RAPID SYSTEMATIC EVIDENCE REVIEW

Automated Accessibility Scores Are Not Accessibility Evidence

What automated web testing can detect, what it misses, and how organizations should interpret the results

Kevin Hintzman

Founder, Wingspan Labs | Published by Wingspan LLC

WINGSPAN LABS RESEARCH BRIEF

July 20, 2026 | AUTOMATED-ACCESSIBILITY-EVIDENCE-REVIEW-20260720-R1

Abstract

Automated accessibility tools are useful. They can scan quickly, repeat checks consistently, and identify many code-detectable issues early. The problem begins when a score is treated as a conclusion rather than a measurement of one tool's configured rules.

This rapid systematic evidence review examined empirical research comparing automated web accessibility evaluation with expert review, manual inspection, known-error benchmarks, cross-tool comparison, and user-centered evidence. Twenty primary studies and one secondary systematic mapping were included in the synthesis. A statistical meta-analysis was considered but rejected because the studies measure different constructs with incompatible denominators and generally do not report poolable uncertainty data.

The recurring result is not that automation lacks value. It is that automation provides partial, tool-dependent evidence. Different tools cover different rules and can disagree on the same page. Manual judgment changes findings. Task-based studies with disabled users reveal barriers, workarounds, and experience effects that scanner outputs cannot establish. The defensible practice is therefore layered: automated checks, experienced manual evaluation, and user-centered testing where the decision requires it.

Executive Summary

A score answers a narrower question than it appears to answer. It reports what a particular tool, version, ruleset, configuration, page sample, and moment in time detected. It does not directly measure whether every visitor can perceive, understand, navigate, and operate the experience.

The evidence does not support one universal automation percentage. Published figures use different denominators: WCAG criteria covered, expert-confirmed violations detected, tool agreement, user problems, or usability outcomes. Pooling them would create a number with no coherent meaning.

Human review is not a ceremonial add-on. Studies show that expert judgment, keyboard and screen-reader testing, and disabled-user task evidence surface issues or consequences not captured by automated checks. Human expertise also affects the quality and reproducibility of manual evaluation.

The right conclusion is a better evidence stack, not less automation. Automated checks should be used continuously for what they can test. Their reports should be paired with explicit scope, unresolved manual checks, experienced review, and user evidence appropriate to the risk and audience.

> **Bottom line: An automated score is evidence about an automated test. It is not, by itself, evidence that a website is accessible.**

1. Why the Score Is So Persuasive

Accessibility scores are attractive because they compress complexity. A red, yellow, or green result looks objective. A number is easy to compare, easy to place in a dashboard, and easy to report to leadership.

But compression changes the question. A scanner must translate accessibility requirements into executable rules. Some requirements are readily testable in code. Others require context, judgment, interaction, or knowledge of whether a person can complete a meaningful task. The score usually reflects only the portion implemented by that tool.

W3C's current guidance is explicit: evaluation tools can quickly identify potential issues and support manual review, but they cannot automatically check every accessibility aspect. Human judgment is required; tools can produce false or misleading results; and tools cannot determine accessibility on their own (W3C Web Accessibility Initiative, n.d.-a).

The title of this paper is intentionally blunt. It does not mean automated findings are worthless. It means the number alone is not sufficient evidence for the larger claim people often attach to it.

Figure 1. Three layers of accessibility evidence.

Original Wingspan Labs synthesis based on the reviewed evidence.

2. What This Review Asked

The review asked three practical questions:

1. What do automated accessibility tools detect reliably enough to be useful?
2. What do they miss, disagree about, or misrepresent relative to expert/manual and user-centered evidence?
3. What should a responsible organization require before it interprets a tool result as evidence of accessibility?

The review searched OpenAlex and Crossref on July 20, 2026, supplemented by backward reference checking and targeted DOI verification. The API searches returned 300 records, representing 278 unique

records after deduplication. A deterministic relevance screen produced 49 candidates; nine additional records were added through targeted and backward searching. The final synthesis includes 20 primary studies and one secondary systematic mapping; 37 screened records were excluded with reasons. The protocol, search log, screening log, and study-level extraction are included with the publication package. This is a rapid review conducted by one evidence reviewer; it is transparent but not exhaustive.

