.