

Evaluating the Quality of UX Design



Structured, evidence-based appraisal of an interface before release



Reduces risk and aligns product, design, research, and engineering



Distinct from usability testing, functional QA, and accessibility compliance



Human-centered quality: usability, accessibility, UX, and avoidance of harm



Key terms: heuristic, validity, reliability, severity, benchmark, evaluator effect



Failure modes: subjective judgment, late timing, single-method reliance



Why Pre-Release UX Evaluation Pays Off



Early validation:

~25% fewer iteration cycles,
millions saved in rework



Broken checkout and form
flows directly cost revenue
and drive abandonment



Shared evidence settles
"what users might do"
debates faster than opinion



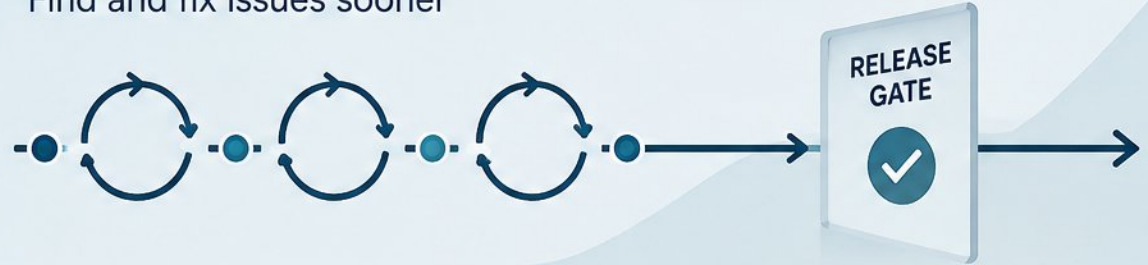
Evaluation runs continuously:
requirements, design,
prototype, build, release,
monitoring



