International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Measuring the Perceived Clutter of Websites

James R. Lewis & Jeff Sauro

To cite this article: James R. Lewis & Jeff Sauro (03 Jun 2024): Measuring the Perceived Clutter of Websites, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2024.2359205

To link to this article: https://doi.org/10.1080/10447318.2024.2359205

Published online: 03 Jun 2024.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2024.2359205

Measuring the Perceived Clutter of Websites

James R. Lewis[a] and Jeff Sauro[b]

[a]MeasuringU, Delray Beach, FL, USA; [b]MeasuringU, Denver, CO, USA

ABSTRACT

Perceived clutter is a potentially important but understudied construct in UX research. In this paper we described the development and assessment of a standardized questionnaire for reliable and valid measurement of perceived clutter of websites. Starting with an initial set of 16 items and two hypothesized factors, a series of exploratory analyses led to a final set of five items, two for the hypothesized construct of Content Clutter (too much irrelevant content like ads and videos) and three for the hypothesized construct of Design Clutter (poor design of relevant information like too much text, an unpleasant layout, or too much visual noise). Confirmatory analyses using an independent dataset showed excellent fit statistics for CFA of the five-item questionnaire and good fit for an SEM of the connections between clutter and other UX constructs. Researchers should exercise caution about generalizing results to other contexts and interfaces, but UX practitioners should be able to use this perceived clutter of websites (PCW) questionnaire when assessing consumer websites.

KEYWORDS

Perceived clutter; standardized usability questionnaires; SUPR-Q; UX-Lite; likelihood-to-recommend; perceived usefulness; perceived ease of use; user experience; UX

1. Introduction

In our user experience (UX) research practice, we have frequently encountered users and designers criticizing website interfaces for being cluttered and stakeholders who worry about the experiential and business consequences of a cluttered website. But what exactly does it mean for a website to appear cluttered?

1.1. The construct of clutter

Dictionary definitions of clutter tend to equate it with messiness or untidiness. As a transitive verb, the Merriam-Webster (n.d.) definition is "to fill or cover with scattered or disordered things that impede movement or reduce effectiveness" and, as a noun, "a crowded or confused mass or collection." The Oxford Dictionary (n.d.) verb and noun definitions are, respectively, "to crowd (a place or space) with a disorderly assemblage of things" and "a crowded and confused assemblage."

These definitions do not address two potential components of clutter. One component is the extent to which the disorganized objects are needed but should be better arranged (e.g., tools in a toolbox). The other is the extent to which some objects are unnecessary and should be discarded (e.g., old candy bar wrappers in a toolbox). This distinction is sometimes brought out in definitions of the word, "declutter" (e.g., "to declutter is to tidy up a mess, especially by getting rid of objects," vocabulary.com, n.d.).

Even in this everyday sense, these two components of clutter suggest different decluttering strategies – (1) reorganize needed objects and (2) discard unnecessary objects.

1.2. Clutter in user interface design

There is a long history of defining and measuring clutter in user interface design, especially for mission-critical applications (e.g., aircraft cockpit displays), drawn from research in disciplines like human factors engineering and perceptual psychology. In most cases this research has focused on objective rather than subjective measurement of clutter.

For example, Tullis (1983) published a human factors review and analysis of the formatting of alphanumeric displays – the types of monochromatic character displays widely used in the 1970s and 1980s. Tullis identified four basic format characteristics: overall density (number of characters displayed divided by total character spaces available), local density (number of filled character spaces near each character), grouping (extent to which items formed well-defined perceptual groups), and layout complexity (extent to which the arrangement of items followed a predictable visual scheme). He explored different ways to objectively measure these characteristics that, along with the reviewed literature, supported several key design recommendations, for example:

- Keep overall density as low as possible while still displaying task-relevant data.
- Use white space to reduce local density.
- Grouping related items is beneficial to performance.

CONTACT James R. Lewis jim@measuringu.com; drjimnfl@gmail.com MeasuringU, Delray Beach, FL, USA

- Layouts are less complex when data are presented in tables rather than narratives.

Using proposed objective measurements of clutter from perceptual psychology, Rosenholtz et al. (2007) evaluated feature congestion (the difficulty of adding to a display a new item that can draw attention), subband entropy (based on clutter being related to visual information in a display), and edge density (the percentage of pixels on a display that are edge pixels). They found these three measures correlated with different empirical measures of search performance (e.g., searching for objects in cluttered maps or on cluttered desks). They also reported that color variability (number of colors and how different they are) affected visual clutter. Design recommendations consistent with this research include:

- To make an object salient, use design features like contrast, color, orientation, and motion.
- Group similar objects together using features like hue, luminance, and size.
- Some use of color can improve search performance but avoid excessive color variability.

Kaber et al. (2008), in the context of advanced cockpit displays, developed a subjective clutter questionnaire. Their participants were four expert test pilots with experience using advanced heads-up displays (HUDs) who rated the clutter of images of a flight approach scenario depicting multiple display conditions. The initial version of the clutter questionnaire contained 14 semantic differential items gleaned from a literature review of display clutter (e.g., sparse/dense, monochromatic/colorful, empty/crowded, ungrouped/grouped). After each trial in the experiment, participants provided a single rating of overall clutter (20-point scale from "low clutter" to "high clutter") and rated each of the 14 semantic differentials regarding their utility for describing clutter (20-point scale from "low" to "high"). Using the eigenvalues-greater-than-one criterion (Cliff, 1988), a Varimax-rotated principal components analysis found the 14 items aligned with four components:

- Global density: not salient/salient, sparse/dense, empty/crowded, low workload/high workload, low attention/high attention
- Feature similarity: redundant/orthogonal, similar/dissimilar
- Feature clarity: unsafe, safe, dull/sharp, indiscernible/discernible, monochromatic/colorful
- Dynamic nature: static/dynamic, ungrouped/grouped, monotonous/variable

In a review of definitions and measurement of display clutter, Moacdieh and Sarter (2015, p. 61) wrote, "Despite the widespread agreement on the harmful nature of 'clutter,' researchers have yet to reach consensus on a definition and a reliable way of manipulating and measuring the phenomenon." Their primary goal was to investigate the literature of the effects of clutter on visual search performance for definitions and metrics. Common definitions include display density (number of entities on a screen), display layout (arrangement, nature, and color of entities), target background/distractor similarity, task irrelevance (both essential and nonessential entities are displayed), and performance/attentional costs. Approaches to measurement include image processing, performance evaluation, eye tracking, and subjective evaluation (perceived clutter).

In the Moacdieh and Sarter (2015) review, the most commonly reported subjective evaluation of clutter was a measure of overall clutter captured with a single rating scale. A notable exception was the standardized questionnaire developed by Kaber et al. (2008), which has been used extensively in research on clutter in aircraft displays (e.g., Kaber et al., 2013). Despite the clear value of the Kaber questionnaire in its intended context (professional pilots familiar with aircraft displays and associated technical terminology such as redundant/orthogonal), it does not seem to be well suited to the context of assessing the perceived clutter of websites.

1.3. Perceived clutter in the specific context of website design

The term "clutter" seems to be part of the website design vernacular, evident in online articles by UX practitioners discussing the topic of decluttering websites (not peer reviewed). For example, Crowley (2017) listed three characteristics believed to lead to perceived clutter: too much content on the screen, content not logically organized, and too much visual noise due to imagery and contrasts. Hughes (2022) recommended that website designers carry out some spring cleaning, have a clear linking strategy, improve content and site readability, and use more white space. Saxena (2021) advised against having too much text and too many options. Even though typical user goals and behaviors with websites (e.g., browsing for information, making online purchases) differ from those of pilots using displays to land aircraft, many of these website design recommendations are consistent with the design guidance implied by clutter research in other domains.

A search of the peer-reviewed literature specifically targeting standardized questionnaires for the assessment of perceived website clutter did not return any relevant results. We did, however, find research in the fields of marketing and advertising that are relevant regarding the extent to which online ads contribute to the perception of clutter on websites, a continuation of lines of research originally conducted on magazines and television (Speck & Elliott, 1997) in which a primary objective metric is the proportion of advertisements in the total space of a medium (Kim & Sundar, 2010).