3. What the Evidence Shows

3.1 Automated tools detect important problems - and only a subset

Automated checks are well suited to repeatable rules: missing attributes, certain markup defects, some contrast conditions, and other machine-detectable patterns. They are valuable in development pipelines precisely because they can rerun those checks consistently.

The empirical studies also show a hard ceiling on what a particular tool run can establish.

- Vigo, Brown, and Conway (2013) benchmarked six tools against expert-derived ground truth. At most half of the WCAG 2.0 success criteria were covered. Completeness ranged from 14% to 38%, and tools that found more violations could trade that gain for lower correctness through false positives.
- Abduganiev (2017) compared eight tools across 52 webpages and manually re-evaluated the sample. Maximum criterion coverage was 32.4%; completeness ranged from 10% to 59%; average correctness was 70.7%; and agreement among tools was low.
- Fischer, Lundell, and Gamalielsson (2025) tested six open-source engines on 550 Swedish public-sector landing pages. The engines collectively covered about one sixth of WCAG success criteria, and the authors concluded that limited coverage prevents exclusive reliance on automated tools for conformance assessment.

These figures should not be averaged. They use different tool generations, page samples, benchmarks, and denominators. Their shared implication is narrower and more durable: scanner coverage is bounded and tool dependent.

3.2 The tool you choose changes the result

If a score were a stable measurement of overall accessibility, tools assessing the same page should agree reasonably well. They often do not.

Centeno and colleagues (2006) explained that tools can interpret non-formalized guidance differently and combine automated, semi-automated, and manual checks in different ways. More recent empirical work confirms that this remains a live issue.

- Ismailova and Inal (2022) ran six tools on 41 government websites and found that the tools generated different evaluation data for the same sites. Their practical conclusion was that some tools are complementary rather than interchangeable.
- Kelsey-Adkins and Thompson (2022) compared four command-line accessibility tools and reported very poor inter-rater reliability.

- Alsaeedi (2020) found different error sets from WAVE and Siteimprove on six university homepages. Its proposed comparison metric used the union of tool-reported issues as the reference set - useful for comparing outputs, but not an independent human ground truth.
- Fischer and colleagues (2025) traced differences not only to coverage but also to interpretation, documentation inconsistency, and ordinary software implementation defects.

A single score therefore needs provenance: tool name, exact version, ruleset, configuration, pages scanned, crawl conditions, and date. Without those facts, the number cannot be reproduced or meaningfully compared.

3.3 Human judgment changes the finding

Some tools explicitly label findings as potential issues because the code alone cannot decide them. Two large case studies show that resolving those issues manually changes the evaluation result.

Tzimas and Katsanos evaluated 441 university department websites in an earlier study and 112 Greek public-hospital websites in a later study. In both, manual inspection of tool-flagged potential issues materially affected guideline-based results (Tzimas & Katsanos, 2021, 2024).

Manual evaluation is not automatically perfect. Brajnik (2008) found differences in correctness and reliability between conformance testing and a barrier walkthrough among 12 novice evaluators. Brajnik, Yesilada, and Harper (2011) later showed that evaluator expertise matters: 19 experts produced more valid, reliable, consistent, and efficient barrier-walkthrough results than 57 nonexperts. In that experiment, three experts were enough to find all benchmark problems; 14 nonexperts were needed to reach the same level.

The practical implication is not simply "have a human look at it." It is to define the manual method, use experienced reviewers, document pages and tasks, and preserve unresolved uncertainty.

3.4 User experience and rule conformance overlap - but they are not the same construct

Task-based user evidence answers a different question: can people use the experience, in context, with their technologies, strategies, and goals?

