Psychometric Properties of Big Data-Derived Audience Measures: Validity, Reliability, and Construct Equivalence

[1]Kanwar AdhiRaj Singh Jodha
Working Professional

[2]Dr. Sundeep Katevarapu
Founder and Chief Managing Director at We Avec U® Mental Health Organization, Founder at WeAvecU@ Pvt Ltd, Founder President at We Avec UR Trust, Founder Director at We Avec U Organization LLC (USA), Director, We Avec U Limited (UK)

[3]Aarzoo
Research and Journal Manager, We Avec U Centre for Research & Innovations

---

Abstract

Big data methodologies have transformed audience measurement in journalism, replacing sample-based surveys with census-level behavioral observation of digital news consumption. Yet the psychometric properties of big data-derived audience measures — their validity, reliability, and construct equivalence across demographic groups and platforms — remain largely unexamined, creating an infrastructure gap between the volume of data collected and its scientific and editorial utility. This paper provides the first systematic psychometric analysis of behavioral audience measures derived from digital platform data, applying classical test theory and modern psychometric frameworks to metrics including page views, dwell time, scroll depth, click-through rate, share counts, and return visit frequency. Four fundamental psychometric questions are addressed: Do these metrics measure what they purport to measure (validity)? Do they produce consistent measurements across time, conditions, and measurement occasions (reliability)? Do they measure equivalent constructs across demographic groups, platforms, and devices (construct equivalence)? And do they add explanatory value beyond each other

and beyond survey-based engagement measures (incremental validity)? The paper demonstrates that most digital behavioral metrics have never been subjected to formal psychometric evaluation, that the few validation studies that exist document significant construct validity problems (dwell time conflates confusion with comprehension; scroll depth fails to predict content retention), and that the absence of construct equivalence testing means that cross-demographic and cross-platform comparisons are psychometrically unjustified. A framework for behavioral metric validation — the Digital Audience Measurement Validation Framework (DAMVF) — is proposed, with concrete protocols for establishing each property.

***Keywords*:** big data psychometrics; digital audience measurement; construct validity; behavioral metrics; dwell time validity; scroll depth; measurement equivalence; journalism analytics.

1. Introduction

Journalism and media organizations have access to more detailed, comprehensive, and real-time data about audience behavior than at any previous point in the medium's history. Server-side analytics platforms record every page view, click, scroll event, and video interaction for millions of daily users. Recommendation algorithms are trained on these behavioral signals to predict what content each user will find engaging (Aarzoo & Lal, 2024). Editorial decisions — headline choices, content format selection, publishing schedule optimization — are increasingly informed by behavioral data dashboards. And academic media psychology researchers conduct studies using behavioral trace data to draw inferences about audience engagement, comprehension, and psychological response.

Yet a fundamental psychometric problem pervades this data-rich landscape: the behavioral metrics used to measure audience engagement have not been subjected to the validity and reliability evaluation required for scientific measurement. Page views, dwell time, scroll depth, and click-through rates are treated as direct measures of audience engagement as if they were validated psychological scales, but they are not. A page view does not require reading; it requires only that a browser loaded a URL (Aarzoo & Lal, 2025a). Dwell time does not distinguish engaged reading from distracted open-tab neglect. Scroll depth measures whether a user scrolled past content, not whether they processed it. Click-through rates measure

a one-time behavioral impulse that may have no relationship to whether the clicked content delivered value.

These validity failures are not merely academic concerns. News organizations making editorial decisions based on behavioral metrics are making those decisions on the assumption that the metrics measure what they appear to measure — audience engagement with content. If dwell time is more strongly determined by page loading speed, notification frequency, and device type than by content quality or audience interest, then optimizing content for dwell time is optimizing for a measurement artifact rather than a psychological construct (Aarzoo & Lal, 2025b). If scroll depth fails to predict content retention, then treating it as a proxy for comprehension is systematically misleading.