Cheapest fixes happen
in discovery — not in a
single pre-release check

Early Evaluation Path

Find and fix issues sooner



Late Check Path

Issues found later mean costly rework



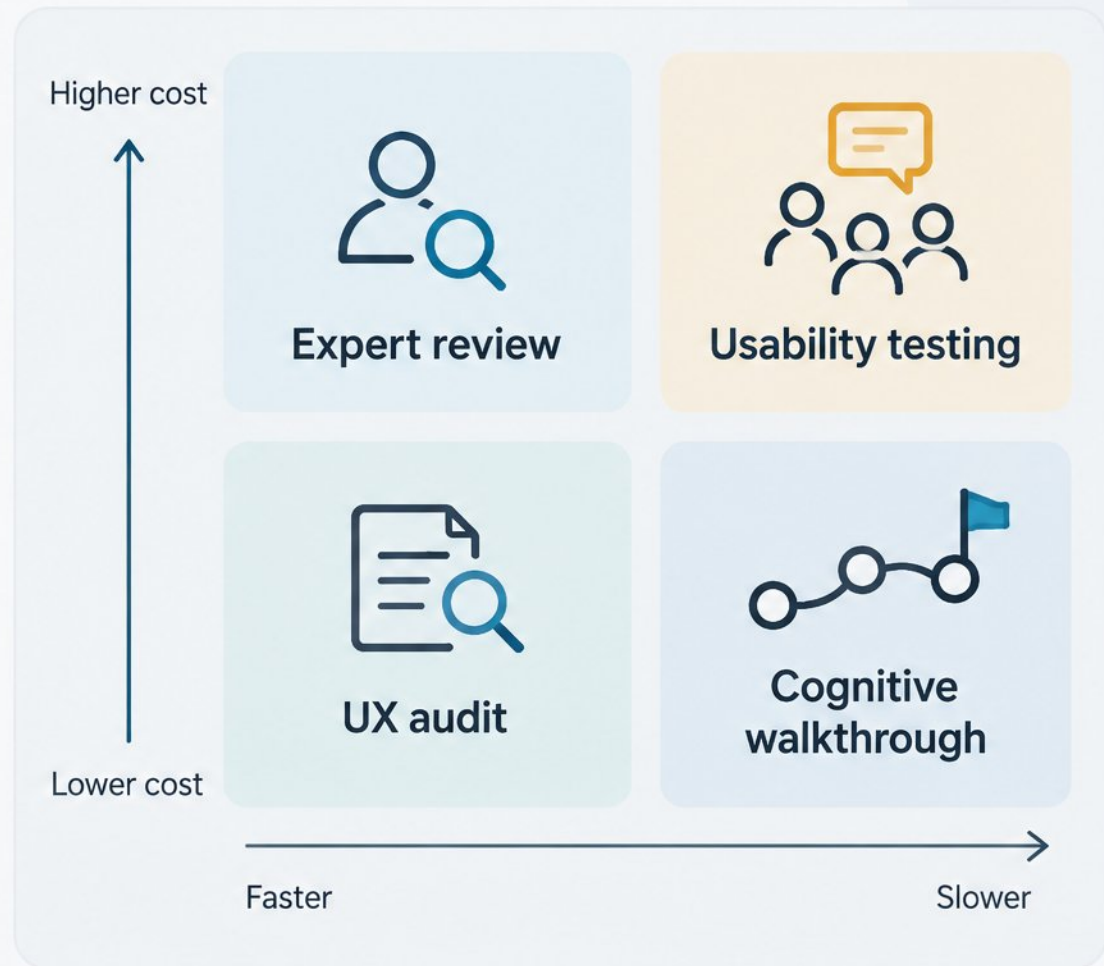
Quality Dimensions and Criteria

- Multi-dimensional model: usability, accessibility, consistency, IA, visual hierarchy, trust, error recovery, performance
- Anchor dimensions to recognized frameworks: ISO 9241-210, ISO 9241-220, Nielsen's heuristics, WCAG 2.2
- Turn abstract dimensions into observable review questions and acceptance criteria
- Weight dimensions by product context, user risk, and release stage
- Document the criteria matrix so coverage gaps stay visible
- WCAG 2.2: four principles — perceivable, operable, understandable, robust



Methods: Expert Review, Testing, and Audits

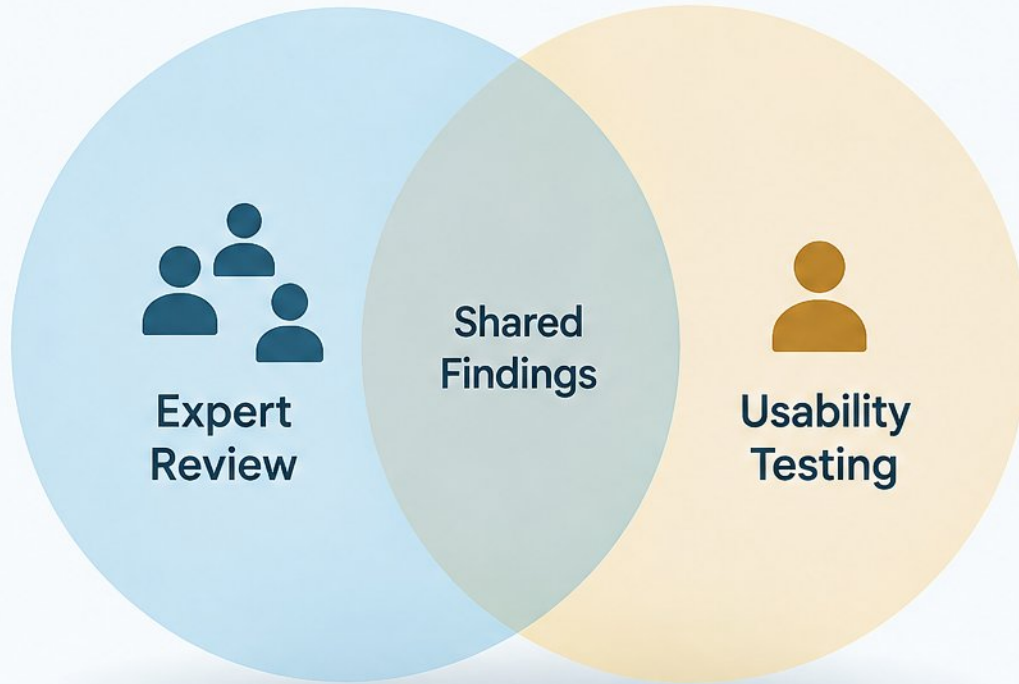
- **Expert review:** heuristic evaluation, cognitive walkthrough, UX audit — fast, no users needed
- Heuristic evaluation is broad; cognitive walkthrough traces one task step by step
- **Usability testing** is the only method observing real behavior; finds severe, recurring issues
- **Match method to question:** audit for site-wide problems, testing for drop-off causes
- **Formative** small-sample work finds problems; **summative** larger samples measure performance



Triangulate methods

Get a fuller, more reliable understanding.

