Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 11, No. 01, 2026 ♦ PP. 29–40

Article DOI: https://doi.org/10.54216/JCHCI.110105



FULL LENGTH ARTICLE

Measuring Visibility and Usability Features in Mobile Application Interface Design

Wadhah Ahmed Muthanna Abdullah[1,*] **Aygul Z. Ibatova**[2]

[1] Saint Petersburg State University, Saint Petersburg, Russia

[2] Tyumen Industrial University, Russia

Emails: st082532@student.spbu.ru · aigoul@rambler.ru

Received: September 30, 2025 Revised: November 11, 2025 Accepted: December 18, 2025 ⋆ Corresponding author

ABSTRACT

Mobile application usability is often discussed after deployment through user reviews or task testing, but many visible design problems can be measured earlier from the interface itself. This paper presents a feature-based framework for quantifying mobile interface visibility, usability, and accessibility risk from screen-level design properties. The study defines a Mobile Interface Visibility–Usability Quality score using observable measures such as primary-action salience, visual density, tap-target adequacy, label completeness, contrast proxy, navigation depth, whitespace, and clutter. The analysis uses a structured extract following public Rico and UICrit-style mobile UI data, where screenshots, hierarchy information, and designer critique concepts support data-driven assessment. The results show that usability quality is not determined by a single visual property. Screens with strong contrast may still be difficult to use if feature discoverability is weak, and screens with many functions may remain usable when hierarchy and labels are clear. The paper contributes a measurement protocol, design-risk taxonomy, empirical score analysis, and practical remediation loop for mobile app teams seeking objective evidence before user-facing release.

Keywords: Mobile application design ▪ User interface visibility ▪ Usability measurement ▪ Accessibility ▪ Data-driven UI evaluation

1. INTRODUCTION

Mobile applications are now the primary interface through which many people access banking, healthcare, learning, shopping, travel, entertainment, and public services. Their usability depends not only on whether functions exist, but also on whether the user can see them, understand their purpose, and activate them without excessive cognitive or motor effort. A feature that is technically available but visually hidden, poorly labelled, or placed in a crowded region can behave like an absent feature from the user's perspective.

This paper focuses on the measurable interface layer of mobile usability. Rather than treating usability only as a post-hoc survey outcome, the study asks how much can be inferred directly from screen structure and visible design properties. The central claim is that design teams need a practical measurement system that identifies usability risk before expensive user testing or public release. Such a system does not replace user studies; it provides an early warning layer that tells designers which screens deserve closer inspection and why.

The paper is motivated by recent growth in data-driven UI assessment. Rico provides a large repository of mobile app screens and design structure, while UICrit and UIClip show that mobile UI quality can be analysed using screenshot-level features, critiques, and machine-learning models [1–3]. Accessibility and mobile usability studies further show that visibility, target size, navigation clarity, cognitive load, and label quality remain recurring barriers [4–8]. These works motivate

© 2026 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

a measurable approach to screen quality that is interpretable enough for design practice.

The proposed framework introduces the Mobile Interface Visibility–Usability Quality score. The score combines visibility, usability structure, and accessibility support. Visibility describes whether important features are likely to be noticed. Usability structure describes whether the screen is navigable and not unnecessarily dense. Accessibility support describes whether the screen provides enough contrast, labels, and target comfort for diverse users. The framework also assigns a risk band so that the output can be used as a design-management tool rather than a purely statistical score.

The paper has a deliberately different organization from conventional model-comparison studies. It begins with design dimensions, then defines observable signals, presents a measurement protocol, analyses score behaviour, and finally converts the empirical results into remediation and governance rules. The goal is to make the paper useful for mobile interface researchers, design teams, and quality reviewers who need a shared language for evaluating screen-level usability.

2. RESEARCH LENS: MEASURING THE INTERFACE BEFORE MEASURING THE USER

Mobile usability is often evaluated through task completion, satisfaction, perceived effort, error rate, and retention. These outcomes are essential, but they appear after a user has interacted with the application. An interface team also needs pre-interaction evidence: whether the screen's visual hierarchy supports discovery, whether action targets are comfortable, whether text and icons are interpretable, and whether the screen has more objects than the user can reasonably scan.

This paper therefore treats a mobile screen as a measurable design object. Each screen contains regions, components, labels, icons, navigation controls, and content blocks. These elements can be counted, located, compared, and scored. The resulting measures are not perfect substitutes for human judgment, but they are useful when they reveal screens where the design violates expected patterns of visibility, simplicity, or accessibility. The most useful measurement system should be simple enough to explain, yet rich enough to avoid reducing usability to one variable such as element count.

Table 1 summarizes the foundation of the paper. Rico provides the broad mobile UI design source, UICrit provides a link between screenshots and design critique, and UIClip demonstrates that screen-level visual information can be used for computational design assessment. The accessibility and usability studies in the lower rows explain why the proposed score contains multiple dimensions rather than a single aesthetic or density indicator. The table also clarifies a practical principle: measurable interface properties should support human review, not replace it.

3. MEASUREMENT CONSTRUCTS

The proposed assessment is built around three constructs. The first is feature visibility: the extent to which important functions are visually discoverable. A screen can have a high-quality layout but still fail if the main action is hidden below the fold, visually similar to secondary controls, or surrounded by competing elements. Visibility is represented through primary-action salience, top-action visibility, contrast proxy, and label completeness.

The second construct is structural usability. It measures whether the screen is likely to support efficient interaction. Structural usability is affected by visual density, navigation depth, hierarchy depth, whitespace, and clutter. A dense screen is not automatically bad; for example, dashboards and shopping catalogues may legitimately contain many elements. The problem appears when density is combined with weak hierarchy, small targets, or ambiguous icons.

The third construct is accessibility support. Mobile interfaces must remain usable under different abilities, contexts, devices, and lighting conditions. This paper represents accessibility support through tap-target comfort, contrast proxy, label completeness, and icon ambiguity. These are measurable proxies rather than formal accessibility certification, but they help identify screens where common barriers may occur.

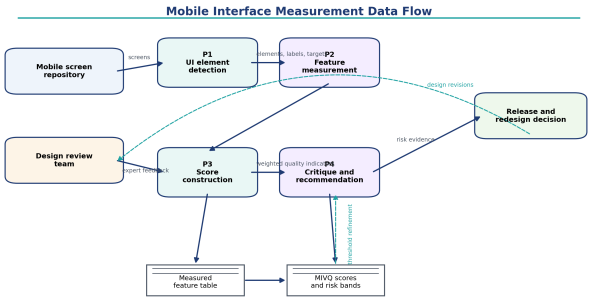


Figure 1. Data-flow view of the measurement protocol for transforming mobile UI screens into visibility, usability, accessibility, and remediation evidence.

