Construction of Large-Scale Digital Accessibility Evaluation Indicator System for the Elderly and People with Disabilities

Liu Xincheng[1]*

[1] The University of Sydney, NSW 2006, Australia
* Correspondence: The University of Sydney, NSW 2006, Australia.

https://doi.org/10.53104/j.acad.res.adv.2026.06005

Citation: Liu, X. (2026) 'Construction of Large-Scale Digital Accessibility Evaluation Indicator System for the Elderly and People with Disabilities', *Journal of Academic Research and Advances*, *2*(2), pp. 51-60.

Copyright: © 2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Digital services have been deeply integrated into public life, and the problem of digital exclusion among the elderly and people with disabilities has become increasingly prominent. Existing accessibility evaluations mostly rely on WCAG specifications, focusing on technical compliance detection for sensory-related barriers, while failing to sufficiently cover local scenarios involving age-related cognitive decline and multiple-disability intersections. Moreover, high costs of manual assessment hinder large-scale screening. Based on the capability-demand matching theory and universal design theory, this paper constructs an initial indicator pool through literature coding and qualitative user analysis. The Delphi-CRITIC-grey relational combined analysis is adopted for indicator screening and subjective-objective weight assignment, establishing a digital accessibility evaluation indicator system consisting of 5 primary-level, 18 secondary-level, and 62 tertiary-level quantitative observation indicators (Martins, B., & Duarte, C., 2024). Reliability, validity and discrimination tests are completed. Empirical evaluation is conducted on 127 high-frequency domestic public digital services, yielding an average composite sample score of 58.72 (W3C Web Accessibility Initiative, 2012). Cognitive comprehensibility, fault tolerance and security guarantee constitute major weak points. Traditional automated WCAG tools overestimate the real-world accessibility level for composite groups of the elderly and people with disabilities. Supporting a hybrid evaluation mode combining automated parsing and sampling-based manual review, this system can provide quantitative evidence for optimization of digitally accessible products and the formulation of industrial evaluation specifications.

Keywords: digital accessibility; information accessibility; elderly population; people with disabilities; indicator system; large-scale evaluation; combined weighting

1. Introduction

1.1 Research Background

With the development of mobile internet and digital government affairs, webpages, applications and mini-programs have become major channels for the public to obtain public services. By the end of 2025, China's population aged 60 and above exceeded 310 million (Chiriboga-Casanova, W., Medina-Medina, N., & Paderewski-Rodríguez, P., 2025), and the registered population with disabilities reached more than 85 million. Some groups suffer from sensory degradation, mobility limitations and cognitive decline. Deficiencies in digital accessibility create barriers to service access and exacerbate social exclusion among the elderly and people with disabilities. Current digital accessibility retrofitting mostly focuses on code compliance and ignores use-related obstacles caused by elderly cognitive deterioration. International WCAG 2.2 standards and automated assessment tools prioritize technical compliance, rooted in overseas scenarios. They poorly adapt to cognitive load and operation habits of domestic elderly users and cannot represent real-world user experience. Industry and regulatory authorities urgently demand quantitative evaluation tools capable of batch assessment that integrate both technical indicators and user experience.

1.2 Domestic and Overseas Research Status

Overseas research mainly focuses on accessibility assessment tools, interactive experiments with elderly users and accessibility measurement models. WCAG-based automated tools show insufficient capacity for cognitive-barrier identification; user controlled experiments are limited in sample size and cannot support horizontal comparison across multiple products; parameters of existing measurement models are adapted to overseas scenarios and deviate from behavioural characteristics of domestic elderly users. Domestic studies concentrate on barrier investigations, small-scale evaluations and tool development. Most directly adopt WCAG items, lacking quantitative indicators and weight systems oriented toward cognitive ageing. No complete quantitative evaluation framework integrating technical detection, local characteristics and subjective experience has been constructed.