Using a standardized questionnaire they developed for assessing consumer reaction to online ads (specifically, the constructs of perceived intrusiveness, irritation, informativeness, and entertainment value), Edwards et al. (2002) reported that ads perceived as intrusive elicited irritation and ad avoidance. Interruptive ads that occur during an

online shopping task have been found to increase primary task time with early interruptions more disrupting than late interruptions (Xia & Sudharshan, 2002).

Forced presentation of ads irritates users especially when ads are not skippable but, when ad clutter is high, skippability doesn't reduce irritation (Senarathna & Wijetunga, 2023). Experimental manipulation of ad location and relevance found that both factors affect perception of ad clutter (Kim & Sundar, 2010). Brinson et al. (2018) investigated why consumers install ad blockers, noting that "To discourage the use of ad blockers, publishers and ad industry leaders have been experimenting with a variety of methods to improve users' experiences—from decluttering websites to developing less intrusive ad formats" (p. 138). Brinson et al. found concerns about information privacy influence attitudes toward personalized advertising when messages are hypertargeted based on too many layers of personal data – ads often described as "creepy."

Based on this research, web design guidelines relevant to advertisement include:

- Reduce perceived ad intrusiveness by increasing ad relevance and value (interesting and entertaining).
- Avoid haphazard presentation of ads with regard to location and relevance.
- Time the presentation of online ads to avoid disrupting users' primary tasks.
- To reduce disruption when placing ads on a website, take into consideration where users probably are in their primary task (preferably at the end of the task).
- Use technologies to make ads relevant to users but avoid overly direct messages based on too much personal data, especially from third parties.

In short, website designers face numerous challenges with regard to the management of perceived clutter. An effective ad strategy is critical for many websites, and failing to strike an appropriate balance between corporate and user needs can lead to negative impressions of the website and its parent enterprise. Website designers must also deal with more traditional design elements associated with perceived clutter, such as density, white space, logical grouping, layout complexity, and color.

1.4. The SUPR-Q measurement framework

The impact of perceived clutter on website users and stakeholders (poor design leading to unhappy users leading to poor business consequences) is very different from the consequences of objective clutter in aircraft displays (poor design increasing the likelihood of crashes). Ad clutter, a potentially major driver of perceived clutter for websites, is a nonissue in mission critical displays (e.g., fighter pilots don't subscribe to the free ad-supported version of their HUDs). The unique aspects of website use are reflected in the development of standardized UX questionnaires for the assessment of website quality (Sauro & Lewis, 2016).

In our practice, we use the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q, n.d.) in periodic assessments of key websites in major business sectors (e.g., Apartments.com, Realtor.com, Redfin, Trulia, and Zillow in the real estate sector). Shown in Figure 1, the SUPR-Q measures four website UX factors with eight questions: Usability (easy to use, easy to navigate), Trust (trustworthy, credible), Appearance (attractive, clean and simple), and Loyalty (likelihood to revisit, likelihood to recommend). For details on SUPR-Q development and scoring, see Sauro (2015).

In addition to its usefulness as a single measure of the UX of websites, the components of the SUPR-Q can be used in a framework in which Usability, Trust, and Appearance are antecedents of Loyalty. That framework can be extended to include additional consequences like Brand Attitude (e.g., "How would you describe your attitude toward this company?" anchored with "1: Very Unfavorable" and "7: Very favorable") and additional antecedents (e.g., perceived clutter).

1.5. Research goals

Our primary research goals were to:

- Create an initial clutter questionnaire capable of measuring two hypothesized components of perceived website clutter: content clutter and design clutter.
- Streamline the initial clutter questionnaire by identifying the best items to retain.
- Assess the predictive value of the clutter questionnaire in the SUPR-Q measurement framework.

2. Methods

This section describes the initial item set for the clutter questionnaire, the participants, and the data collection procedure.

2.1. Initial set of clutter items

Consistent with the literature review, we hypothesized that there are at least two factors that might contribute to the perceived clutter of websites: content clutter and design clutter.

We expected content clutter to be driven by the presentation of nonrelevant ads and videos occupying a considerable percentage of display space and having negative emotional consequences (e.g., annoying). Considering the components of the everyday conception of clutter, these would be the candy wrappers in the toolbox – items that website users would prefer to discard, perhaps with ad blockers.

Our conception of design clutter is that it is driven by issues with the presentation of potentially relevant content that make it difficult to consume (e.g., insufficient white space, too much text, illogical layout). Analogous with the everyday definition of clutter, this is the content that should be retained but needs reorganization.

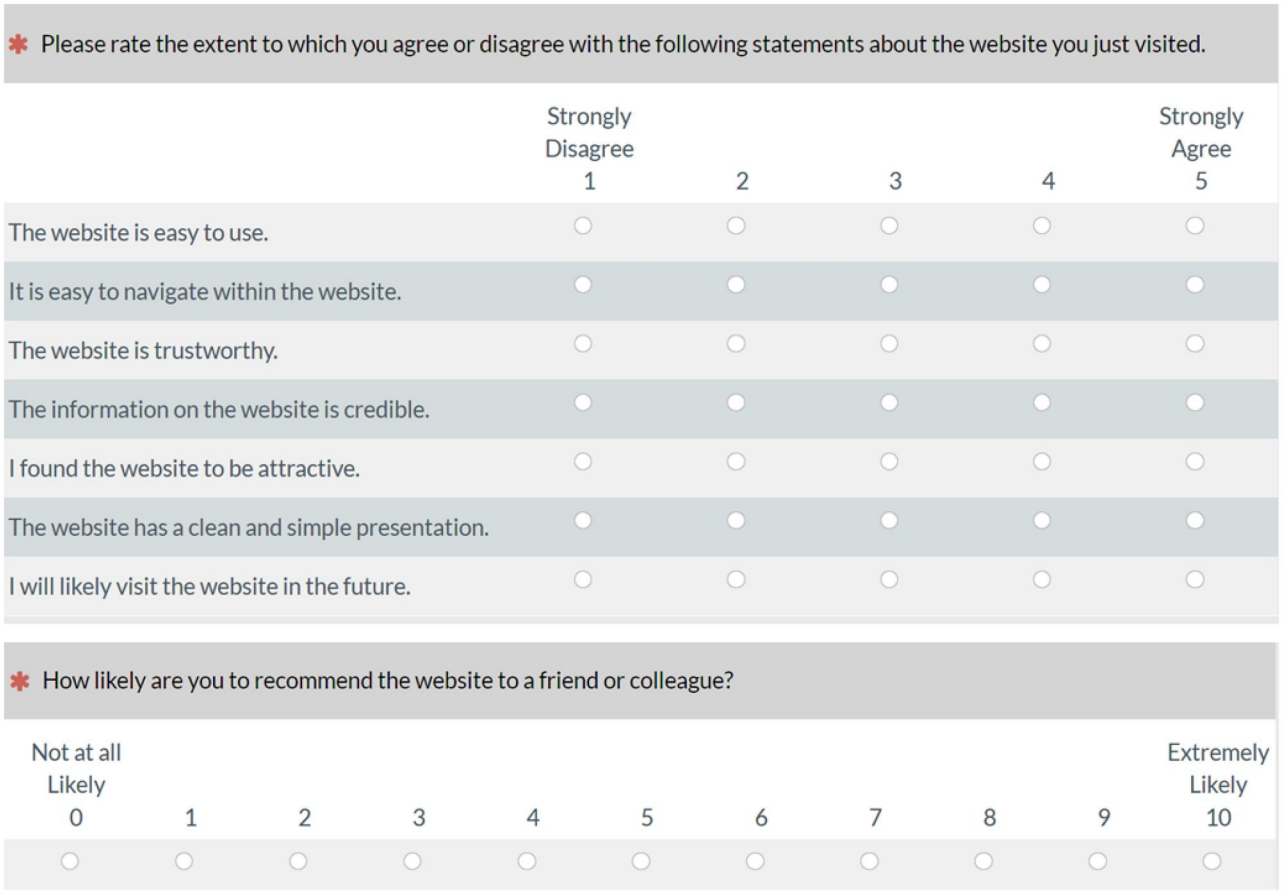


Figure 1. The SUPR-Q questionnaire.