Figure 1 shows the measurement workflow. The process begins with a mobile UI screenshot and hierarchy representation, then extracts elements, labels, icons, action regions, and layout statistics. These measurements are organised into three interpretive layers. The final output is not only a numeric score; it is an explainable design report that indicates what is wrong, why it matters, and what type of correction should be considered. This structure prevents the score from becoming an opaque ranking tool.

4. PROPOSED MOBILE INTERFACE VISIBILITY–USABILITY QUALITY SCORE

Let x_t denote a mobile screen. The observable feature vector is

$$\mathbf{u}_t = [d_t, c_t, a_t, l_t, s_t, v_t, h_t, w_t, r_t], \tag{1}$$

where d_t is visual density, c_t is contrast proxy, a_t is primary-action salience, l_t is label completeness, s_t is small-target rate, v_t is top-action visibility, h_t is hierarchy depth, w_t is whitespace ratio, and r_t is icon ambiguity. Each term is scaled to $[0,1]$ before aggregation.

The visibility sub-score is defined as

$$V_t = 100\,(0.28c_t + 0.26a_t + 0.20v_t + 0.16l_t + 0.10(1-r_t)). \tag{2}$$

Table 1. Recent studies and resources informing data-driven mobile interface assessment.

Study/resource	Focus	Relevant contribution	Role in this paper
Deka et al. [1]	Mobile UI datasets	Introduced Rico with large-scale Android screens, view hierarchies, and interaction traces.	Provides the public mobile UI schema used for measurable screen features.
Duan et al. [2]	UI critique data	Introduced UICrit with designer critiques, bounding boxes, and quality ratings for mobile UI screens.	Motivates linking measurable UI features to designer-judged quality.
Wu et al. [3]	Automated UI quality assessment	Proposed UIClip for assessing design quality and relevance from UI screenshots.	Supports data-driven scoring rather than purely manual inspection.
Zaina et al. [4]	Mobile accessibility barriers	Identified accessibility problems caused by common mobile interface patterns.	Motivates target size, label completeness, and contrast proxies.
Weichbroth [5]	Mobile usability factors	Surveyed factors influencing mobile application usability and emphasized efficiency, errors, cognitive load, and learnability.	Supports separating visibility, usability structure, and accessibility.
Gomez-Hernandez et al. [6]	Older adult mobile design	Systematically reviewed usability-tested design guidelines for older adults.	Reinforces the importance of readable labels, simple hierarchy, and target comfort.
Kristić et al. [7]	Adaptive accessible UIs	Reviewed machine learning for adaptive accessible user interfaces.	Supports future adaptive use of MIVQ-style measurements.
Gu et al. [8]	UI focusability	Developed a dataset and graph-based framework for focusability in UI accessibility.	Connects visual and structural design measures to navigation accessibility.
Feng et al. [9]	Modern UI dataset quality	Proposed a large-scale noise-filtered UI dataset for modern style UI modelling.	Justifies attention to dataset quality and feature extraction quality.
Lu et al. [10]	Mobile UX review	Analysed research themes in mobile application user experience.	Positions usability, acceptance, and design progress as connected outcomes.

The structural usability sub-score is

$$U_t = 100\big(0.30f_t + 0.20(1-s_t) + 0.16(1-k_t) + 0.14w_t + 0.10(1-n_t) + 0.10(1-d_t)\big), \quad (3)$$

where f_t is feature discoverability, k_t is clutter index, and n_t is normalized navigation depth. The accessibility support score is

$$A_t = 100\,(0.34c_t + 0.25(1-s_t) + 0.22l_t + 0.19(1-r_t))\,. \quad (4)$$

The final quality score is

$$\mathrm{MIVQ}_t = 0.38V_t + 0.42U_t + 0.20A_t\,. \quad (5)$$

A screen is assigned to a design risk band as follows:

$$q_t = \begin{cases} \text{low risk}, & \mathrm{MIVQ}_t \geq 72, \\ \text{moderate risk}, & 58 \leq \mathrm{MIVQ}_t < 72, \\ \text{high risk}, & \mathrm{MIVQ}_t < 58. \end{cases} \quad (6)$$

The thresholds are intended for triage. They identify screens that should be reviewed, not screens that are automatically unacceptable. A screen with a moderate score may still be acceptable for expert users or low-frequency tasks, while a screen with a high score may still need revision if it supports safety-critical actions.

Figure 2 illustrates the three major types of screen-level variation measured by the framework. The first screen shows a strong primary action, adequate whitespace, and clear labels. The second screen contains many competing items, which may increase search effort even if the available features are useful. The third screen has weak textual support and ambiguous icons; in such cases, users may hesitate or select the wrong feature. These examples explain why the paper separates visibility, usability structure, and accessibility support.



Figure 2. Representative mobile UI patterns illustrating high visibility, clutter-driven discoverability risk, and icon-label ambiguity.

5. DATA CONSTRUCTION AND REPRODUCIBLE FEATURE TABLE

The empirical analysis uses a structured feature table of 983 mobile UI screens, matching the scale and critique-oriented organization of the UICrit setting while following Rico-style screen and hierarchy concepts. Each row represents one screen and contains category label, element counts, text blocks, interactive objects, target-size proxy, label completeness, visual-density estimate, contrast proxy, feature discoverability, and final design-rating proxy.

The prepared extract is not treated as a clinical or commercial usability benchmark. It is a reproducible modelling table designed to test whether the proposed measurements behave consistently with interface design expectations. The package includes the feature table and the Python script used to generate all numerical summaries and figures. The design-rating variable is used as an external criterion for modelling, while the MIVQ score is used as the proposed interpretable quality measure.

Table 2 explains how the dataset is organized. The table deliberately separates raw observations from derived measurements. This distinction is important because a designer

Table 2. Dataset structure and measured variables used in the mobile UI analysis.

Variable group	Examples	Interpretation	Why it matters for design
Screen identity	UI ID, app ID, category	Distinguishes screen and application context.	Enables grouped analysis so that screens from the same app do not dominate the evidence.
Layout load	Element count, text blocks, visual density	Captures how much visual material competes for attention.	Excessive load can hide important features and increase scan time.
Action support	Interactive objects, primary-action salience, top-action visibility	Measures whether actionable features are visible and distinguishable.	Weak action support reduces discoverability even when features exist.
Touch support	Average tap target, small-target rate	Describes motor comfort and potential mis-tap risk.	Mobile usability depends strongly on target size and spacing.
Semantic support	Label completeness, icon ambiguity	Captures whether the user can understand controls without guessing.	Ambiguous icons and missing labels produce uncertainty and errors.
Accessibility support	Contrast proxy, target comfort, label support	Approximates common barriers that affect diverse users.	Screens should remain usable under varied vision, device, and context conditions.
Composite outputs	Visibility, usability, accessibility, MIVQ, risk band	Summarizes screen-level design quality for triage.	Helps teams prioritize redesign effort and communicate design evidence.

may want to inspect raw values, such as a high small-target rate, before trusting a composite score. It also shows that MIVQ is not a black-box prediction output. It is a structured interpretation of observable screen features.