1.3 Research Questions

Synthesizing existing theories and practical status quo, three core research questions are proposed:

1) For composite groups of the elderly and people with disabilities, how to build a digital-accessibility indicator pool compatible with machine automatic detection and manual experience evaluation to remedy the deficiency of traditional frameworks in covering cognitive-decline scenarios.
2) How to integrate expert experience and objective data to complete indicator screening and weight calculation, forming a quantitative evaluation system with sound reliability and validity applicable to hundreds-of-sample-scale assessments.
3) Conduct empirical evaluation based on real-world public digital-service samples. What are the main existing shortcomings of domestic public-digital-service accessibility, and what assessment biases exist in traditional automated WCAG evaluation tools?

1.4 Research Innovations

Targeting the above-mentioned research questions, innovations lie in theoretical, methodological and empirical dimensions:

- **Theoretical dimension**: Introduce elderly cognitive-decline dimension into the formal evaluation framework. Strengthen the evaluation status of cognitive comprehensibility and fault-tolerance guarantee based on the four WCAG principles, correcting the traditional bias of prioritizing sensory aspects while neglecting cognition, and adapting to usage characteristics of domestic elderly-disability intersecting groups.

- **Methodological dimension**: Combine Delphi expert experience and CRITIC objective data to realize combined weighting. All tertiary indicators are equipped with implementable quantitative observation rules. A hybrid evaluation mode integrating automated parsing and sampling-based manual review is established, balancing assessment efficiency and evaluation completeness for large-scale evaluation scenarios.
- **Empirical dimension**: Complete empirical evaluation of 127 public-digital-service samples, quantitatively reveal accessibility shortcomings of domestic digital services, verify the overestimation bias of traditional automated WCAG tools through comparison, and provide empirical evidence for industrial rectification and industry specifications.

The technical route proceeds sequentially: literature review-barrier-mechanism analysis-initial indicator-pool construction-multi-round indicator screening-combined weighting-reliability-validity testing-large-scale empirical evaluation-result discussion and optimization suggestions.

2. Relevant Concepts and Theoretical Foundations

2.1 Definition of Core Concepts

Key research concepts are clearly defined to clarify connotations and boundaries for subsequent indicator construction and empirical evaluation.

- **Digital Accessibility**: The comprehensive attribute of digital interfaces (webpages, APPs, mini-programs) enabling information perception, executable commands, comprehensible content, error fault tolerance and compatibility with assistive technologies for users with sensory, motor or cognitive impairments. Different from narrow-sense disability-oriented accessibility, it emphatically covers intersecting composite usage scenarios of the elderly and people with disabilities.
- **Large-Scale Digital Accessibility Evaluation**: Distinct from small-sample user experiments. For dozens to hundreds of digital-product samples, program-driven automated parsing of interface elements is adopted supplemented by stratified-sampling standardized manual review. Composite accessibility scores are calculated in batches to support cross-product horizontal comparison and barrier-distribution statistics.
- **Elderly-Disability Composite Group**: Includes adult users with visual, hearing or physical disabilities, as well as elderly users suffering from sensory degradation, declined motor control and working-memory decay. Some individuals are subject to multiple functional constraints simultaneously.

2.2 Basic Theoretical Support

Relevant basic theories and specifications are sorted out to interpret theoretical adaptation logic for designing dimensions and observation items of the indicator system.

- **Capability-Demand Matching Theory**: A matching relationship exists between users' physiological-cognitive capability set and digital-interface supply capabilities. When interface interaction requirements exceed users' capability thresholds, digital-usage barriers emerge. Interfaces shall lower rigid requirements for perception precision, operation proficiency and memory capacity to narrow capability-demand gaps.
- **Cognitive Load Theory**: The elderly exhibit reduced working-memory capacity and sustained-attention capability. Highly complex interfaces cause cognitive overload and raise operation-error rates. Information

architecture, pop-up frequency and operation steps directly determine cognitive-load levels.

- **Universal Design Theory**: Products shall be designed for maximum inclusive usability for all populations, adapting to users of diverse capability levels without special retrofitting. It provides theoretical support for indicator comprehensiveness in this study.
- **Four Underlying WCAG 2.2 Principles**: Perceivable, Operable, Understandable, Robust. Taken as the reference baseline for dimension division in this research, they are expanded and revised according to characteristics of domestic elderly users.