The first iteration of the perceived website clutter (PWC) questionnaire included one item for overall clutter, six for content clutter, and ten for design clutter (for a screenshot of the entire questionnaire as used in our surveys, see Appendix Figure A1). The format for overall clutter was an 11-point agreement item ("Overall, I thought the website was too cluttered," 0: Strongly disagree, 10: Strongly agree). The format for content and design clutter were five-point agreement items (1: Strongly disagree, 5: Strongly agree). The short labels and item wording for the content and design clutter items were:

- **Content_ALot**: These types of content made up a lot of the clutter.
- **Content_TooMany**: There were too many ads or videos.
- **Content_Space**: These types of content took up too much space.
- **Content_Distracting**: These types of content were distracting.
- **Content_Irrelevant**: These types of content were irrelevant.
- **Content_Annoying**: These types of content were annoying.
- **Design_HardToRead**: The text was hard to read.
- **Design_SmallFont**: The font size was too small.
- **Design_DistractingColors**: The colors were distracting.
- **Design_UnpleasantLayout**: The layout was unpleasant.
- **Design_WhiteSpace**: There wasn't enough white space.
- **Design_TooMuchText**: There was too much text.
- **Design_NotLogical**: The content was not logically organized.
- **Design_Disorganized**: The layout was disorganized.
- **Design_VisualNoise**: There was too much visual noise.
- **Design_HardToStart**: It was hard for me to find what I needed to get started.

2.2. Participants

The participants were members of an online consumer panel, all from the United States who participated in retrospective consumer surveys conducted from 2022 through 2023 (57 websites in eight sectors). Table 1 shows the sectors, sample sizes, dates, gender distributions, and age distributions for the surveys. The total sample size was 2,761 with roughly even distributions of females and males and ages under and over 35 years old.

Respondents volunteered to participate in this research and were paid for participation by the online consumer panel. Participants cannot be identified from their survey responses and, as is typical in consumer surveys, there was no risk associated with participation. We complied with the ethical standards of the Human Factors and Ergonomics Society (HFES, 2020) and the User Experience Professionals Association (UXPA, n.d.).

Table 1. Summary of participant gender and age for eight consumer surveys.

Sector	*n*	Date	Websites	Female (%)	Male (%)	Under 35 (%)	35 or older (%)
Real Estate	269	April 2022	5	48	51	48	52
Travel Aggregator	452	April 2022	9	48	51	48	52
Business Info	183	July 2022	3	46	53	42	58
Domestic Air	350	May 2022	7	48	49	58	42
International Air	200	May 2022	5	53	46	61	39
Ticketing	234	June 2022	5	45	52	40	60
Clothing	550	December 2022	13	52	45	48	52
Wireless	523	January 2023	10	47	50	40	60
Overall	2,761	–	57	49	49	48	52

2.3. Data collection procedure

The eight surveys shown in Table 1 were retrospective studies of the UX of websites in their respective sectors. Some content of the surveys differed according to the nature of the sector being investigated, but all surveys included the SUPR-Q, a brand attitude item, the clutter questionnaire, and basic demographic items. For each survey we conducted screeners to identify respondents who had used one or more of the target websites within the past year, then invited those respondents to rate one website with which they had prior experience. On average, respondents completed the surveys in 10–15 min (there was no time limit). The websites included in each survey are listed below.

- **Real Estate** (Sauro et al., 2022, June 14): Apartments.com, Realtor.com, Redfin, Trulia, Zillow
- **Travel Aggregator** (Sauro et al., 2022, August 30): Booking.com, Expedia, Google Travel, Kayak, Orbitz, Priceline, Travelocity, Tripadvisor, Trivago
- **Business Info** (Sauro et al., 2022, September 27): Google Reviews, Tripadvisor, Yelp
- **Domestic Air** (Sauro et al., 2022, July 26): Alaska Airlines, American Airlines, Delta, Frontier, JetBlue, Southwest, United
- **International Air** (Sauro et al., 2022, July 26): Air Canada, Air France, British Airways, Lufthansa, Ryanair
- **Ticketing** (Sauro et al., 2022, October 17): AXS, SeatGeek, StubHub, Ticketmaster, Vivid Seats
- **Clothing** (Sauro et al., 2023, January 24): Anthropologie, Athleta, Banana Republic, Gap, H&M, Kohl's, Lululemon, Macy's Neiman Marcus, Nordstrom, Old Navy, Urban Outfitters, Zara
- **Wireless** (Schiavone et al., 2023, February 21): AT&T, Boost Mobile, Cricket Wireless, Google Fi, Mint Mobile, Spectrum Mobile, Straight Talk Wireless, T-Mobile, Verizon, Xfinity Mobile

3. Results

Unless otherwise specified, statistical analyses used SPSS Version 23 (including AMOS Version 23 for confirmatory factor analysis and structural equation modeling). To support independent exploratory and confirmatory analysis, we split the sample into two datasets by assigning every other respondent to an exploratory ($n = 1{,}381$) or confirmatory ($n = 1{,}380$) sample by sector and website in the order in which respondents completed the surveys. These sample sizes ensured that we far exceeded the recommended sample sizes for exploratory factor analysis, multiple regression, confirmatory factor analysis, and structural equation modeling, even after splitting the sample (Tonidandel et al., 2015).

Table 2. Exploratory factor analysis of the clutter items.

Item	Content	Design
Content_ALot	***.855***	.011
Content_TooMany	***.883***	−.034
Content_Space	***.881***	.035
Content_Distracting	***.897***	.016
Content_Irrelevant	***.774***	.033
Content_Annoying	***.892***	−.015
Design_HardToRead	−.114	***.832***
Design_SmallFont	−.084	***.778***
Design_DistractingColors	.024	***.765***
Design_UnpleasantLayout	.039	***.829***
Design_WhiteSpace	.086	***.723***
Design_TooMuchText	.063	***.776***
Design_NotLogical	.061	***.803***
Design_Disorganized	.035	***.844***
Design_VisualNoise	.219	***.664***
Design_HardToStart	.000	***.795***

Note: Bold values indicate factor loadings greater than .600.

3.1. Exploratory analyses

3.1.1. Factor analysis

A parallel analysis (O'Connor, 2000) of the clutter items indicated retention of two factors. Table 2 shows the alignment of items (identified with by item code) with factors from maximum likelihood factor analysis and Promax rotation (KMO = 0.95). Content and design items aligned as expected with Content and Design factors. The reliabilities (coefficient alpha) for the Content and Design factors were both 0.95; the overall reliability was 0.96.

3.1.2. Item analysis

Item loadings were especially high for content items due to high item correlations which is good for scale reliability, but indicates an opportunity to improve scale efficiency by removing some items. The situation was similar but not quite as extreme for the design items.

A common strategy for deleting items is to identify those with lower factor loadings. For example, for the Content factor the lowest item loading was for Content_Irrelevant (.774) and for the Design factor was Design_VisualNoise (.664). However, because we collected a measure of overall perceived clutter (Overall Clutter), we were able to use an alternative strategy of backward elimination regression

analysis to select the subset of clutter and design items that were best at accounting for variation in Overall Clutter.

3.1.3. Item retention

Backward regression of the six content items retained three: Content_ALot, Content_Space, and Content_Distracting, accounting for 35.5% of variation (adjusted-R^2) in Overall Clutter. Backward regression of the 10 design items plus deletion of items with negative beta weights retained three: Design_UnpleasantLayout, Design_TooMuchText, and Design_VisualNoise, accounting for 39% of variation (adjusted-R^2) in OverallClutter.

Backward regression of these six items revealed some evidence of variance inflation, and in this combination Content_Distracting no longer made a significant contribution to the model. After removing Content_Distracting, the remaining five items accounted for almost half of the variation in OverallClutter (adjusted-$R^2 = 45\%$) and all variance inflation factors (VIF) were less than 4. The reliabilities (coefficient alpha) for the revised Content and Design factors were, respectively, 0.91 and 0.88; the overall reliability was 0.90.