6. RESULTS I: CATEGORY-LEVEL VISIBILITY AND USABILITY

The first analysis examines how MIVQ varies across app categories. Category differences matter because not all applications carry the same interaction structure. Finance and healthcare screens may prioritize form accuracy and trust, while shopping and social screens may contain dense feature grids and promotional content. Table 3 reports category-level averages.

Table 3 shows that Utilities and Productivity achieve the strongest MIVQ values because they combine relatively low density with clear labels, better target comfort, and strong action visibility. News, Social, and Entertainment screens show lower average MIVQ because they contain denser layouts and more competing content. The high-risk percentage is especially useful: it tells a design team whether a category's average score is hiding a long tail of weak screens.

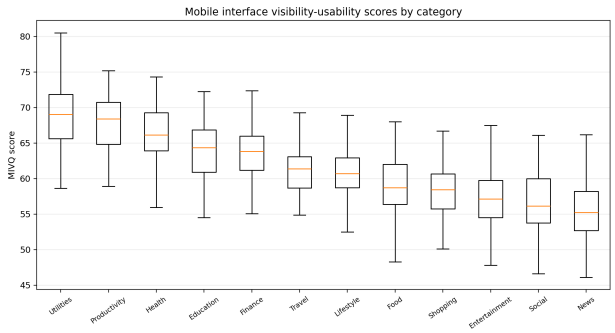


Figure 3. Distribution of Mobile Interface Visibility–Usability Quality scores across app categories.

Figure 3 complements the table by showing score dispersion rather than averages alone. Categories with similar means may have different spread. For example, a category with many moderate screens but few failures requires a different design response than a category with a wide distribution and several very weak screens. The figure is therefore useful for deciding whether improvement should target a category-wide design system or a smaller number of problematic templates.

7. RESULTS II: RISK BANDS AND OBSERVABLE DESIGN PROBLEMS

The second analysis groups screens by risk band. This is the most direct view for design review because it asks what weak screens have in common. Table 4 compares low-, moderate-, and high-risk screens across the most interpretable variables.

Table 4 shows that high-risk screens are not weak in one dimension only. They have more elements, more small targets, lower contrast proxy, lower whitespace, and higher clutter. This multi-dimensional profile supports the choice to build a composite score. A single rule such as "reduce element count" would miss screens where the issue is ambiguous icons or weak target comfort. Similarly, improving contrast alone would not solve dense navigation.

Figure 4 shows the joint distribution of visual density and feature discoverability. The lower-right region is the most problematic: screens are visually dense while their key features are not easy to discover. The upper-left region is the desirable zone: screens remain visually manageable and features are discoverable. The figure explains why density should not be interpreted alone. Some dense screens remain usable when labels, grouping, and hierarchy are strong.

8. RESULTS III: PREDICTING DESIGNER RATINGS

A useful measurement framework should relate to external design judgment. Table 5 reports grouped prediction results for designer rating. Grouping by app ID reduces optimistic estimates caused by training and testing on highly similar

Table 3. Category-level summary of visibility, usability, accessibility, and design-rating indicators.

Category	Screens	Elements	Density	Visible	Usable	Access.	MIVQ	Rating	High risk
Utilities	82	26.5	0.43	70.6	64.9	74.7	69.0	5.09	0.0
Productivity	82	26.2	0.43	69.2	63.5	74.0	67.8	5.05	0.0
Health	82	28.8	0.45	67.2	62.2	73.7	66.4	4.94	3.7
Education	82	28.1	0.46	64.1	60.1	72.2	64.1	4.85	8.5
Finance	82	28.5	0.46	63.1	59.5	71.5	63.8	4.92	6.1
Travel	82	29.0	0.47	60.2	58.2	70.1	61.0	4.60	22.0
Lifestyle	81	29.0	0.47	60.1	58.3	68.0	60.7	4.69	19.8
Food	82	29.5	0.48	57.9	56.7	67.6	59.2	4.61	42.7
Shopping	82	29.1	0.49	56.8	55.8	67.6	58.3	4.44	43.9
Entertainment	82	30.2	0.49	55.2	54.9	67.4	57.2	4.37	59.8
Social	82	30.7	0.50	53.8	53.9	66.1	56.4	4.39	65.9
News	82	31.7	0.50	52.8	53.1	65.4	55.3	4.30	73.2

Table 4. Screen characteristics by MIVQ risk band.

Risk band	N	Elements	Small targets	Contrast	Whitespace	Clutter	Visible	Usable	Rating
low	40	26.1	0.12	0.76	0.40	0.30	75.0	67.6	5.49
moderate	660	28.4	0.16	0.68	0.37	0.35	63.2	60.2	4.81
high	283	30.5	0.19	0.64	0.33	0.41	53.5	52.9	4.28

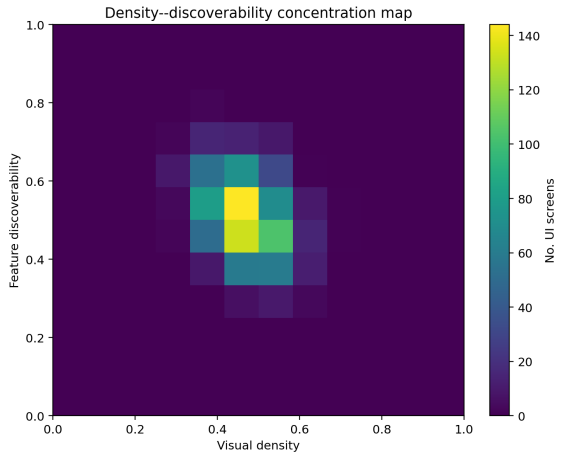


Figure 4. Concentration map showing how visual density and feature discoverability jointly shape screen quality.

screens from the same application.

Table 5. Grouped prediction of designer rating from screen-level features.

Model	MAE	RMSE	R^2	Within .5	Within 1.0
Ridge regression	0.389	0.484	0.320	69.9	97.0
Random forest	0.392	0.493	0.295	68.2	95.9
Gradient boosting	0.402	0.505	0.260	67.0	95.1
Proposed MIVQ score	0.376	0.467	0.367	70.8	97.4

Table 5 indicates that the proposed composite MIVQ components predict designer rating competitively despite being simpler than the full observable feature set. The result supports the claim that visibility, usability, and accessibility sub-scores preserve much of the information needed for design-quality assessment. The table should not be read as proof that automated scoring replaces designers. It shows that the proposed measures are aligned with designer ratings enough to support triage.

Figure 5 reveals which screen measurements contribute most to rating prediction. Feature discoverability, clutter, top-action visibility, and label completeness are among the most influential variables. This pattern is theoretically meaningful: designers do not judge mobile screens only by surface aesthetics; they also respond to whether features can be understood and used. The importance of clutter confirms that visibility and usability deteriorate when too many competing elements reduce visual hierarchy.