2.3 Analysis of Constraints for Large-Scale Evaluation

For large-scale scenarios, the indicator system must satisfy two hard constraints:

1) Part of the indicators can be automatically collected through page DOM parsing and interface-element identification.
2) For experience-oriented indicators unrecognizable by machines, standardized scoring rules shall be formulated to support unified manual scoring by evaluators and reduce subjective bias.

The system shall balance technical achievability and completeness of user experience and avoid total reliance on high-cost in-depth user experiments.

3. Analysis of User-Barrier Mechanisms and Construction of Initial Indicator Pool

3.1 Deconstruction of Digital-Usage Barriers Among the Elderly and People with Disabilities

Barrier sources are classified into four categories: sensory, physical-motor, cognitive and external-context barriers. According to empirical statistics from the 127 public-digital-service samples in this research (Chen, X., & Sun, K., 2025), detection rates of various barriers differ:

1) **Sensory barriers (detection rate 62.2 %)**: Including difficulties in visual-contrast discrimination, font recognition limitations, and absence of text alternatives for audio information. They constitute common basic accessibility defects.
2) **Physical-motor barriers (detection rate 54.3 %)**: Manifested as insufficient clicking precision, inability to complete complex gestures, and limited tolerable operation duration, exerting significant impacts on mobility-impaired users.
3) **Cognitive barriers (detection rate 76.4 %)**: Concentrated in memory decay, difficulties in understanding multi-step tasks, and obstacles in reading prompt information. This category yields the highest detection rate among samples and creates prominent problems for elderly users.
4) **External-context barriers (detection rate 48.0 %)**: Including strong-light environments, noisy surroundings and performance constraints of outdated terminals, representing usage restrictions arising from environment-device coupling.

Note: Detection rate refers to the proportion of all 127 evaluated samples presenting at least one barrier of the corresponding category.

Five sources feed the indicator pool: decomposition of WCAG 2.2 clauses, extraction of items from highly-cited domestic-overseas accessibility literature, coding of semi-structured interviews with 21 elderly and disabled users, in-depth-interview records of 8 industry experts (Thomas, J., Almidani, L., Swenor, B. K., & Varadaraj, V., 2024), and collation of detection items of mainstream accessibility-evaluation tools. Following five principles: comprehensiveness, measurability, pertinence, discrimination and large-scale adaptability, an initial indicator pool containing 5 primary-level dimensions, 23 secondary-level dimensions and 87 tertiary-level observation indicators is preliminarily established. Primary-level dimensions include Perceptual Accessibility, Operational Executability, Cognitive

Comprehensibility, Fault-Tolerance and Security Guarantee, and Assistive-Technology Compatibility.

Table 1. Overview of Initial Indicator-Pool Structure

Primary-Level Dimension	Number of Secondary-Level Indicators	Number of Tertiary-Level Indicators	Indicator-Type Distribution
Perceptual Accessibility	6	22	14 automated, 8 manual
Operational Executability	5	19	11 automated, 8 manual
Cognitive Comprehensibility	5	18	7 automated, 11 manual
Fault-Tolerance and Security Guarantee	4	14	6 automated, 8 manual
Assistive-Technology Compatibility	3	14	8 automated, 6 manual

3.2 Establishment of Preliminary Indicator-Screening Criteria

The initial indicator pool suffers from semantic overlap among entries, prominent qualitative characteristics of partial indicators and insufficient quantifiable detectability, failing empirical requirements for large-scale standardized evaluation. To guarantee scientific validity, independence and operability of the indicator system, standardized screening criteria are preset:

- Eliminate inefficient indicators with importance scores below 3.5. For indicators with variation coefficients greater than 0.35 and high divergence in expert reviews, optimize wording or delete them directly to resolve cognitive disagreements.
- Restructure purely-qualitative indicator entries lacking implementable observation standards and clarify quantitative judgement rules.
- Merge duplicate indicators with high semantic overlap and crossed evaluation dimensions to streamline the indicator system.