Evidence Strength and Triangulation



Expert review and usability testing each miss about half of the other's findings



One heuristic evaluation is the weakest method; use three to five expert evaluators



Heuristics catch skill and rule-based errors; user testing catches knowledge failures



Record whether each finding is a prediction or an observation



Sequence: expert review first, usability testing to validate, audits as baseline



Expert Review



Usability Testing



Audits

Evaluating Accessibility and Inclusion

- Accessibility is essential UX quality, not an optional add-on
- WCAG 2.2: W3C standard defining testable success criteria
- Four principles: perceivable, operable, understandable, robust
- Conformance means meeting success criteria at A, AA, or AAA — not techniques
- Automated tools fully cover only ~22% of A/AA criteria
- About 55% of criteria require human evaluation
- Scanners check alt text exists — not accuracy, focus order, or clarity
- Inclusive design addresses cognitive load, language, motor demands
- Testing with users with disabilities is the strongest evaluation

WCAG 2.2

Four principles for accessible experiences



perceivable



operable



understandable



robust

Evaluation Coverage of A/AA Criteria

22%

24%

55%

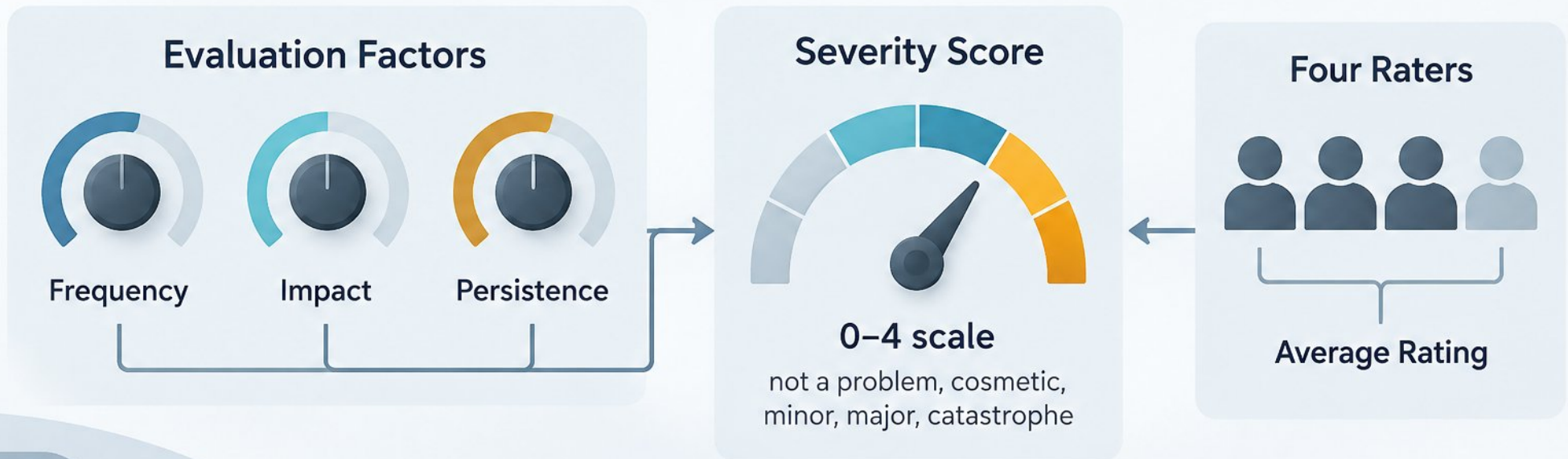
Fully automated
by tools

Partially covered
by tools

Require human
evaluation

Severity, Prioritization, and Scoring

- Nielsen's model scores frequency, impact, and persistence together
- 0–4 scale: not a problem, cosmetic, minor, major, catastrophe
- Single rater correlations average only about .24 — use four raters
- Report frequency and severity separately; they are largely independent
- Calibrate ratings, check distributions, map to impact x effort



Metrics and Benchmarks

- Behavioral metrics show what users do; attitudinal metrics explain why
- Key benchmarks: 78% task success, SUS 68, SEQ 5.5/7, 0.7 errors per task
- Common reference points, not universal pass marks — report raw scores and context
- SUS is not a percentage; report raw score, percentile, and grade separately
- UMUX-Lite and UX-Lite estimate SUS (.81–.83) but are separate instruments with exact wording
- Sample size: 5–8 for formative studies, 20–40 for quantitative benchmarks
- Report confidence intervals, never bare numbers