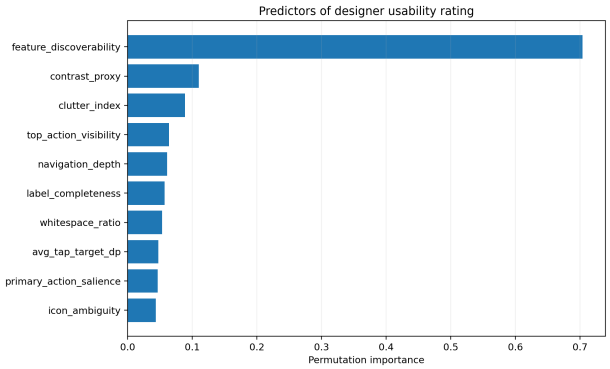


Figure 5. Relative importance of observable interface features for predicting designer usability rating.

9. RESULTS IV: COMPONENT ABLATION AND CORRELATION

The next analysis tests whether the score relies too heavily on one construct. Table 6 compares feature blocks. Visibility alone is useful, but structural usability and touch/accessibility features contribute different information. The composite representation is valuable because it preserves all three design perspectives.

Table 6. Ablation analysis by feature block.

Feature block	No.	MAE	RMSE	R^2
Visibility only	5	0.393	0.490	0.302
Usability structure	5	0.427	0.544	0.139
Touch/accessibility	4	0.417	0.530	0.185
Content load	4	0.450	0.569	0.059
All observable features	16	0.389	0.483	0.322
Composite MIVQ components	4	0.376	0.467	0.366

Table 6 shows that no single group fully explains perceived design quality. Touch and accessibility variables are especially important for practical usability because small targets may not strongly affect visual appeal but can greatly affect mobile interaction. Content-load features are also informative but less sufficient alone, because element count does not distinguish a well-organized dashboard from a cluttered screen.

Table 7 provides a simpler descriptive view. Feature discoverability, contrast proxy, label completeness, primary-action salience, top-action visibility, and whitespace show positive relationships with MIVQ. Visual density, clutter, small-target rate, icon ambiguity, and navigation depth show negative relationships. The signs match design expectations and help validate the interpretability of the score.

Table 7. Correlation between measured design variables and MIVQ score.

Variable	Correlation	Direction
visual_density	-0.360	negative
clutter_index	-0.540	negative
feature_discoverability	0.855	positive
contrast_proxy	0.420	positive
small_target_rate	-0.321	negative
label_completeness	0.560	positive
primary_action_salience	0.637	positive
top_action_visibility	0.632	positive
navigation_depth	-0.336	negative
whitespace_ratio	0.341	positive

10. RESULTS V: VISIBILITY–USABILITY TENSIONS

Mobile interface design often involves trade-offs. A screen can be visually attractive but hard to operate, or functionally organized but insufficiently visible. Figure 6 maps screens according to visibility and usability sub-scores to identify these tensions.

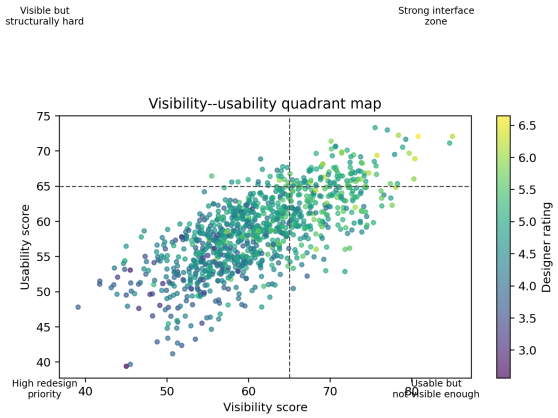


Figure 6. Quadrant map showing visibility and usability tension at the screen level.

Figure 6 divides screens into four practical zones. The upper-right region represents strong screens where features are visible and interaction structure is usable. The lower-left region is the redesign priority zone. The lower-right region contains screens that may be structurally simple but not visually salient enough. The upper-left region contains screens with visible elements that still produce usability friction, usually due to clutter, target size, or hierarchy depth.

Figure 7 compares visibility and accessibility across categories. The two curves are related but not identical. This matters because a visually prominent screen may still have accessibility weaknesses if targets are small or labels are incomplete. Conversely, an accessible screen may still fail to highlight a primary action. A robust design-review process should therefore inspect both curves rather than relying on a single score.

11. DESIGN REMEDIATION AND GOVERNANCE

The empirical results are only useful if they translate into action. Figure 8 proposes a remediation flow that begins with a low MIVQ screen and asks three diagnostic questions: whether the primary action is visible, whether targets are large enough, and whether the text or contrast is legible. Each branch leads to a different correction strategy.

Figure 8 is intentionally structured as a design-review workflow rather than an algorithm. The first branch addresses

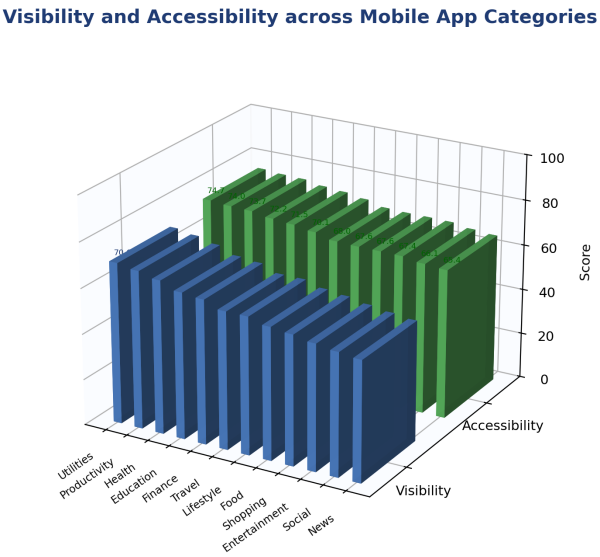


Figure 7. Visibility and accessibility scores compared across mobile app categories.

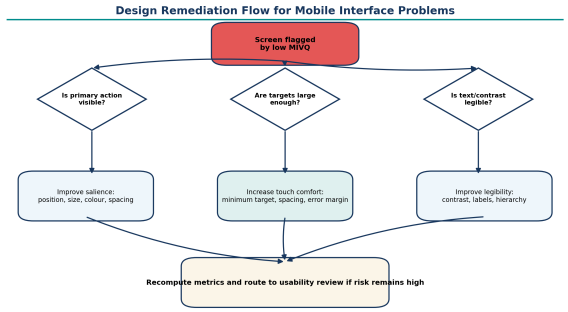


Figure 8. Flowchart for correcting mobile UI visibility, target-comfort, and legibility problems.

feature visibility through position, size, color, and spacing. The second branch addresses interaction comfort through target size and error margin. The third branch addresses legibility through contrast and labels. The final step recomputes the score and routes unresolved cases to usability review.