After systematic sorting and pre-processing of the initial indicator pool, a standardized Delphi expert-consultation questionnaire is compiled. Two-round multi-cycle expert surveys are implemented to further verify indicator rationality and provide reliable data support for subsequent indicator weighting and system construction.

4. Indicator Screening and Combined Weighting Based on Delphi-CRITIC-Grey-Relational Analysis

This core methodological chapter accomplishes indicator refinement, weight calculation and system-validity verification. The expert sample contains 16 respondents: 4 university scholars majoring in human-computer interaction, 5 accessibility-engineering practitioners, 3 gerontological-sociology researchers and 4 disability-rights practitioners. The average expert authority coefficient reaches 0.81, satisfying Delphi-method requirements.

4.1 Explanation of Method Principles

Delphi Method: Conduct multi-round anonymous questionnaires for experts to score indicator importance and measurability. Calculate expert coordination coefficients to judge opinion convergence and realize indicator deletion, revision and merging.

Grey Relational Analysis: Calculate correlation degrees between each tertiary indicator and overall accessibility level to identify key indicators exerting high influence on comprehensive evaluation. The grey-relational-coefficient formula:

$$\xi_i(k)=\frac{\min_i\min_k|x_0(k)-x_i(k)|+\rho\max_i\max_k|x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)|+\rho\max_i\max_k|x_0(k)-x_i(k)|}$$

Where: $x_0(k)$ denotes reference sequence; $x_i(k)$ denotes comparison sequence; ρ denotes distinguishing coefficient set as 0.5 in this paper; $\xi_i(k)$ denotes relational coefficient. The mean value is further computed to obtain grey relational degree r_i.

CRITIC Objective Weighting Method: Compute objective weights utilizing both indicator variation degrees and inter-indicator conflict relationships. Information quantity C_j of each indicator:

$$C_j=\sigma_j\sum_{i=1}^{m}(1-r_{ij})$$

Where σ_j represents standard deviation of the j-th indicator; r_{ij} denotes correlation coefficient between indicator i and indicator j. Objective-weight formula:

$$w_j^o=\frac{C_j}{\sum_{j=1}^{n}C_j}$$

Subjective-Objective Combined Weighting: Linearly combine subjective weights w_j^s obtained via Delphi and CRITIC objective weights w_j^o. Combined weight:

$$w_j=\alpha w_j^s+(1-\alpha)w_j^o$$

Drawing on existing similar indicator-system research, α=0.5 is adopted to balance expert subjective experience and objective-data characteristics.

4.2 Two-Round Delphi Consultation and Indicator Refinement

16 valid questionnaires are recovered in the first-round consultation with a 100 % expert positive-response coefficient. Importance mean values and variation coefficients of all 87 indicators are calculated. Seven indicators of insufficient importance and six indicators with excessive divergence are deleted. Four groups of duplicate entries are merged, and wordings of 21 indicators are revised. Seventy tertiary indicators enter the second-round consultation.

After the second-round consultation, Kendall's coordination coefficient W=0.726 with significance p<0.001, indicating convergent expert opinions. After two-round screening, 5 primary-level dimensions, 18 secondary-level dimensions and 62 tertiary-level observation indicators are finally confirmed.

Table 2. Final Primary-Level Dimensions and Combined-Weight Results

Primary-Level Dimension	Combined Weight	Evaluation Attribute	Brief Core Connotation
Perceptual Accessibility	0.214	Automated + Manual	Visual-auditory information perceivability for users with disabilities
Operational Executability	0.227	Automated + Manual	Completion of interactive actions by mobility-limited users
Cognitive Comprehensibility	0.241	Mainly manual review	Content-logic adaptation to age-related cognitive decline
Fault-Tolerance and Security Guarantee	0.176	Automated + Manual	Mechanisms for error prompts, revocation and risk warning
Assistive-Technology Compatibility	0.142	Mainly automated detection	Adaptation capacity for screen readers, magnifiers and other assistive tools

4.3 Key-Indicator Identification via Grey Relational Analysis

Grey relational calculation is performed on the simulated matrix of indicator evaluation samples. Grey-relational-degree values of tertiary indicators range from 0.527 to 0.841. High-correlation top-ranked indicators include mandatory pop-up blocking, operation-step hierarchy, readability of error prompts, screen-reader-label adaptation, and

dynamic-font-scaling capability. These factors exert decisive influences on overall user experience of the elderly-disability composite group.