- Power and colleagues (2012) observed 32 blind users completing tasks on 16 websites, producing 1,383 problem instances. Only 50.4% of those user problems were covered by WCAG 2.0 success criteria. This does not diminish WCAG; it shows that a guideline mapping and a lived task barrier are different units of evidence.
- Kumar and Owston (2016) compared automated checks with student-centered evaluation of online learning units. Nearly all students encountered at least one barrier, while the tools did not predict those barriers and instead flagged potential barriers that were not relevant to the participating students.
- AlSaeed and colleagues (2020) combined four automated tools with task testing by four visually impaired teachers on three e-government systems. Automated checks identified prominent technical issues, while participants exposed screen-reader, navigation, task-completion, training, confidence, independence, and privacy consequences.

- Vollenwyder and colleagues (2023) randomized 66 participants with visual impairments and 65 without visual impairments to an online shop with low or high standards conformance. Quantitative usability and user-experience outcomes did not differ significantly, although open-ended responses suggested more positive experiences for visually impaired participants and fewer negative experiences for participants without visual impairments on the higher-conformance version. The authors recommended complementing conformance approaches with user-centered and participatory methods.

No one study should be universalized. Together they establish a boundary: standards, expert review, and user experience are related sources of evidence, not interchangeable proxies.

3.5 Newer automation and AI do not erase the boundary

Newer tools can improve rule breadth, text analysis, or reporting. That progress should be evaluated, not assumed.

Ara and Sik-Lanyi (2025) compared an algorithmic scoring environment with 80 ratings from 40 participants across 20 Hungarian university pages. Their score aligned with participant perceptions for many pages. The result is promising for that tool and sample, but the participant ratings were general perception scores, not a disabled-user or expert conformance ground truth.

Palmer and Oswal (2024) examined three generative-AI website builders using automated scanning and expert screen-reader testing. Their study identified accessibility problems in both the creation workflows and generated products, illustrating why code generation does not remove the need for expert interaction testing.

Hartman and Gorichanaz (2025) installed three AI-powered accessibility overlays on two synthetic websites containing known errors. Automated scores improved, but keyboard and screen-reader testing showed that major errors remained. The study is small, but its design exposes a crucial measurement risk: the score can improve while the underlying interaction remains materially inaccessible.

4. Why This Review Does Not Publish a Pooled Percentage

Statistical meta-analysis requires studies to estimate sufficiently comparable effects. These studies do not.

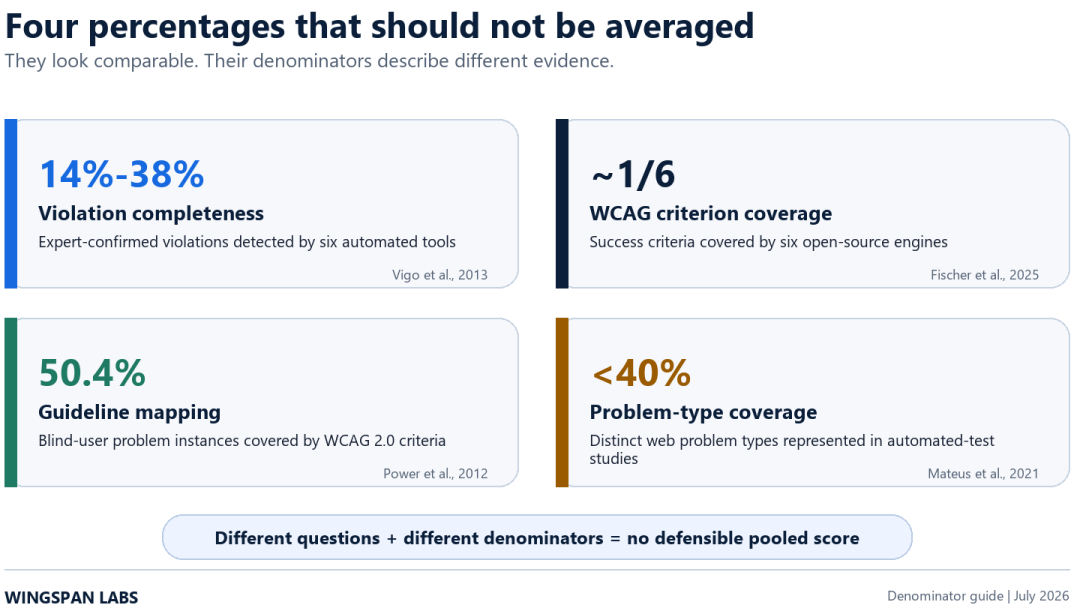

Figure 2. Four percentages that answer different questions.