3.1.4. Exploratory validity

For the exploratory research, the method of consulting the literature and expert brainstorming to arrive at the initial item set established content validity for the clutter questionnaire (Nunnally, 1978). The expected alignment of items with factors in the factor analysis is evidence of construct validity. Evidence of concurrent validity of the clutter factors comes from their significant correlations with the single-item measure of overall clutter (Content Clutter: $r(1,379) = 0.60$; $p < 0.0001$; Design Clutter: $r(1,379) = 0.61$, $p < 0.0001$).

3.2. Confirmatory analyses

3.2.1. Initial item set

Figure 2 shows the item loadings for a confirmatory factor analysis (CFA) assuming no structure in the items (i.e., a one-factor model). Figure 3 shows the same items in a two-factor model (Content and Design).

There are many ways to assess the quality of CFA. Following the recommendations of Jackson et al. (2009), we focused on Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Bayesian Information Criterion (BIC). There are guidelines for good levels of model fit for CFI (>0.90) and RMSEA (<0.08), but not for BIC which is used to compare models (smaller is better).

For the one-factor model, the CFI was 0.74, RMSEA was 0.20, and BIC was 6,166. For the two-factor model, the CFI was 0.92, RMSEA was 0.11, and BIC was 2,144. Thus, accounting for the Content/Design two-factor structure led to better fit statistics including an acceptable level of CFI, but RMSEA was greater than 0.08.

3.2.2. Final item set

Figure 4 shows the CFA for the five items retained during the exploratory analyses (for the final version of the questionnaire with these five items, see Appendix Figure A2). The fit statistics for this model were excellent, with a CFI of 0.997, RMSEA of 0.047, and BIC of 96, using just the five items that were retained from the exploratory analysis of the separate independent set of data:

- **Content_ALot**: These types of content [ads, videos, suggested posts] made up a lot of the clutter.
- **Content_Space**: These types of content [ads, videos, suggested posts] too up too much space.
- **Design_UnpleasantLayout**: The layout was unpleasant.
- **Design_TooMuchText**: There was too much text.
- **Design_VisualNoise**: There was too much visual noise.

3.2.3. Structural equation model

Figure 5 is a structural equation model (SEM, all links significant, $p < 0.0001$) depicting how the final version of the clutter questionnaire's Content Clutter and Design Clutter factors drive Overall Clutter which, in turn, drives the SUPR-Q attitudinal factors of Appearance and Usability (both effects were significant, but the direct effect of Overall Clutter on Appearance was about five times that of the direct effect on Usability). The SUPR-Q Trust, Appearance, and Usability factors (attitudinal) have direct effects on the SUPR-Q Loyalty (behavioral intention) factor and also influence Loyalty through their effect on Brand Attitude, ultimately accounting for 62% of the variation in Loyalty ratings. The model has good fit statistics (CFI: 0.98, RMSEA: 0.08, BIC: 296).

3.2.4. Confirmatory validity

The CFA models confirmed the construct validity of the two-factor structure identified in the exploratory analyses. The SEM demonstrated convergent validity (significant beta weights for relationship of Content Clutter and Design Clutter with Overall Clutter) and divergent validity (strongest effect of Overall Clutter on Appearance, no significant effect of Overall Clutter on Trust).

3.4. Sensitivity and range analyses

Using the full dataset ($n = 2{,}761$), we conducted ANOVAs to check the sensitivity (significance of the main effect of website) of the three clutter metrics, all of which were statistically significant:

- **Content Clutter**: Mean of Content_ALot and Content_Space ($F(55, 2760) = 6.3$, $p < 0.0001$, $\eta^2 = 0.11$)
- **Design Clutter**: Mean of Design_UnpleasantLayout, Design_TooMuchText, and Design_VisualNoise ($F(55, 2760) = 9.5$, $p < 0.0001$, $\eta^2 = 0.16$)
- **Overall Clutter**: The one-item measure of overall clutter ($F(55, 2760) = 3.9$, $p < 0.0001$, $\eta^2 = 0.07$)

Along with sensitivity to manipulation, we assessed the range of these metrics across websites (after rescaling to a

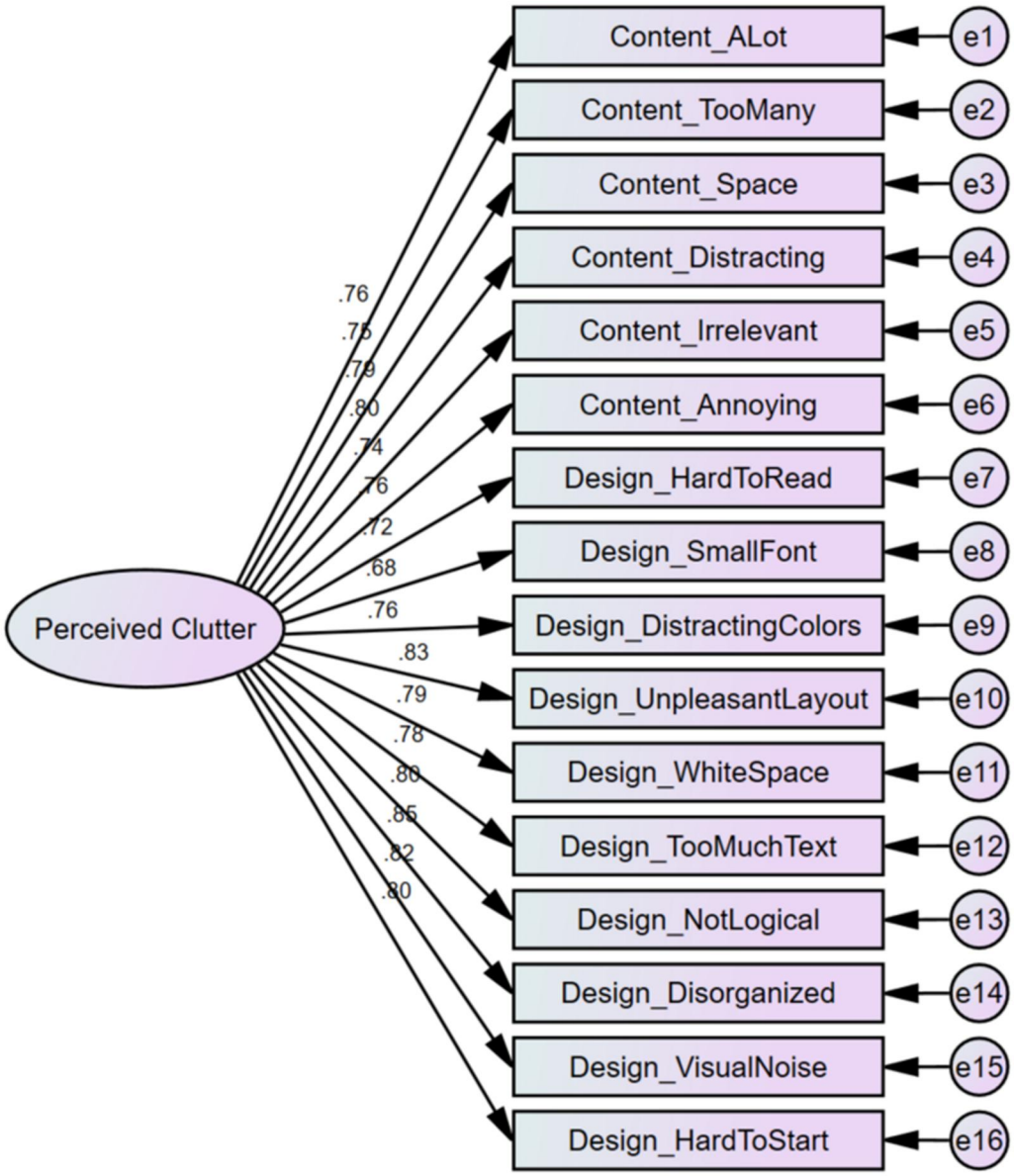


Figure 2. One-factor CFA model with initial item set.

common 0–100-point scale for ease of comparison) to get a sense of the extent to which the dataset included websites with different levels of clutter. The distributions are shown in Figure 6 and summarized in Table 3.