4.4 Reliability and Validity Tests of the Indicator System

Reliability, content validity and discriminant validity tests are implemented to verify reliability, effectiveness and discrimination capability of the constructed digital-accessibility evaluation indicator system.

- **Reliability Test**: Cronbach's α coefficient is calculated. The overall system yields α=0.834 , and α values of each primary-level dimension range from 0.712 to 0.789, all above the 0.7 benchmark threshold, proving satisfactory internal indicator consistency.
- **Content Validity**: The average Delphi expert authority coefficient reaches 0.81, demonstrating sound content validity for the questionnaire and indicator system.
- **Discriminant Validity**: Six digital products with distinct accessibility levels are selected for pre-evaluation. Composite scores range from 41.36 to 82.51. The system can effectively differentiate accessibility grades of different products.

4.5 Construction of Large-Scale Hybrid-Evaluation Implementation Workflow

The complete evaluation workflow contains four phases:

1) **Batch automated parsing**: Capture interface-element parameters and score machine-recognizable indicators.
2) **Risk-sample marking**: Push abnormal entries output by automation for manual review.
3) **Stratified-sampling manual evaluation**: Perform standardized scoring for cognitive and experience-oriented indicators with a 25 % sampling ratio.
4) **Weighted composite-score calculation**: Output sub-item diagnostic results. Composite-score formula:

$$S=\sum_{i=1}^{5} w_i\left(\sum_{j=1}^{n_i} w_{ij}s_{ij}\right)$$

Where: S denotes sample composite accessibility score; w_i represents weight of primary-level dimension i; w_{ij} denotes weight of corresponding secondary-level dimension j ; s_{ij} stands for standardized score of secondary-level dimension j.

5. Large-Scale Empirical Evaluation for Public Digital Services

5.1 Empirical-Sample and Evaluation-Environment Settings

Specification for empirical-experiment design and sample quality control:

1) **Sample Selection and Distribution**

Empirical samples are high-frequency public digital services used by domestic elderly and disabled groups, totalling 127 valid samples covering diverse high-frequency service scenarios: 32 government-affairs services, 27 medical-health services, 24 life-payment services, 23 news-information services and 21 e-commerce services. Sample carriers cover three mainstream service forms: mobile APPs, webpages and WeChat mini-programs. Sample coverage and representativeness satisfy requirements for large-scale horizontal evaluation.

2) **Evaluation Implementation and Quality Control**

This research adopts the hybrid evaluation mode of "automated parsing + manual review". Automated page-parsing scripts extract interface-element and data-information in a standardized manner. Manual evaluation strictly follows unified scoring rules. Two professional evaluators score independently and conduct cross-validation. The Kappa coefficient for evaluation consistency reaches 0.792, indicating favourable inter-rater consistency, effective avoidance of subjective bias and reliable standardized evaluation data.

3) **Control-Group Experimental Design**

To quantitatively verify differentiated advantages of this indicator system, mainstream automated

WCAG evaluation tools are introduced for parallel comparative evaluation on the same 127 samples (W3C Web Accessibility Initiative, 2023). Result differences between the two evaluation approaches are analysed to explore limitations of traditional accessibility-evaluation tools and application value of this system.

5.2 Analysis of Overall Sample-Evaluation Results

Sample composite scores range from minimum 34.17 to maximum 84.63, with mean 58.72 and standard deviation 9.46. Only 14.17 % of all samples score higher than 70; 70.08 % fall within the 45-70 score interval; 15.75 % score below 45. Results indicate that overall accessibility of domestic public digital services for the elderly-disability composite group remains moderately low.