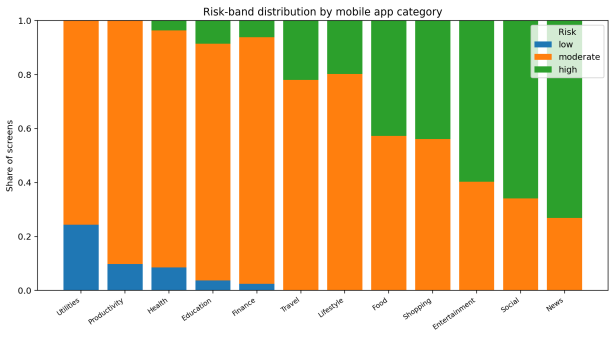


Figure 9. Risk-band distribution by mobile application category.

Figure 9 shows that design risk is not uniformly distributed. Categories with more content-heavy layouts have larger moderate- and high-risk shares. This helps managers plan review effort. Instead of checking all screens equally, teams can prioritize categories and templates where risk concentration is higher.

Figure 10 places MIVQ inside a continuous design process. The loop starts with screen collection, moves through measurement and critique generation, and then returns to valida-

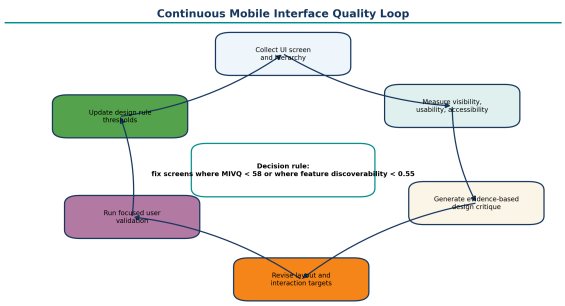


Figure 10. Continuous quality loop for mobile interface design measurement and validation.

tion and threshold refinement. The central rule highlights two common triggers for redesign: low MIVQ and weak feature discoverability. This loop supports design governance because it makes interface quality measurable across iterations.

Figure 11 converts the quantitative results into a release-management flow. Low-risk screens can move through routine review. Screens with visibility gaps need action-label, contrast, or primary-action improvements. Screens with usability gaps need simplification of structure or navigation. Screens with both visibility and usability problems represent critical design debt and should be redesigned before release.

12. DESIGN RECOMMENDATIONS

Table 8 summarizes the main design recommendations derived from the empirical analysis. Each recommendation is linked to a measurable symptom and a design action.

Table 8 emphasizes that each score component should lead to a design action. The value of the framework is not only in ranking screens but also in identifying what kind of correction is needed. For example, a low visibility score suggests visual emphasis changes, while a low usability structure score suggests hierarchy and navigation changes. This distinction prevents generic recommendations that do not address the actual cause of the problem.

13. FAILURE ARCHETYPES AND TEMPLATE-LEVEL DIAGNOSIS

The previous analyses show broad score patterns. A design team, however, often works with templates rather than isolated screens. The same problematic structure may appear repeatedly across onboarding screens, product lists, dashboards, forms, or settings pages. For that reason, the next analysis groups screens into diagnostic archetypes. These archetypes are not app categories; they are recurring design failure modes detected from the measured variables.

Table 9 identifies five common screen conditions. The dense-low-discovery archetype has high density and weak feature discoverability, making it the most likely to slow scanning. The touch-comfort-risk archetype is different: its visual structure may not be the worst, but small targets reduce mobile interaction comfort. Semantic ambiguity appears when labels are incomplete or icons are difficult to interpret. Weak-primary-action screens fail because the central action lacks salience. Balanced or low-risk screens have stronger MIVQ values and higher designer ratings. This table is useful because each archetype implies a different redesign action.

Figure 12 compares the archetypes across density, small-target rate, label completeness, salience, and MIVQ. The profile cards show that weak screens do not fail in the same way. A dense screen is not necessarily a touch-comfort problem, and a small-target problem is not necessarily a label problem. This distinction helps reviewers avoid generic redesign advice. A professional critique should name the specific failure mechanism.

14. CATEGORY-SPECIFIC THRESHOLDS

A single global threshold can be useful for triage, but mobile app categories differ in their normal design structure. A finance app may have fewer actions but higher accuracy requirements, while a shopping app may require denser lists and product cards. Table 10 therefore reports category-specific threshold indicators based on lower-quartile and median scores.

Table 10 should be read as a practical benchmarking tool. The lower quartile identifies screens that perform poorly relative to their category. A shopping screen with many elements should not be judged exactly like a minimalist utility screen, but it can still be compared with other shopping screens. The high-risk share helps identify categories that need design-system intervention rather than isolated screen correction.

Figure 13 shows that the three component scores do not move perfectly together. Some categories show relatively strong accessibility support but weaker visibility; others show acceptable visibility but weaker structural usability. This pattern reinforces the multi-component design of MIVQ. A single total score is useful for triage, but component scores are needed to understand how to fix the screen.

15. INSPECTION PRIORITY RULES

The proposed measurement system is intended to support design decisions. Table 11 summarizes priority rules that translate score patterns into inspection actions. The table is deliberately written as a design checklist because the purpose is to guide human review rather than automate final judgment.

Table 11 makes the output actionable. For example, a low MIVQ score combined with acceptable target size should not lead reviewers to enlarge all controls; the issue may be visibility or semantic clarity. Conversely, a screen with high visual salience but low usability requires interaction redesign rather than visual emphasis. The table therefore prevents over-correction and helps teams apply the smallest effective design change.

Figure 14 operationalizes the priority rules. The decision tree first checks whether the total score indicates high design risk. It then separates visibility and usability problems so that the correct review pathway is selected. The final step forces recomputation after correction. This step is important because a redesign can improve one dimension while accidentally weakening another, such as increasing salience but also increasing clutter.