Task Success



Behavioral
Metric

Errors per Task



Behavioral
Metric

SUS



Attitudinal
Metric

SEQ

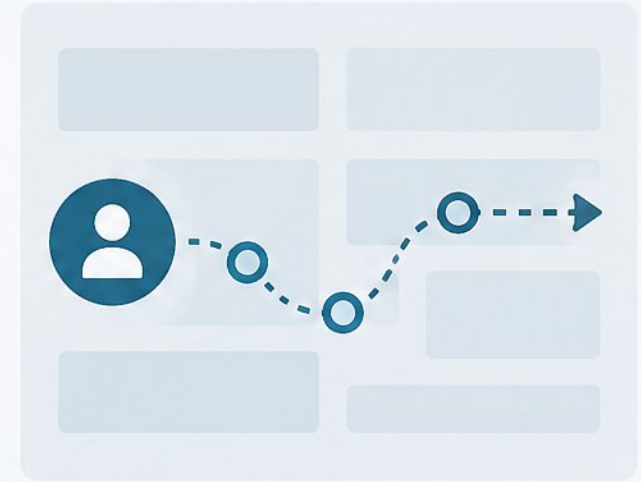
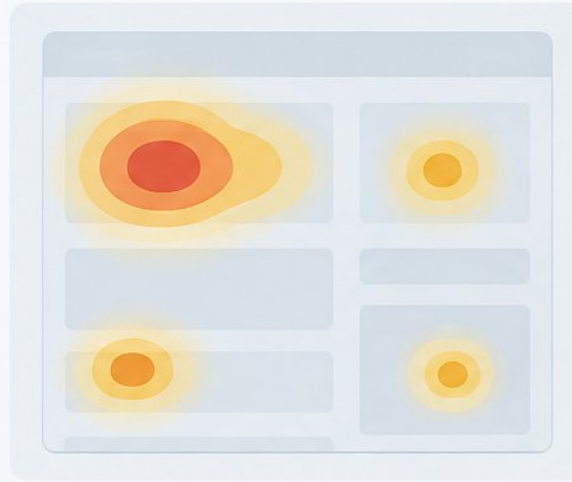
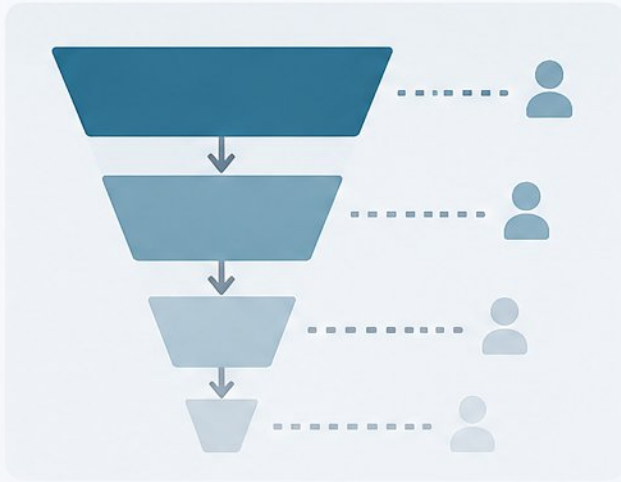


Attitudinal
Metric



UMUX-Lite and UX-Lite estimate SUS (.81–.83) but are separate instruments with exact wording

Behavioral Data: Analytics, Heatmaps, and Replays



Analytics and funnels: how many, and where users drop off



Heatmaps: what a crowd does on one page, not sequence



Session replay: what one specific user saw and did



Traps: sampling bias, scroll thresholds, dead clicks, dynamic content



Privacy: verify masking, consent, retention; report aggregated findings

Reviewer Calibration and Bias Control

- 1 Evaluator agreement across methods ranges from about 5% to 65% — manage, don't eliminate
- 2 Watch confirmation, anchoring, halo, recency, expertise, and novelty bias
- 3 Labeling identical prototypes 'new' can shift preference by up to 77%
- 4 Calculate and report inter-rater reliability instead of assuming consensus
- 5 Rubrics and codebooks define severity levels with worked examples
- 6 Calibrate: rate independently, compare, discuss gaps of two points or more, converge

Two Evaluators' Ratings



Common Biases to Watch



Opinion

Observation

Documenting Findings and Reporting

- 1 Structure reports by how decisions get made: summary, scope, severity-ordered findings, recommendations
- 2 Each finding stands alone: problem-based title, marked screenshot, observation, impact, one recommendation
- 3 Separate observation, interpretation, and recommendation; describe the design element, not blame
- 4 Rate severity with three independent raters and use the mean; gaps of two points signal vague findings
- 5 Include what worked, a backlog table with owners, method limits, and a retest plan

Report Outline

- Summary
- Scope
- Findings (severity-ordered)
- Recommendations

Finding Example

Problem-based finding title



Observation

Impact

Recommendation

Evidence Spectrum

Focus on what you can see and verify.