Table 3. Statistical Scores of Each Primary-Level Dimension

Primary-Level Dimension	Mean Score	Standard Deviation	Score Ranking
Perceptual Accessibility	63.41	10.27	2
Operational Executability	61.25	9.83	3
Cognitive Comprehensibility	52.68	11.34	5
Fault-Tolerance and Security Guarantee	54.33	10.71	4
Assistive-Technology Compatibility	66.74	12.56	1

The assistive-technology-compatibility dimension achieves the highest average score yet presents the largest sample dispersion; screen-reader adaptation performance varies drastically across applications. The cognitive-comprehensibility dimension yields the lowest score and constitutes the most prominent shortcoming, which is exactly the area inadequately covered by traditional WCAG standards.

5.3 Comparative Analysis of Different Categories of Digital Services

Mean composite score for government-affairs services = 62.37; medical-health services = 57.14; life-payment services = 58.42; news-information services = 56.86; e-commerce services = 55.29. Government-affairs platforms demonstrate relatively better overall performance but still show obvious defects in cognitive adaptation and error fault tolerance. E-commerce and news-information services suffer most severely from cognitive-load problems caused by rampant pop-ups and complicated information layout. Top-ranked high-frequency defects include pop-ups lacking simple closing mechanisms, excessive page hierarchies, overly-specialized error-prompt wording, layout chaos triggered by font scaling, and missing descriptive labels for images.

5.4 Comparison Against Traditional Automated WCAG Evaluation Results

The average score output by automated WCAG tools for identical samples reaches 71.46, significantly higher than the 58.72 mean value from this paper's system. The correlation coefficient between the two sets of evaluation results is merely 0.613. Focusing on code-level technical compliance detection, WCAG tools fail to identify core usage obstacles for elderly users such as cognitive overload, pop-up interference and cumbersome operation workflows, and therefore overestimate real-world accessibility for the elderly-disability composite group. Two typical samples are analysed as case studies:

1) **Sample A (composite score 79.22)**: Complete basic accessibility technical adaptation, supporting adaptive font-size adjustment, one-click pop-up-closing entries, and plain-language error prompts that effectively reduce cognitive burden, delivering satisfactory overall user experience.
2) **Sample B (composite score 42.65)**: Frequent mandatory pop-ups, lengthy business-operation workflows, heavily-specialized error prompts, partial screen-reader-adaptation failures. Superimposed multiple interactive defects create prominent usage barriers for the elderly and people with disabilities.

Case comparisons intuitively verify the strong problem-diagnosis capacity of this indicator system.

6. Discussion

6.1 Interpretation of Indicator-System Weights and Evaluation Results

Combined-weight results show that cognitive comprehensibility obtains the highest weight of 0.241 among all primary-level dimensions. For the elderly-disability composite group, cognitive-adaptation importance is no less significant than sensory and operational accessibility retrofitting, representing a relatively weak point of traditional international standards. Perceptual accessibility and operational executability obtain weights of 0.214 and 0.227 respectively, continuing traditional accessibility-research priorities. The 0.176 weight for fault-tolerance and security guarantee reflects practical significance of fault-tolerance mechanisms amid frequent elderly-user operation errors. Assistive-technology compatibility with weight 0.142 serves as a basic guarantee item. Nevertheless, pure screen-reader adaptation cannot guarantee satisfactory practical user experience.

Empirical results indicate that most domestic digital services accomplish basic assistive-technology adaptation yet generally underinvest in interactive design for cognitively-declining users. Many products pass WCAG code-detection while still creating usage barriers for elderly users. The root cause lies in treating accessibility merely as screen-reader compatibility during development, lacking systematic considerations for cognitive load and operation fault tolerance. The hybrid evaluation mode proposed herein assigns basic code-detection tasks to automation and supplements cognitive-experience-dimension assessment via manual sampling, balancing evaluation costs and completeness and suiting large--scale industrial-screening scenarios.