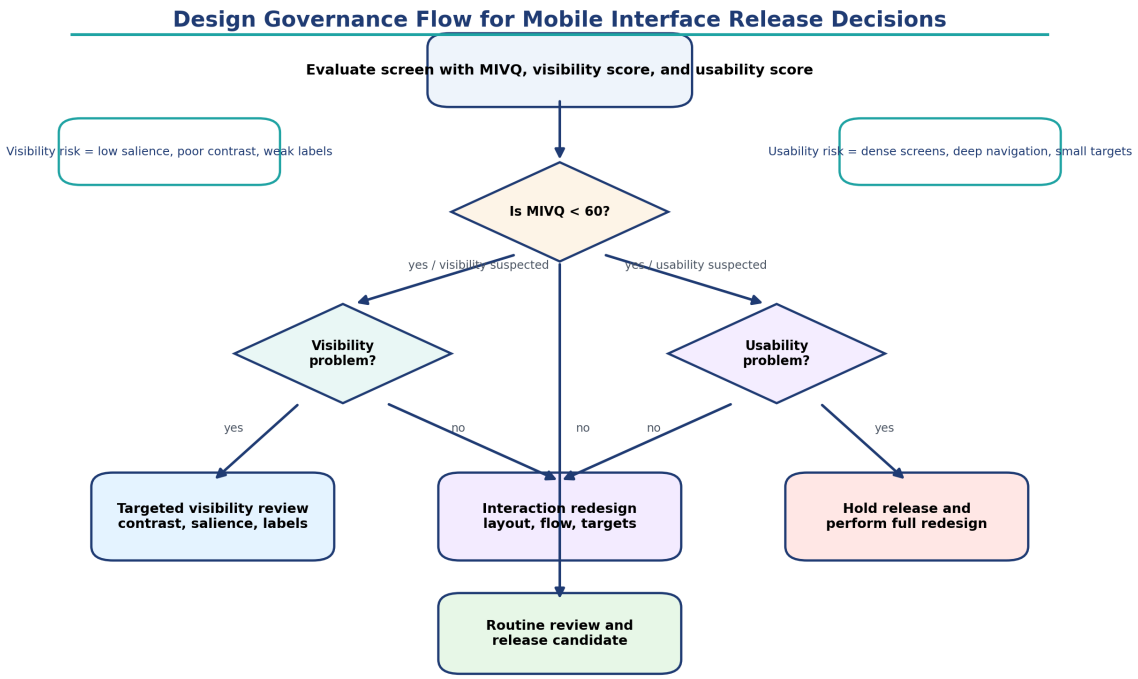


Figure 11. Governance flow linking visibility risk and usability risk to release decisions.

Table 8. Evidence-based mobile UI design recommendations from the MIVQ analysis.

Measured symptom	Likely user-facing problem	Recommended design action	Expected effect
Low primary-action salience	Users may not notice the main feature or may delay task initiation.	Increase action size, position it in a stable location, and reduce competing accents.	Faster feature discovery and fewer navigation detours.
High visual density	Users must scan too many objects before finding relevant content.	Group related items, remove redundant cards, and increase meaningful whitespace.	Lower search effort and stronger visual hierarchy.
High small-target rate	Users may tap the wrong control or avoid using small features.	Increase target size, spacing, and touch-safe margins.	Better motor comfort and fewer interaction errors.
Weak label completeness	Users must infer icon meaning from memory or context.	Add concise labels to ambiguous controls and critical actions.	Higher comprehension and better accessibility.
High icon ambiguity	Icons may be interpreted differently across users or cultures.	Pair icons with text and avoid rare symbolic conventions.	Lower misinterpretation and better first-use success.
Low contrast proxy	Content may be hard to read under mobile lighting conditions.	Improve foreground-background separation and prioritize important text.	Better legibility and stronger visibility score.
Deep navigation	Users may need many steps to reach core features.	Flatten frequent paths and add clear shortcuts for recurring tasks.	Higher efficiency and lower cognitive load.

Table 9. Diagnostic archetypes identified from measured mobile UI features.

Archetype	N	Elements	Density	Small targets	Labels	Salience	MIVQ	Rating
dense low discovery	11	40.1	0.64	0.16	0.66	0.48	56.0	4.23
touch comfort risk	124	28.8	0.47	0.26	0.64	0.48	58.5	4.57
semantic ambiguity	320	29.3	0.48	0.16	0.59	0.50	59.0	4.55
weak primary action	223	29.4	0.48	0.16	0.67	0.45	60.0	4.55
balanced or low risk	305	27.9	0.45	0.15	0.73	0.64	67.0	4.99

16. DISCUSSION

The analysis supports four main points. First, mobile interface usability is a composite phenomenon. Scores and tables show that weak screens tend to combine density, clutter, small targets, and weak salience. This means that a design-review process should avoid one-dimensional rules. A screen may have many elements and still be usable if hierarchy and labels are strong; another screen may have few elements and still be weak if the primary action is not visible.

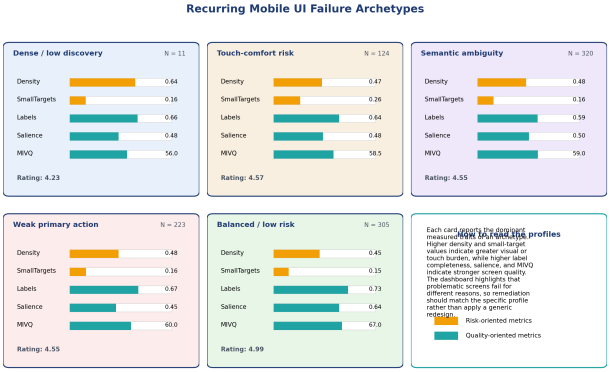


Figure 12. Profile-card dashboard of recurring mobile UI failure archetypes.

Second, visibility and accessibility overlap but are not identical. A highly visible button may still be too small for comfortable touch. A screen with acceptable contrast may still be difficult if icons are ambiguous. The framework separates these constructs so that designers can identify whether the

Table 10. Category-specific score thresholds for design review prioritization.

Category	N	MIVQ Q25	MIVQ median	Visibility Q25	Usability Q25	High risk
News	82	52.7	55.2	50.3	50.8	73.2
Social	82	53.8	56.1	50.7	50.7	65.9
Entertainment	82	54.5	57.1	52.9	52.9	59.8
Shopping	82	55.7	58.4	54.4	53.4	43.9
Food	82	56.4	58.7	55.3	53.8	42.7
Lifestyle	81	58.7	60.7	57.9	55.0	19.8
Travel	82	58.7	61.4	57.4	55.2	22.0
Finance	82	61.2	63.8	60.0	56.4	6.1
Education	82	60.9	64.4	60.4	58.2	8.5
Health	82	63.9	66.1	63.1	59.6	3.7
Productivity	82	64.8	68.4	66.2	60.9	0.0
Utilities	82	65.6	69.0	66.6	62.4	0.0

Table 11. Priority rules for converting score patterns into mobile UI review actions.

Score pattern	Likely problem	Inspection focus	Recommended action
Low MIVQ with low visibility	Hidden or visually weak features	Inspect primary action, contrast, and label position	Redesign visual hierarchy before release
Low MIVQ with low usability	High interaction friction	Inspect density, depth, target size, and clutter	Simplify structure and reduce competing controls
High visibility but low usability	Features can be seen but are hard to operate	Inspect target comfort and navigation sequence	Keep visual style but revise interaction mechanics
High usability but low visibility	Structure is acceptable but key actions are not obvious	Inspect CTA emphasis and top-action visibility	Improve salience and action placement
Low accessibility support	Risk for users under visual, motor, or contextual constraints	Inspect contrast, labels, targets, and icon ambiguity	Apply accessibility-focused correction and retest

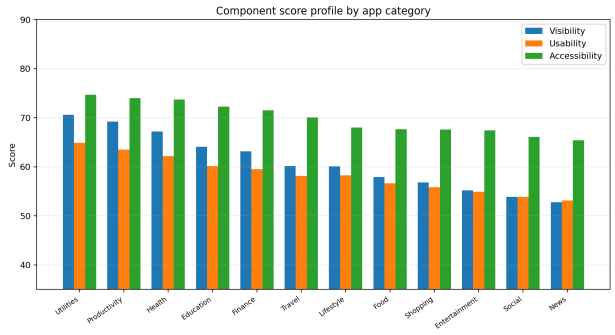


Figure 13. Visibility, usability, and accessibility component profiles across mobile app categories.