Design Clutter scores tended to run lower than Content Clutter scores, a 10-point difference in medians (50th percentiles). For Content Clutter and Design Clutter the range of scores was slightly more than half of the possible range of the metric. The range for Overall Clutter was a little more restricted, covering about 40% of the possible range of the metric. The 5th–95th percentiles for the metrics were from 20 to 51 for Content Clutter, 12 to 41 for Design Clutter, and 20 to 45 for Overall Clutter. None of the websites had a mean score on these metrics higher than 65.

4. Summary and discussion

The perceived clutter of websites is a potentially important but understudied construct in UX research. In this paper we described the development and assessment of a standardized questionnaire for its reliable and valid measurement.

4.1. Hypothesized structure and initial items

Starting with the linguistic definition of clutter, we hypothesized a logical distinction between content and design clutter applied to website design. Content clutter is made up of elements (e.g., ads and videos) that are not directly relevant to user tasks, making them potential candidates for discarding. Design clutter is the consequence of specific elements of poor general design that have significant statistical relationships with overall impressions of clutter. The items in our initial set were included because we found advocates for them in the non-peer reviewed literature on decluttering websites and/or in the peer-reviewed literature on display clutter and ad clutter. After brainstorming with our team of UX researchers, we developed an initial questionnaire with one item for the overall assessment of perceived clutter

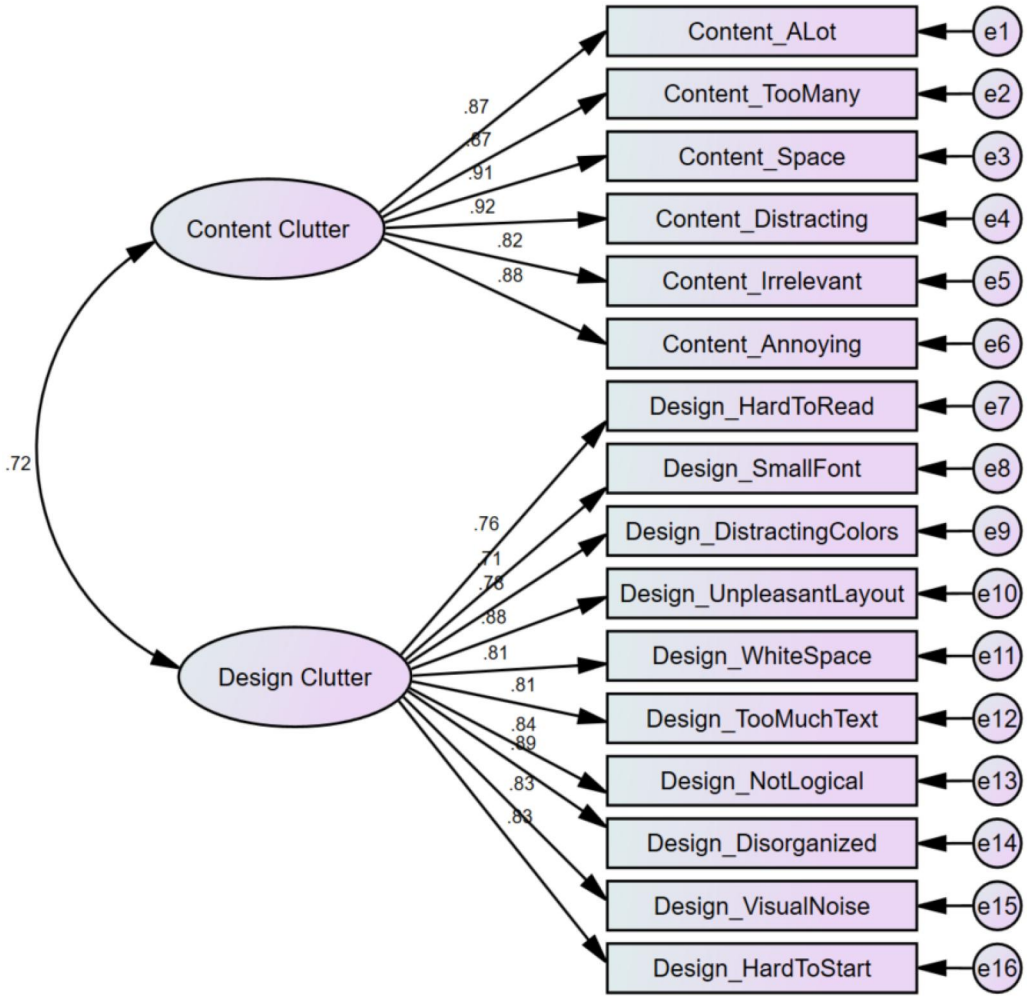


Figure 3. Two-factor CFA model with initial item set.

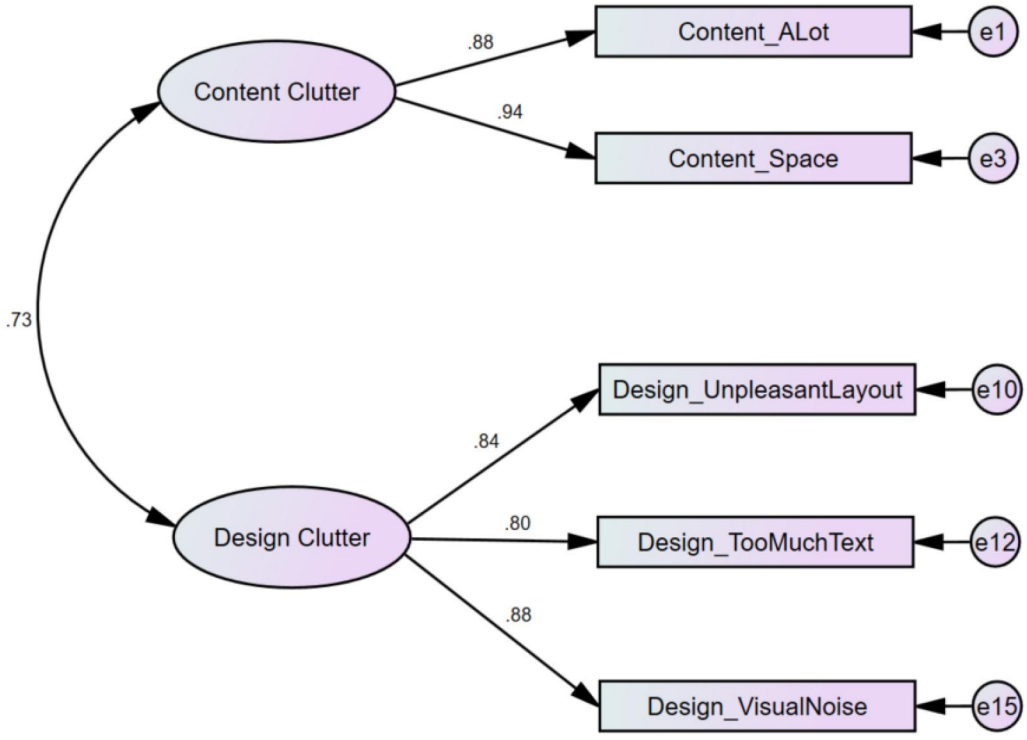


Figure 4. Two-factor CFA model with final five-item set.