6.2 Practical Implications

For internet-product developers, this system can serve as an accessibility self-inspection checklist, emphasizing cognitive-simplification and error-fault-tolerance design instead of only completing code-level compliance rectification. For industrial regulators and standard-setters, this indicator system delivers a quantifiable scoring framework supporting horizontal comparative evaluation of diverse digital services and provides a reference foundation for subsequent industry evaluation specifications. For evaluation-tool developers, the 62 tertiary observation indicators can act as the underlying indicator library for new-generation large--scale accessibility-assessment tools. Subsequent algorithm iterations can further raise the automation-recognition proportion of cognitive-oriented indicators.

6.3 Research Limitations

Several limitations remain within this study. Empirical samples concentrate on mainstream domestic webpages, APPs and mini-programs, with insufficient coverage for IoT terminals such as smart TVs and wearable devices. Cognitive-oriented indicators still rely on sampling-based manual review at present, and full-automated detection for all indicators faces technical bottlenecks. The Delphi expert sample size has upper limits; future work may expand expert and real-user samples to further calibrate indicator thresholds and weight parameters.

6.4 Future-Research Prospects

Follow-up research can advance along four directions:

1) **Algorithm level**: Integrate interface-semantic recognition and large-model text-understanding algorithms to raise the automation-recognition ratio for cognitive-comprehensibility indicators.
2) **Evaluation-object expansion**: Incorporate IoT smart terminals into evaluation objects and

construct a unified multi-terminal digital-accessibility evaluation framework.

3) **Time-dimension research**: Conduct longitudinal follow-up evaluation to reveal evolutionary characteristics of digital-product accessibility across version iterations.
4) **Indicator-parameter optimization**: Further calibrate discrimination critical thresholds for each indicator via behavioural experiments with elderly-disability users.

7. Conclusions

Aiming at practical challenges that existing accessibility-evaluation frameworks cannot adapt to large-scale evaluations for the elderly-disability composite group, this paper builds an initial indicator pool from multi-source materials. Indicator screening and subjective-objective combined weighting are completed via Delphi-CRITIC-grey-relational analysis. A digital-accessibility evaluation system containing 5 primary-level dimensions, 18 secondary-level dimensions and 62 quantifiable tertiary-level observation indicators is established and passes reliability, validity and discrimination tests. Empirical evaluation is implemented for 127 domestic public digital services, yielding an average sample composite score of 58.72 (W3C Web Accessibility Initiative, 2023). Cognitive comprehensibility, fault-tolerance and security guarantee constitute major current shortcomings. Traditional automated WCAG evaluation tools significantly overestimate real-world accessibility for the elderly-disability composite group. This paper proposes a large--scale hybrid-evaluation workflow combining automated parsing and sampling-based manual review. It can be applied both to batch accessibility-level screening of digital services and to provide quantitative-tool support for product accessibility retrofitting and industry-specification construction.

References

Chen, X., & Sun, K. (2025) 'Digital disparities among older adults with and without disabilities in the U.S.: An examination of the census data', *Innovation in Aging*, 9(Suppl 2), igaf122. https://doi.org/10.1093/geroni/igaf122.2585

Chiriboga-Casanova, W., Medina-Medina, N., & Paderewski-Rodríguez, P. (2025) 'Accessible web design for older adults: Challenges and solutions', *ACM Transactions on Accessible Computing*, 18(3), Article 14, pp. 1-50. https://doi.org/10.1145/3763243

Martins, B., & Duarte, C. (2024) 'Large-scale study of web accessibility metrics', *Universal Access in the Information Society*, 23(1), pp. 411-434. https://doi.org/10.1007/s10209-022-00956-x

Thomas, J., Almidani, L., Swenor, B. K., & Varadaraj, V. (2024) 'Digital technology use among older adults with vision impairment', *JAMA Ophthalmology*, 142(5), pp. 445-448. https://doi.org/10.1001/jamaophthalmol.2024.0467

W3C Web Accessibility Initiative. (2012) *Research report on web accessibility metrics*. World Wide Web Consortium. Available at: https://www.w3.org/TR/accessibility-metrics-report/

W3C Web Accessibility Initiative. (2023) *Web Content Accessibility Guidelines (WCAG) 2.2*. World Wide Web Consortium. Available at: https://www.w3.org/TR/WCAG22/

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