Values are preserved from Vigo et al. (2013), Fischer et al. (2025), Power et al. (2012), and Mateus et al. (2021).

One number describes success-criterion coverage. Another describes the proportion of expert-confirmed violations a tool found. Another describes user problems that map to guideline criteria. Another counts problem types represented across studies. Treating them as versions of the same percentage would create false precision.

The Cochrane Handbook warns that meta-analysis can seriously mislead when study designs, outcome types, bias, and heterogeneity are not addressed. It also requires study-level effects to be expressed in a common way before they are combined (Deeks et al., 2024). Because that condition is not met, this review uses denominator-preserving quantitative examples and a structured narrative synthesis.

This no-pooling decision is a result, not a failure. It tells decision-makers that there is no defensible universal answer to "What percentage of accessibility does automation catch?"

5. How to Read an Automated Accessibility Score

Before a score reaches a board report, procurement memo, or public claim, ask six questions.

1. What exactly was scanned?

Was it one page, a page sample, a logged-in workflow, a generated state, or the full site? A homepage score cannot stand in for a transaction, form, document, or account workflow that was never tested.

2. Which tool, version, rules, and configuration produced it?

Tool outputs vary. Version changes can add rules, fix defects, or change scoring. A number without its technical provenance is not reproducible evidence.

3. What could the tool decide automatically?

The report should separate definite automated failures from potential issues and manual checks. A clean automated result may mean "no failures among the implemented automated rules," not "no accessibility barriers."

4. Who resolved the judgment calls?

Potential issues should be inspected by an experienced reviewer using a defined method. The report should identify unresolved items rather than converting uncertainty into a passing score.

5. What interaction testing was completed?

Keyboard navigation, focus behavior, zoom and reflow, screen-reader use, error recovery, touch targets, and dynamic states are interaction questions. Automated evidence can prioritize them; it cannot substitute for performing them.

6. What claim is the score being used to support?

"The scanner found no failures in its automated rules" is a bounded technical statement. "The website is accessible," "WCAG compliant," or "legally safe" is a much larger conclusion requiring different evidence and, where applicable, qualified legal advice.

6. A Practical Three-Layer Evidence Model

Layer 1: Automated checks throughout development

Use automation early and often for repeatable code-detectable rules. Preserve tool/version data and integrate checks into regression work. Treat the output as an issue detector, not a certification engine.

Layer 2: Experienced manual evaluation

Review representative pages, templates, components, and end-to-end workflows. Resolve potential issues. Test structure, meaning, keyboard behavior, focus, dynamic changes, responsive reflow, and assistive-technology interaction appropriate to scope.

Layer 3: Task-based testing with disabled users

When the decision concerns actual use, include people whose access needs match the product and tasks. Observe completion, workarounds, confidence, errors, effort, and consequences. Do not use a small participant group to claim universal accessibility; use the evidence to identify barriers and improve the experience.

W3C's reporting template follows the same basic logic: conformance evaluation combines semi-automated tools with manual evaluation by an experienced reviewer and documents which pages and methods were used (W3C Web Accessibility Initiative, n.d.-b).

7. What a Credible Report Should Contain

A meaningful accessibility report should make its evidence inspectable. At minimum, it should state:

- the product, routes, states, and documents in scope;
- representative page and workflow selection;
- evaluation date and version under review;
- WCAG version and target level, if conformance is being evaluated;
- automated tools, versions, rulesets, settings, and known limitations;
- definite failures, potential issues, manual checks, and unresolved items;
- manual methods, reviewer experience, browsers, devices, and assistive technologies;
- user-testing participants and tasks, when included;
- severity and prioritization method;
- remediation status and regression evidence;
- claims that the evidence does and does not support.