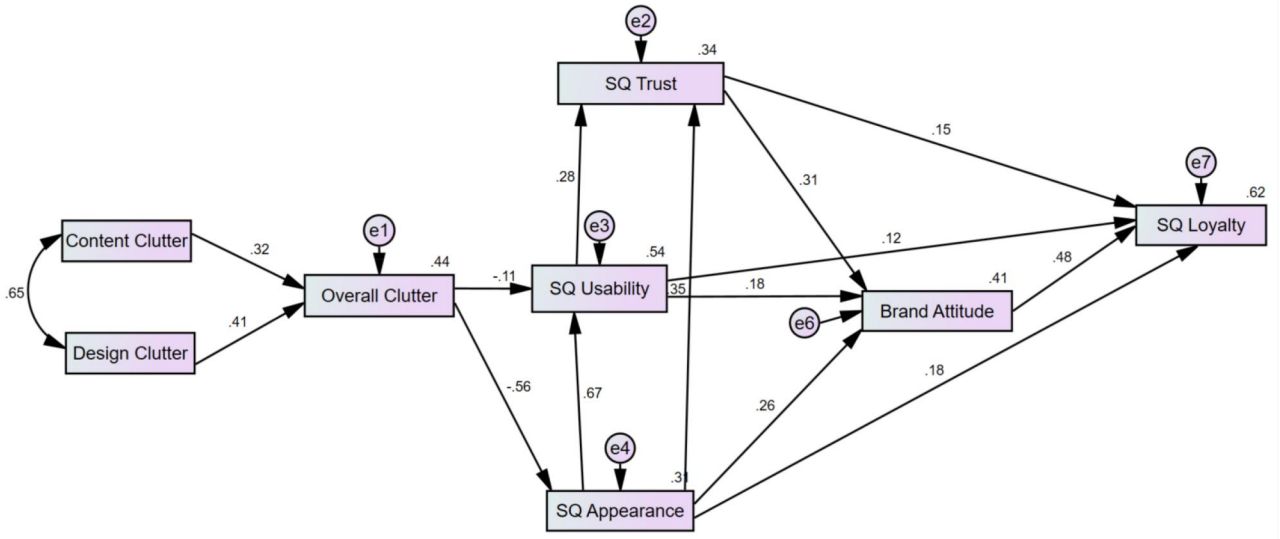


Figure 5. Structural equation model of influence of perceived clutter on SUPR-Q factors.

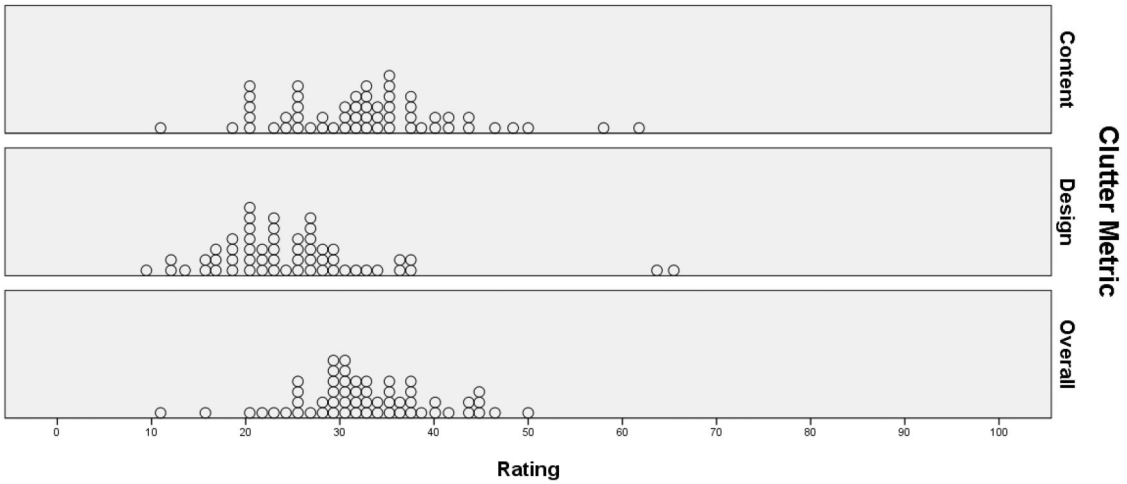


Figure 6. Dotplots of the distributions of Content Clutter, Design Clutter, and Overall Clutter across the websites included in the consumer surveys.

Table 3. Summary of distributions for Content Clutter, Design Clutter, and Overall Clutter after conversion to a 0–100-point scale.

Clutter metric	Min	5th %ile	10th %ile	25th %ile	50th %ile	75th %ile	90th %ile	95th %ile	Max	Range
Content	11	20	21	26	33	37	44	51	62	51
Design	9	12	16	20	23	29	36	41	65	56
Overall	11	20	23	29	32	38	44	45	50	39

(Overall Clutter), six items for Content Clutter, and ten items for Design Clutter (see Appendix Figure A1). As is typical in the development of standardized questionnaires, we started with more items than we expected to keep.

4.2. Data collection

We included the initial clutter questionnaire in eight retrospective consumer surveys conducted between April 2022 and January 2023. Each survey targeted a specific sector and, in total, we collected 2,761 responses to questions about the UX of 57 websites. These questions included the initial version of the clutter questionnaire, the SUPR-Q questionnaire, and a brand attitude item. The sample had roughly equal representation of gender and age (split at 35 years old).

The data were split into two groups with roughly equal numbers of responses for each sector and website. The purpose of this split was to have separate datasets for exploratory and confirmatory analyses.

4.3. Exploratory analyses

We started the exploratory analyses with a parallel analysis of the content and design clutter items, which, as expected, indicated retention of two factors. Exploratory factor analysis of those items showed the expected alignment of items with the two factors. The reliability of the resulting content

and design scales were very high (.95), indicating a reasonable opportunity to reduce the number of items in the scales to achieve more efficient measurement while maintaining acceptably high reliability.

Our strategy for identifying the items to retain was to conduct backward stepwise regressions for each of the two sets of items (content and design) to see which items worked best to account for variation in the Overall Clutter metric. This process resulted in the retention of five items, two related to content clutter and three related to design clutter. The reliabilities of the resulting Content Clutter and Design Clutter scales were, respectively, 0.91 and 0.88. Thus, this exercise led to scales that had slightly lower but still high levels of reliability with five items instead of 16. In addition to the acceptable levels of reliability, the method of questionnaire construction and evaluation, factor analysis, and correlation with overall clutter established, respectively, content, construct, and concurrent validity.

4.4. Confirmatory analyses

Switching to the data we set aside for confirmation, a two-factor CFA of the five-item version of the clutter questionnaire had excellent fit statistics (CFI = 0.997, RMSEA = 0.047, BIC = 96), better than a similar two-factor CFA of the 16-item version (CFI = 0.92, RMSEA = 0.11, BIC = 2,144).

We then built an SEM that demonstrated the strengths of the relationships between Content Clutter and Design Clutter with Overall Clutter and the connections between Overall Clutter and the SUPR-Q Trust, Usability, and Appearance scales, followed by the connections between those SUPR-Q scales and the constructs of Brand Attitude and SUPR-Q Loyalty. The fit statistics for the SEM were good (CFI = 0.98, RMSEA = 0.08, BIC = 296), ultimately accounting for 62% of variation in the Loyalty scale. Together, the CFA and SEM analyses provided additional confirmation of the construct validity, convergent validity, and divergent validity of the two-factor model of perceived clutter.

Focusing on the clutter constructs in the model, together Content Clutter and Design Clutter accounted for 44% of variation in Overall Clutter, with a beta weight of 0.32 for Content Clutter and 0.41 for Design Clutter. The weights were significant ($p < 0.0001$), but left 56% of variation in Overall Clutter unaccounted for, leaving open the possibility of additional, as yet undiscovered, clutter factors that would account for some of its remaining variability.

The link between Overall Clutter and Trust was not significant, but its links with Appearance and Usability were ($p < 0.0001$). As expected, Overall Clutter had more influence (about five times as much) on ratings of Appearance than Usability (-.11 for Usability, −.56 for Appearance). The links between Overall Clutter and SUPR-Q factors were negative because higher ratings of clutter indicate poorer UX while higher ratings of SUPR-Q factors indicate better UX.

4.5. Sensitivity and range analyses

The websites we surveyed in this research were all from well-known professionally designed commercial websites. Our sensitivity analyses of Content Clutter, Design Clutter, and Overall Clutter showed significant variation in the means of these metrics by website, but our analyses of the ranges of these values showed that none of them, after rescaling values to 0–100-point scales, had any clutter score greater than 65. The observed ranges of Content Clutter and Design Clutter covered about half of the possible range of those metrics; Overall Clutter covered about 40% of its possible range.