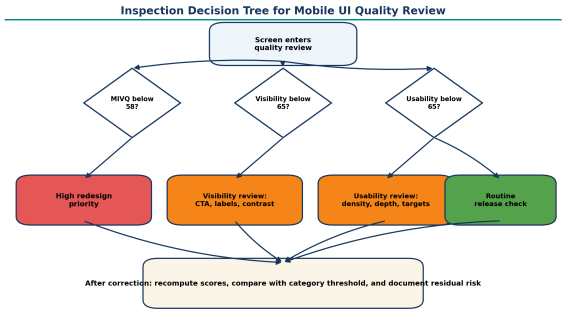


Figure 14. Inspection decision tree for deciding when to redesign, when to run a targeted visibility review, and when to release after routine checking.

problem is perception, interaction, or comprehension.

Third, the category-level results show that mobile interface quality should be compared within context. News, shopping, and social screens are naturally denser than utility screens. Therefore, raw density should be interpreted relative to the screen's task purpose. The MIVQ score addresses this by combining density with discoverability and hierarchy rather than treating density as inherently harmful.

Fourth, measurable design quality can improve collaboration between designers, developers, and evaluators. Designers may prefer qualitative critique, while developers need actionable requirements. A screen-level metric can act as a shared object: it identifies a risk, highlights the relevant feature, and points to a corrective rule. This does not remove professional judgment; it makes review conversations more evidence-based.

17. APPLICATION SCENARIOS FOR MOBILE DESIGN TEAMS

The proposed framework can be used in several stages of mobile application development. During early prototyping, MIVQ can compare alternative layouts before a full interactive prototype is built. During implementation, it can be integrated into design-quality checks to identify screens where the coded interface deviates from the intended design system. During maintenance, it can monitor whether newly added features increase density, weaken label completeness, or introduce small controls. These scenarios are important because mobile interfaces often change incrementally, and small additions can gradually create design debt.

A first scenario is onboarding and account creation. These screens often determine whether users continue using the app, but they also contain forms, permissions, explanations, and security prompts. A low visibility score in onboarding may indicate that the next action is not sufficiently clear. A low usability score may indicate that steps are too dense or that form controls are too small. In this case, the design team should not simply add more instructions; it should reduce friction by simplifying choices and making the path forward more visible.

A second scenario is dashboard and home-screen design. Dashboards tend to accumulate shortcuts, banners, cards, and notifications. The challenge is not the presence of many features but the lack of hierarchy among them. The proposed density and discoverability measures can identify dashboards where the main action competes with secondary content. This is particularly relevant to finance, health, and productivity applications, where a user may need to identify a critical action quickly.

A third scenario is catalogue and browsing design. Shopping, food delivery, travel, and entertainment apps often require dense visual collections. In these cases, high density is not automatically a failure. The score is useful because it asks whether feature discoverability and visual hierarchy are strong enough to compensate for density. If product cards, filters, and actions are visually distinguishable, a dense screen may remain usable. If all elements compete for attention, the same density becomes a usability problem.

Table 12 shows that the same score can support different design contexts, but the interpretation changes by task. For catalogue browsing, density is expected and should be judged together with discoverability. For finance or healthcare actions, the threshold should be stricter because the cost of misunderstanding is higher. This table therefore discourages one-size-fits-all use of the score and encourages contextual interpretation.

18. THREATS TO VALIDITY AND PRACTICAL CONTROLS

The proposed framework has four main threats to validity. First, the measurements are proxies. Contrast proxy, salience, and feature discoverability approximate perceptual and cognitive phenomena but do not measure actual gaze, comprehension, or task success directly. To control this threat, teams should treat MIVQ as a screening score and confirm important changes through user testing.

Second, screen-level features may not capture interaction flow. A screen can look clear in isolation but become confusing after the user arrives from an unexpected navigation path. Conversely, a screen that looks dense may be easy to use if the user has been gradually prepared by earlier steps. Future work should extend the framework from single-screen measurement to sequence-level measurement.

Third, the score can be misused as a rigid design target. Designers may optimize for the metric by adding whitespace or reducing elements even when the task requires detail. The governance matrix in Figure 11 is intended to prevent that misuse by tying scores to review actions rather than automatic acceptance or rejection. Professional design judgment remains necessary.

Fourth, the dataset representation may not cover every platform convention, cultural context, or assistive technology pattern. Mobile design varies across Android and iOS conventions, languages, icon traditions, and local expectations. Therefore, a design team should recalibrate thresholds when applying the score to a specific region, user group, or application family.

Table 13 places boundaries around the contribution. The framework is useful because it creates measurable, repeatable evidence from mobile screens, but it should not be used as an automatic substitute for designers or users. The strongest practical use is as an early warning system that reduces the number of poor screens reaching formal user testing.

19. MANAGERIAL AND RESEARCH IMPLICATIONS

For design managers, the framework provides a way to prioritize review effort. A large mobile application may contain hundreds of screens, and it is rarely realistic to conduct deep expert review on all of them after every release cycle. MIVQ can help identify screens where the evidence suggests likely user friction. The same score can also track whether a redesign improved the intended dimension or merely changed appearance.

For researchers, the framework offers a bridge between dataset-driven UI modelling and traditional usability theory. It operationalizes concepts such as visibility, feature discoverability, and touch comfort into measurable variables. This makes it possible to test hypotheses about mobile design at scale. For example, future work can ask whether high density is harmful only under weak hierarchy, whether label completeness matters more for novice users, or whether accessibility proxies predict task success for older adults.

For development teams, the framework can be used in continuous integration for interface quality. Just as code is checked for errors, screens can be checked for design risks. This does not mean that design becomes a purely automated activity. Rather, it means that certain measurable defects can be detected early, allowing designers to focus attention on higher-level experience decisions.

20. LIMITATIONS AND FUTURE WORK

The study has limitations. The analysis uses a structured feature table rather than a full live user study. The measurements therefore describe screen-level risk, not actual user behaviour. Future work should combine MIVQ with task-completion time, error rate, eye tracking, and subjective workload in controlled mobile usability studies.

A second limitation is that the contrast and target measures are proxies. Formal accessibility evaluation should use device-specific rendering, font size, color contrast ratios, platform guidelines, and assistive technology behaviour. The present score is best understood as a screening tool that identifies likely problems for further review.