An accessibility score without that context is easier to read. It is also easier to misuse.

8. Questions to Ask a Vendor or Internal Team

1. Which pages, workflows, documents, and states did you test?
2. Which automated tools and versions did you use?
3. Which findings required manual judgment, and who made those decisions?
4. What keyboard, zoom/reflow, screen-reader, and responsive checks were completed?
5. Were disabled users involved? If so, which tasks and access needs were represented?
6. How were false positives, false negatives, and potential issues handled?
7. What remains untested?
8. What evidence supports any claim of conformance, accessibility, or legal readiness?
9. How will fixes be regression-tested after content or code changes?
10. Can another qualified reviewer reproduce the result from the report?

9. Claims Boundaries

This review supports the statement that automated accessibility testing is useful but incomplete and must be interpreted in context.

It does **not** establish:

- the accessibility or WCAG conformance of Wingspan Labs, HorizonSong, or any client website;
- legal compliance under the ADA, Section 508, the European Accessibility Act, or another law;
- a universal automation coverage percentage;

- superiority of a particular scanner, consultant, or methodology;
- customer outcomes, conversion improvement, or market demand;
- that one user study represents all disabled people.

10. Review Limitations

This was a rapid review, not a preregistered exhaustive systematic review. One reviewer conducted screening and extraction. The search used broad scholarly APIs rather than subscription databases, and several older studies were available only through detailed abstracts. Tool versions, WCAG editions, website technologies, samples, and outcome definitions varied substantially. Publication bias is possible. Some studies were conducted by tool developers or evaluated their own proposed method. The evidence base is stronger for visual and screen-reader-related web barriers than for the full range of cognitive, motor, hearing, speech, and intersectional access needs.

Those limitations are another reason not to manufacture a pooled score.

Conclusion

Automated accessibility tools should be part of serious web practice. They are fast, repeatable, scalable, and capable of finding real defects. But their most persuasive output - the score - is also the easiest output to overread.

A responsible interpretation is simple:

- use automated checks for the rules they can test;
- preserve the tool's scope and uncertainty;
- add experienced manual evaluation;
- include disabled-user task evidence when the decision concerns real use;
- make claims no broader than the evidence.

The question is not whether to automate accessibility testing. The question is whether the organization will mistake a useful instrument reading for the condition it is trying to understand.

References

Abduganiev, S. G. (2017). Towards automated web accessibility evaluation: A comparative study. *International Journal of Information Technology and Computer Science, 9*(9), 18-44. DOI 10.5815/ijitcs.2017.09.03

Acosta-Vargas, P., Salvador-Ullauri, L. A., & Lujan-Mora, S. (2019). A heuristic method to evaluate web accessibility for users with low vision. *IEEE Access, 7*, 125634-125648. DOI 10.1109/ACCESS.2019.2939068

Alsaeedi, A. (2020). Comparing web accessibility evaluation tools and evaluating the accessibility of webpages: Proposed frameworks. *Information, 11*(1), 40. DOI 10.3390/info11010040

AlSaeed, D., Alkhalifa, H., Alotaibi, H., Alshalan, R., Al-Mutlaq, N., Alshalan, S., Bintaleb, H. T., & AlSahow, A. M. (2020). Accessibility evaluation of Saudi e-government systems for teachers: A visually impaired user's perspective. *Applied Sciences, 10*(21), 7528. DOI 10.3390/app10217528

Ara, J., & Sik-Lanyi, C. (2025). Automated evaluation of accessibility issues of webpage content: Tool and evaluation. *Scientific Reports, 15*, 9516. DOI 10.1038/s41598-025-92192-5

Brajnik, G. (2008). A comparative test of web accessibility evaluation methods. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 113-120). DOI 10.1145/1414471.1414494

Brajnik, G., Yesilada, Y., & Harper, S. (2011). The expertise effect on web accessibility evaluation methods. *Human-Computer Interaction, 26*(3), 246-283. DOI 10.1080/07370024.2011.601670