For the eight surveys we conducted, our focus was to gather information about top websites in their sectors, so we did not focus on including websites with unusually high levels of clutter. There is some possibility that inclusion of very cluttered websites might have led to different analytical solutions. That said, our exploratory and confirmatory analyses are appropriate for the types of websites we typically study in our consumer surveys, and may work well when assessing very cluttered websites. We definitely did not see evidence of ceiling or floor effects with these clutter metrics.

4.6. Limitations and future research

There are four key limitations of this study, with implications for future research.

4.6.1. Research methodology

First, our data were collected in retrospective consumer surveys and not in task-based usability studies of consumer websites. We believe that the final five-item version of the clutter questionnaire could be of value in the post-task section of usability studies, especially if there are concerns about the effect of clutter on the user experience.

For example, Lewis and Mayes (2015) developed long and short forms of the Emotional Metrics Outcome (EMO) questionnaire using data from two retrospective consumer surveys with sample sizes of 2,600 for the first survey and 1,000 for the second. The next year, Lewis et al. (2015) used the EMO in a large-sample unmoderated usability study ($n = 471$) and confirmed the EMO structure with this independent set of data collected using a different research methodology. It would be reasonable to attempt this type of replication in future research with the clutter questionnaire.

4.6.2. Type of interface

It is an open question whether these results would generalize to other web-like interfaces, for example, mobile web or mobile apps. There is nothing in the content of the five clutter items that seems obviously problematic in these other contexts of use, but this is another potential topic for future research. Given its measurement of the contribution of ads to perceived clutter, it is unlikely that this questionnaire would be useful in ad-free contexts.

4.6.3. Possible range restriction

None of the clutter scores obtained in this research exceeded 65 (on a 0–100-point scale where larger numbers indicate more clutter and a poorer experience). A potential future research project would be to exert more control over the amount of clutter in the rated websites to get a more direct estimate of clutter scale sensitivities with particular emphasis on measuring perceived clutter in very cluttered websites.

4.6.4. Number of clutter factors

We developed two clutter factors, Content Clutter and Design Clutter. In both exploratory and confirmatory analyses, metrics based on these factors accounted for about 45% of the variation in Overall Clutter ratings. That is a very good model, but with 55% of variation unaccounted for there appears to an opportunity to conduct research (e.g., cognitive interview studies) to discover other factors that would increase the explanatory power of the model (e.g., further reduce RMSEA).

5. Conclusions

This paper presents a new standardized questionnaire for the measurement of perceived clutter in websites. After exploratory and confirmatory analysis, the final version of the questionnaire had five items measuring two clutter factors: Content Clutter (influenced by the amount of screen space take by irrelevant ads or videos) and Design Clutter (influenced by poor design of relevant content such as too much text, an unpleasant layout, or too much visual noise). These factors accounted for a significant amount (45%) of the variability in a concurrently collected item for Overall Clutter, which in turn was shown to significantly account for variation in Appearance and Usability constructs.

Like other standardized UX questionnaires, the number on a clutter scale doesn't provide specific guidance on exactly what to fix but can provide high-level indications. For example, high scores on Content Clutter indicate different interventions from high scores on Design Clutter.

We expect UX researchers and practitioners to be able to use this version of the clutter questionnaire when the research context is similar to the websites we studied in our consumer surveys. We don't anticipate serious barriers to using the clutter questionnaire in other similar contexts including task-based studies, mobile apps, and very cluttered web/mobile UIs, but because that research has not yet been conducted, UX researchers and practitioners should exercise due caution.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

James R. Lewis http://orcid.org/0000-0002-3295-5392

References

Brinson, N. H., Eastin, M. S., & Cicchirillo, V. J. (2018). Reactance to personalization: Understanding the drivers behind the growth of ad blocking. *Journal of Interactive Advertising*, *18*(2), 136–147. https://doi.org/10.1080/15252019.2018.1491350

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, *103*(2), 276–279. https://doi.org/10.1037/0033-2909.103.2.276

Crowley, L. (2017, June 20). *Is my website too cluttered?* Eyequant. https://www.eyequant.com/resources/is-my-website-too-cluttered/

Edwards, S. M., Li, H., & Lee, J.-H. (2002). Forced exposure and psychological reactance: Antecedents and consequences of the perceived intrusiveness of pop-up ads. *Journal of Advertising*, *31*(3), 83–95. https://doi.org/10.1080/00913367.2002.10673678

HFES. (2020, July 15). *Code of ethics*. Human Factors and Ergonomics Society. https://www.hfes.org/about-hfes/code-of-ethics

Hughes, J. (2022, September 8). *Declutter your website*. Themeisle. https://themeisle.com/blog/remove-clutter-from-website/

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23. https://doi.org/10.1037/a0014694

Kaber, D. B., Alexander, A. L., Stelzer, E. M., Kim, S.-H., Kaufmann., & K., Hsiang. (2008). Perceived clutter in advanced cockpit displays: Measurement and modeling with experienced pilots. *Aviation Space and Environmental Medicine*, *79*, 1–12. https://doi.org/10.3357/ASEM.2319.2008

Kaber, D., Kaufmann, K., Alexander, A. L., Kim, S.-H., Naylor, J. T., Prinzel, L. J., III, Pankok, Jr., C., & Gil, G.-H. (2013). Testing and validation of a psychophysically defined metric of display clutter. *Journal of Aerospace Information Systems*, *10*(8), 359–368. https://doi.org/10.2514/1.I010048

Kim, N. Y., & Sundar, S. S. (2010). Relevance to the rescue: Can "smart ads" reduce negative response to online ad clutter? *Journalism & Mass Communication Quarterly*, *87*(2), 346–362. https://doi.org/10.1177/107769901008700208

Lewis, J. R., Brown, J., & Mayes, D. K. (2015). Psychometric evaluation of the EMO and the SUS in the context of a large-sample unmoderated usability study. *International Journal of Human-Computer Interaction*, *31*(8), 545–553. https://doi.org/10.1080/10447318.2015.1064665

Lewis, J. R., & Mayes, D. K. (2015). Development and psychometric evaluation of the Emotional Metric Outcomes (EMO) Questionnaire. *International Journal of Human-Computer Interaction*, *30*(9), 685–702. https://doi.org/10.1080/10447318.2014.930312

Merriam-Webster (n.d.). *Clutter*. https://www.merriam-webster.com/dictionary/clutter

Moacdieh, N., & Sarter, N. (2015). Display clutter: A review of definitions and measurement techniques. *Human Factors*, *57*(1), 61–100. https://doi.org/10.1177/0018720814541145

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill. https://doi.org/10.1177/0018720814541145

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, *32*(3), 396–402. https://doi.org/10.3758/bf03200807

Oxford Dictionary. (n.d.). *Clutter*. https://www.oed.com/search/dictionary/?scope=Entries&q=clutter

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*(2), 17.1–22. https://doi.org/10.1167/7.2.17

Sauro, J. (2015). SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of Usability Studies*, *10*(2), 68–86. https://dl.acm.org/doi/10.5555/2817315.2817317

Sauro, J., Jenks, S., Short, E., Atkins, D., Lewis, J. R. (2023, January 24). *UX and NPS benchmarks of clothing retail websites*. MeasuringU. https://measuringu.com/clothing-benchmark-2023/

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience* (2nd ed.). Morgan Kaufmann.