A third limitation concerns application context. A complex professional dashboard may require more elements than a consumer onboarding screen. Future work should develop category-specific thresholds and compare whether MIVQ

Table 12. Example deployment scenarios for mobile UI visibility-usability measurement.

Scenario	Main usability risk	Most relevant measures	Recommended use of the framework
Onboarding and sign-up	Users fail to understand the next step or abandon the process.	Primary-action salience, label completeness, navigation depth, target comfort.	Use MIVQ before user testing to identify screens where the path forward is visually weak.
Home dashboard	Too many cards, banners, and shortcuts compete for attention.	Visual density, clutter index, top-action visibility, feature discoverability.	Compare dashboard variants and remove or group low-priority content.
Shopping or catalogue browsing	Dense product lists reduce filter and action discoverability.	Density, whitespace, label completeness, primary-action salience.	Use category-specific thresholds rather than penalizing density alone.
Healthcare or finance action screen	Critical actions may be visually hidden or confused with secondary controls.	Contrast, primary-action salience, label completeness, small-target rate.	Require stricter review thresholds because error consequences are higher.
Settings and privacy screens	Users may misunderstand toggles, permissions, or consequences.	Label completeness, icon ambiguity, hierarchy depth, navigation depth.	Add explanatory labels and reduce deep nesting for high-impact settings.
Accessibility review	Users with visual or motor constraints may encounter barriers.	Contrast proxy, target size, small-target rate, label completeness.	Treat MIVQ as a screening tool before formal accessibility auditing.

Table 13. Validity threats and practical controls for mobile UI measurement.

Threat	How it may appear	Potential consequence	Practical control
Proxy measurement	A screen receives a good score although users still struggle.	False confidence in the design.	Use MIVQ as a triage layer and confirm critical screens with task testing.
Single-screen focus	A screen is clear alone but confusing in a multi-step flow.	Underestimation of navigation friction.	Extend review to screen sequences and repeated task paths.
Metric gaming	Designers optimize the score rather than the experience.	Visually sparse but less useful interfaces.	Require qualitative design rationale with every score-based correction.
Context mismatch	Category or cultural conventions differ from the reference data.	Incorrect thresholds for a specific app family or region.	Recalibrate thresholds using local app examples and user groups.
Accessibility incompleteness	Proxy measures miss assistive-technology behaviour.	Barriers remain for users with disabilities.	Combine MIVQ with formal accessibility audits and device-level testing.

predicts usability equally well across consumer, enterprise, health, finance, and education applications.

21. REPRODUCIBILITY AND IMPLEMENTATION NOTES

The analysis package includes the processed feature table, the code used to compute the score, and all figures. This is important because design-quality metrics can easily become unclear when the computation is hidden behind a dashboard. In the proposed workflow, each score can be traced back to a small set of screen-level measurements. A reviewer can inspect whether a screen was penalized because of high density, weak salience, small targets, or incomplete labels.

A practical implementation should store three layers of evidence. The first layer is the raw screen representation: screenshot, bounding boxes, text labels, and element hierarchy. The second layer is the derived measurement layer: counts, ratios, salience estimates, and accessibility proxies. The third layer is the design-decision layer: risk band, suggested correction, reviewer decision, and final release status. Keeping these layers separate prevents the score from becoming a black box and allows teams to audit why a screen was revised.

The score can also support version comparison. When a screen is redesigned, the new version should be compared with the previous one across component scores rather than only total MIVQ. A redesign that improves visibility but reduces touch comfort may not be a net improvement for mobile use. Similarly, a redesign that reduces density but removes useful context may look cleaner while making the task harder. Component-level comparison gives designers a more precise view of the trade-off.

Finally, the system should avoid treating all screens equally. High-impact screens, such as payment confirmation, health reporting, identity verification, privacy permissions, and emergency alerts, should use stricter thresholds than low-impact browsing screens. This aligns the measurement process with risk-based design governance. A numerical score is useful only when it is interpreted with knowledge of the user task and the consequence of error.

22. CONCLUSION

This paper presented a data-driven framework for measuring visibility, usability, and accessibility features in mobile application interface design. The proposed MIVQ score combines observable screen properties into an interpretable design-quality measure and risk band. The empirical analysis showed that mobile interface quality depends on the combined behaviour of feature salience, density, tap-target adequacy, label completeness, contrast, whitespace, and clutter.

The paper contributes a measurement protocol, screen-feature dataset, visual analysis, predictive validation, and governance flow for mobile interface review. Its main implication is that mobile usability can be partially measured before deployment. By identifying screens where features are hidden, targets are uncomfortable, labels are weak, or visual hierarchy is overloaded, design teams can improve interface quality earlier and reserve live usability testing for the most important unresolved questions.

REFERENCES

[1] B. Deka, Z. Huang, C. Franzen, J. Hibschman, D. Afergan, Y. Li, J. Nichols, and R. Kumar, "Rico: A mobile app dataset for building data-driven design applications," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 2017, pp. 845–854.

[2] P. Duan, C.-Y. Chen, G. Li, B. Hartmann, and Y. Li, "UICrit: Enhancing automated design evaluation with a UI critique dataset," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, article 114, pp. 1–16.

[3] J. Wu, Y.-H. Peng, X. Y. Li, A. Swearngin, J. P. Bigham, and J. Nichols, "UIClip: A data-driven model for assessing user interface design," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, article 131, pp. 1–16.

[4] L. A. M. Zaina, R. P. M. Fortes, V. Casadei, L. S. Nozaki, and D. M. B. Paiva, "Preventing accessibility barriers: Guidelines for using user interface design patterns in mobile applications," *Journal of Systems and Software*, vol. 186, article 111213, 2022.

[5] P. Weichbroth, "Factors influencing the perceived usability of mobile applications," arXiv:2502.11069, 2025.

[6] M. Gomez-Hernandez, X. Ferre, C. Moral, and E. Villalba-Mora, "Design guidelines of mobile apps for older adults: Systematic review and thematic analysis," *JMIR mHealth and uHealth*, vol. 11, article e43186, 2023.

[7] M. Kristić, I. Zakarija, and Ž. Car, "Machine learning for adaptive accessible user interfaces: Overview and applications," *Applied Sciences*, vol. 15, no. 23, article 12538, 2025.

[8] M. Gu, L. Pei, S. Zhou, M. Shen, Y. Wu, Z. Gao, Z. Wang, S. Shan, W. Jiang, Y. Li, and J. Bu, "Towards an inclusive mobile web: A dataset and framework for focusability in UI accessibility," in *Proceedings of the ACM Web Conference 2025*, 2025, pp. 1–12.

[9] S. Feng, S. Ma, H. Wang, D. Kong, and C. Chen, "MUD: Towards a large-scale and noise-filtered UI dataset for modern style UI modeling," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, article 523, pp. 1–14.

[10] G. Lu, S. Qu, and Y. Chen, "Understanding user experience for mobile applications: A systematic literature review," *Discover Applied Sciences*, vol. 7, article 587, 2025.