Base findings on observable evidence, not assumptions.

Embedding Evaluation into Release Workflows



Make design QA a gate: block only on visual regressions, WCAG-AA failures, token drift



Right-size effort: quick heuristics for small changes, deep testing for major releases



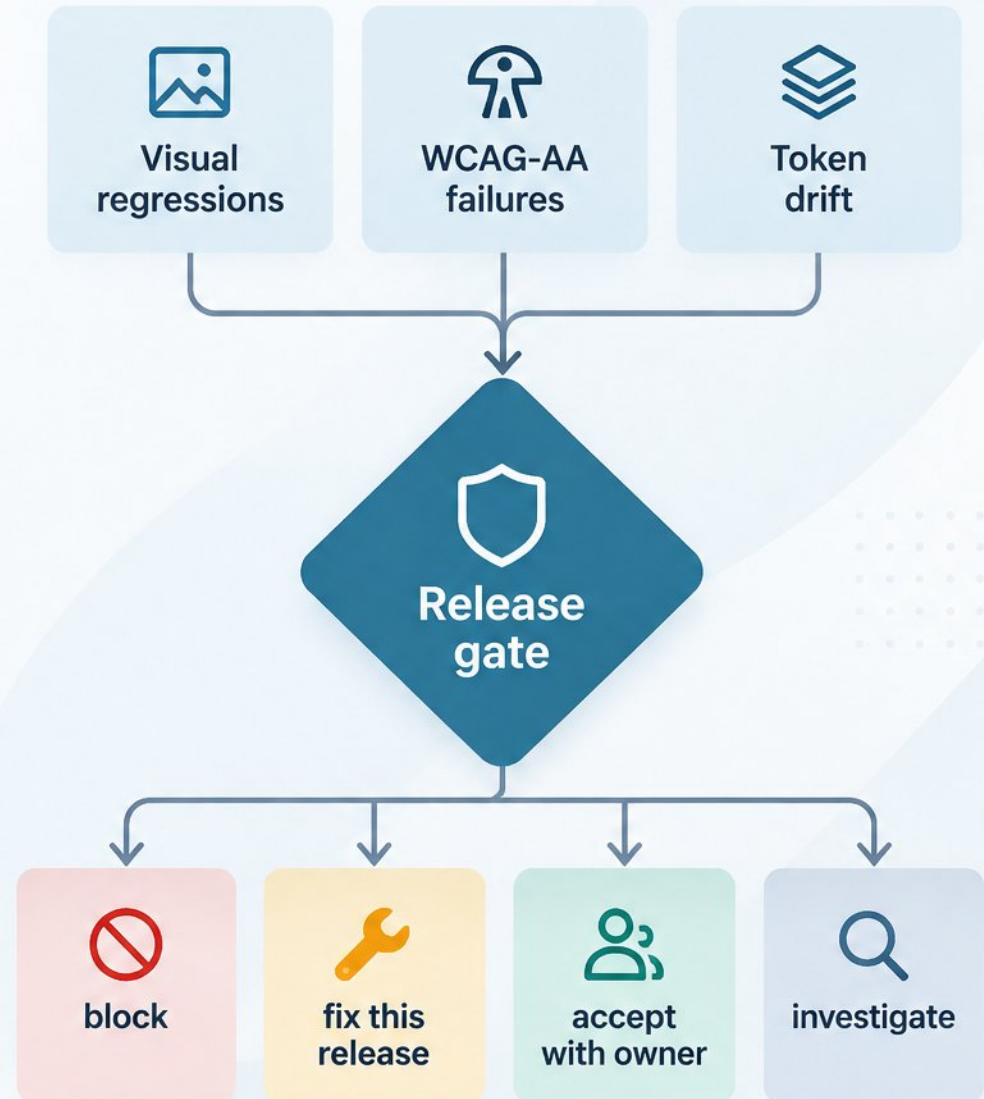
Run discovery one sprint ahead so findings land backlog-ready before commitment



Define readiness dispositions: block, fix this release, accept with owner, investigate



Post-release: verify the fix against the original reproduction, feed learnings to the design system



Putting It Together: A Repeatable Evaluation Practice



- The loop: criteria → inspect → validate → document → fix → monitor
- Choose methods by question, not by habit
- Combine heuristics, walkthroughs, usability testing, and periodic audits
- Guard against single-method reviews and siloed accessibility
- Integrated usability, UX, and accessibility evaluation improves release confidence



Better evaluation. More inclusive experiences. Greater confidence.

PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/b68a1778/public