Sauro, J., Lewis, J. R., Metzler, G. (2022, July 26). *UX and NPS benchmarks of airline websites*. MeasuringU. https://measuringu.com/airlines-benchmark-2022/

Sauro, J., Lewis, J. R., Metzler, G. (2022, September 27). *UX and NPS benchmarks of business information websites*. MeasuringU. https://measuringu.com/business-information-benchmark-2022/

Sauro, J., Lewis, J. R., Short, E. (2022, June 14). *UX and NPS benchmarks of real estate websites*. MeasuringU. https://measuringu.com/real-estate-benchmark-2022/

Sauro, J., Lewis, J. R., Yazvec, M., Nawalaniec, N. (2022, October 17). *UX and NPS benchmarks of ticketing websites*. MeasuringU. https://measuringu.com/ticketing-benchmark-2022/

Sauro, J., Metzler, G., Lewis, J. R. (2022, August 30). *UX and NPS benchmarks of travel aggregator websites*. MeasuringU. https://measuringu.com/travel-benchmark-2022/

Saxena, R. (2021, November 19). *Four ways to avoid a cluttered website*. Kōmaya. https://www.komaya.com/four-ways-to-avoid-a-cluttered-website/

Schiavone, W., Sauro, J., Short, E., Lewis, J. R. (2023, February 21). *UX and NPS benchmarks of wireless service provider websites*. MeasuringU. https://measuringu.com/wireless-benchmark-2023/

Senarathna, T., & Wijetunga, D. (2023). Examining some dynamics related to YouTube ad clutter in a high-clutter context. *South Asian Journal of Marketing*. https://doi.org/10.1108/SAJM-04-2023-0025

Speck, P. S., & Elliott, M. T. (1997). The antecedents and consequences of perceived advertising clutter. *Journal of Current Issues & Research in Advertising*, *19*(2), 39–54. https://doi.org/10.1080/10641734.1997.10524436

SUPR-Q (n.d.). *Product description*. MeasuringU. https://measuringu.com/product/suprq/

Tonidandel, S., Williams, E. B., & LeBreton, J. M. (2015). Size matters … just not in the way you think: Myths surrounding sample size requirements for statistical analysis. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 162–183). Routledge.

Tullis, T. S. (1983). The formatting of alphanumeric displays: A review and analysis. *Human Factors*, *25*(6), 657–682. https://doi.org/10.1177/001872088302500 6

UXPA. (n.d.). *UXPA code of professional conduct*. User Experience Professionals Association. https://uxpa.org/uxpa-code-of-professional-conduct/

Vocabulary.com. (n.d.). *Declutter*. https://www.vocabulary.com/dictionary/declutter

Xia, L., & Sudharshan, D. (2002). Effects of interruptions on consumer online decision processes. *Journal of Consumer Psychology*, *12*(3), 265–280. https://doi.org/10.1207/S15327663JCP1203_08

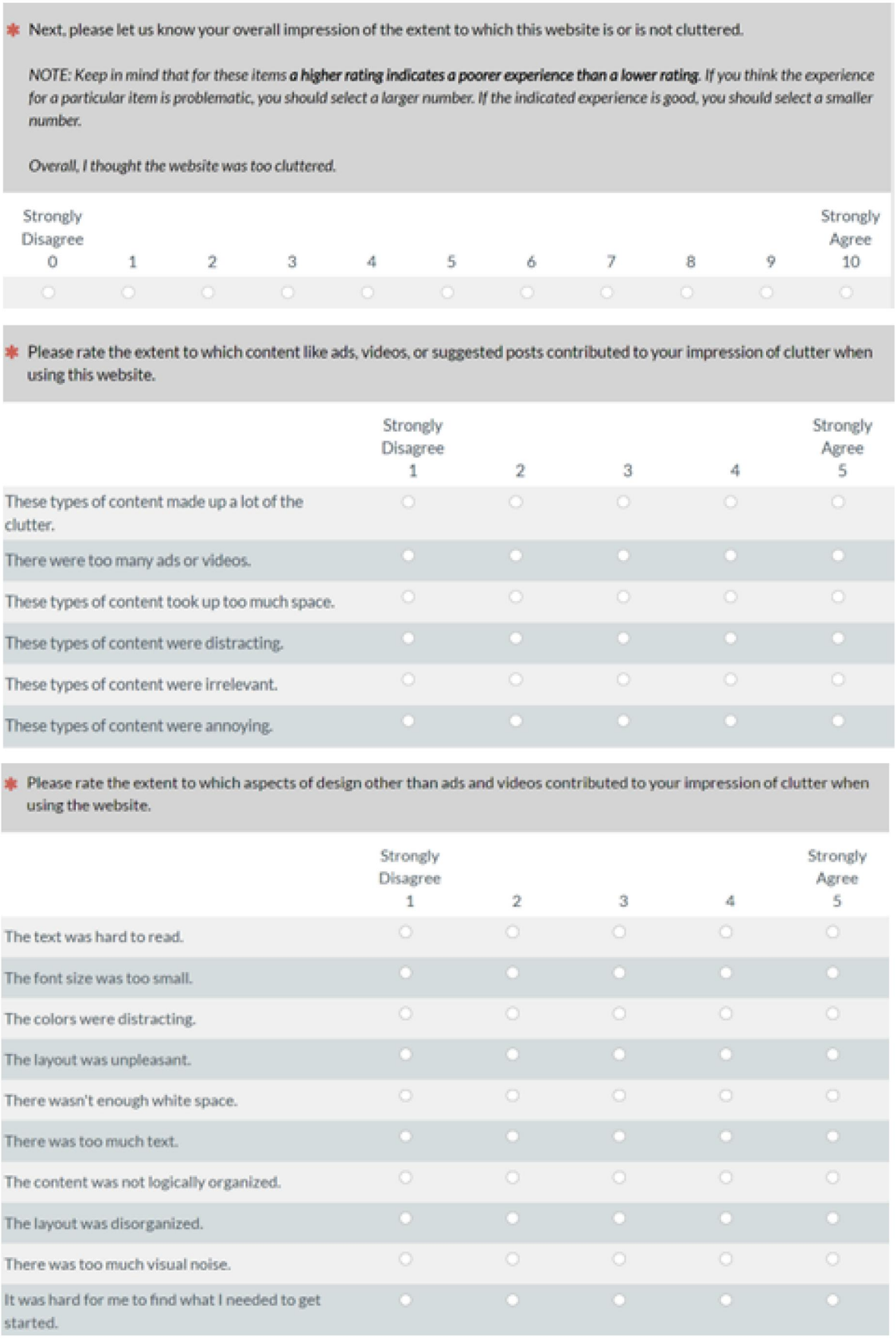
About the authors

James R. Lewis is a distinguished user experience researcher at MeasuringU, an IBM Master Inventor emeritus (over 90 US patents), the author of five books and over 100 peer-reviewed publications, and a member of the Academy of Science, Engineering, and Medicine of Florida.

Jeff Sauro is the founder/CEO of MeasuringU. He is a pioneer in quantifying the user experience, widely recognized for making statistical concepts understandable and actionable, author of over 25 research papers and seven books, including Surveying the User Experience, Benchmarking the User Experience, and Quantifying the User Experience.

Appendix A

This appendix includes the initial and final versions of the PCW questionnaire.



Appendix Figure A1. First iteration of a standardized questionnaire for the measurement of perceived clutter of websites (overall question presented first then, on separate screens, followed by content and design clutter grids with item order randomized in each grid).

* Next, please let us know your overall impression of the extent to which this website is or is not cluttered.

NOTE: Keep in mind that for these items ***a higher rating indicates a poorer experience than a lower rating****. If you think the experience for a particular item is problematic, you should select a larger number. If the indicated experience is good, you should select a smaller number.*

Overall, I thought the website was too cluttered.

Strongly Disagree 0	1	2	3	4	5	6	7	8	9	Strongly Agree 10
○	○	○	○	○	○	○	○	○	○	○

* Please rate the extent to which content like ads, videos, or suggested posts contributed to your impression of clutter when using this website.

	Strongly Disagree 1	2	3	4	Strongly Agree 5
These types of content made up a lot of the clutter.	○	○	○	○	○
These types of content took up too much space.	○	○	○	○	○

* Please rate the extent to which aspects of design other than ads and videos contributed to your impression of clutter when using the website.

	Strongly Disagree 1	2	3	4	Strongly Agree 5
The layout was unpleasant.	○	○	○	○	○
There was too much text.	○	○	○	○	○
There was too much visual noise.	○	○	○	○	○

Appendix Figure A2. Final version of the perceived clutter of websites (PCW) Questionnaire.