Centeno, V. L., Kloos, C. D., Fisteus, J. A., & Alvarez, L. A. (2006). Web accessibility evaluation tools: A survey and some improvements. *Electronic Notes in Theoretical Computer Science, 157*(2), 87-100. DOI 10.1016/j.entcs.2005.12.048

Deeks, J. J., Higgins, J. P. T., Altman, D. G., McKenzie, J. E., & Veroniki, A. A. (2024). Chapter 10: Analysing data and undertaking meta-analyses. In J. P. T. Higgins et al. (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (Version 6.5). Cochrane Handbook, Chapter 10

Fischer, T., Lundell, B., & Gamalielsson, J. (2025). Coverage of web accessibility guidelines provided by automated checking tools. *Universal Access in the Information Society, 24*, 3615-3637. DOI 10.1007/s10209-025-01263-x

Hartman, P., & Gorichanaz, T. (2025). Evaluating AI-powered website accessibility overlays. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 1-4). DOI 10.1145/3663547.3759761

Ismailova, R., & Inal, Y. (2022). Comparison of online accessibility evaluation tools: An analysis of tool effectiveness. *IEEE Access, 10*, 58233-58239. DOI 10.1109/ACCESS.2022.3179375

Kelsey-Adkins, E. R., & Thompson, R. H. (2022). Inter-rater reliability of command-line web accessibility evaluation tools. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. DOI 10.1145/3517428.3550395

Kumar, K. L., & Owston, R. (2016). Evaluating e-learning accessibility by automated and student-centered methods. *Educational Technology Research and Development, 64*(2), 263-283. DOI 10.1007/s11423-015-9413-6

Kumar, S., Jeevitha Shree, D. V., & Biswas, P. (2021). Comparing ten WCAG tools for accessibility evaluation of websites. *Technology and Disability, 33*(3), 195-209. DOI 10.3233/TAD-210329

Mateus, D. A., Silva, C. A., de Oliveira, A. F. B. A., Costa, H., & Freire, A. P. (2021). A systematic mapping of accessibility problems encountered on websites and mobile apps: A comparison between automated tests, manual inspections and user evaluations. *Journal on Interactive Systems, 12*(1), 145-171. DOI 10.5753/jis.2021.1778

Palmer, Z. B., & Oswal, S. K. (2024). Constructing websites with generative AI tools: The accessibility of their workflows and products for users with disabilities. *Journal of Business and Technical Communication*. DOI 10.1177/10506519241280644

Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. DOI 10.1136/bmj.n71

Power, C., Freire, A., Petrie, H., & Swallow, D. (2012). Guidelines are only half of the story: Accessibility problems encountered by blind users on the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 433-442). DOI 10.1145/2207676.2207736

Tzimas, I., & Katsanos, C. (2021). Effect of potential issues flagged by automated tools on web accessibility evaluation results: A case study on university department websites. In *Proceedings of the 25th Pan-Hellenic Conference on Informatics*. DOI 10.1145/3503823.3503845

Tzimas, I., & Katsanos, C. (2024). Do potential issues reported by automated tools affect guideline-based web accessibility evaluation? A case study on Greek public hospitals. In *Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics*. DOI 10.1145/3716554.3716596

Vigo, M., Brown, J., & Conway, V. (2013). Benchmarking web accessibility evaluation tools: Measuring the harm of sole reliance on automated tests. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (pp. 1-10). DOI 10.1145/2461121.2461124

Vollenwyder, B., Petralito, S., Iten, G. H., Bruhlmann, F., Opwis, K., & Mekler, E. D. (2023). How compliance with web accessibility standards shapes the experiences of users with and without disabilities. *International Journal of Human-Computer Studies, 170*, 102956. DOI 10.1016/j.ijhcs.2022.102956

W3C Web Accessibility Initiative. (n.d.-a). *Selecting web accessibility evaluation tools*. Retrieved July 20, 2026, from W3C guidance on selecting evaluation tools

W3C Web Accessibility Initiative. (n.d.-b). *Template for accessibility evaluation reports*. Retrieved July 20, 2026, from W3C accessibility evaluation report template