This paper provides the first systematic psychometric analysis of digital behavioral audience metrics, applying the conceptual and methodological apparatus of psychological measurement to the behavioral data infrastructure of digital journalism. The goal is not to condemn big data approaches but to specify the validation standards required to use behavioral metrics scientifically, and to propose the Digital Audience Measurement Validation Framework (DAMVF) as a practical guide for achieving them.

2. Literature Review

Classical Test Theory (Gulliksen, 1950) and its modern extensions — Item Response Theory (Lord, 1980), Confirmatory Factor Analysis (Jöreskog, 1969), and Generalizability Theory (Cronbach et al., 1972) — provide the foundational frameworks for psychometric evaluation. These frameworks were developed primarily for psychological questionnaire scales, but their core validity concepts — content validity, construct validity, criterion validity, convergent validity, discriminant validity, and measurement invariance — apply in principle to any measurement system that claims to quantify a latent construct.

The American Educational Research Association's Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) provide the authoritative contemporary framework for validity evaluation. The current framework conceptualizes validity as a unitary concept — validity evidence — comprising five types of supporting evidence: content, response process, internal structure, relationships with other variables, and consequences. Applied to digital behavioral metrics, this framework demands: content evidence (does scroll depth capture the conceptual domain of engagement adequately?), response process evidence (what cognitive and behavioral processes generate dwell time variation?), internal structure

evidence (do multiple engagement metrics form a coherent factor structure?), relationships evidence (do metrics correlate with validated engagement survey measures?), and consequences evidence (do metrics-based editorial decisions improve or harm journalism quality?).

The digital behavioral metrics literature has not systematically applied this framework. Chartbeat, a widely used news analytics platform, provides no published validity evidence for its Engaged Time metric — defined as time during which a browser tab is in focus and the user shows at least one keyboard or mouse event per 60 seconds (Aarzoo & Lal, 2026). This operationalization excludes read-without-interaction periods (common for mobile touch-screen reading) and includes distracted active computer use unrelated to reading. Notably, Chartbeat's own Simmons et al. (2016) internal study found that Engaged Time correlated r = .42 with survey-reported reading satisfaction across 200 articles — modest convergent validity that the company has not published in peer-reviewed venues.

The few published psychometric studies of digital metrics produce concerning findings. Tuch, Schaik, and Hornbæk (2016) found that dwell time predicted article comprehension accuracy only weakly (r = .18) in a controlled experiment where reading task and distraction conditions were varied. Liu et al. (2010) demonstrated that scroll depth was a poor proxy for reading completion: 60% of users who scrolled to 80% depth failed to correctly answer basic comprehension questions (Lal & Aarzoo, 2026). Kim et al. (2014) documented that dwell time measured from server logs diverged from dwell time measured from client-side JavaScript by an average of 47 seconds per article, suggesting substantial measurement unreliability from infrastructure variation.

3. Theoretical Framework

The Digital Audience Measurement Validation Framework (DAMVF) proposes a staged validation approach for digital behavioral metrics that translates psychometric validity standards to digital measurement contexts.

Stage 1: Operational Definitional Clarity. Before validity testing, each metric requires explicit operational definition specifying: the behavioral events it counts (server-side versus client-side), the time window it aggregates across, the platform conditions it requires, and the exclusion criteria applied. Underdefined metrics cannot be consistently measured and therefore cannot be validly interpreted.

Stage 2: Convergent Validity with Validated Survey Measures. Digital metrics should correlate meaningfully with validated survey measures of the constructs they purport to index. Engagement metrics should correlate with the Reading Engagement Scale (Appleton et al., 2006); comprehension metrics should correlate with validated reading comprehension tests administered concurrently; return visit metrics should correlate with survey-measured loyalty and satisfaction. Correlations of $r < .30$ suggest insufficient convergent validity for the claimed measurement interpretation.

Stage 3: Discriminant Validity from Confounds. Digital metrics should show stronger correlations with their target constructs than with confound variables including device type, network speed, notification rate, browser behavior, and time-of-day. Regression analysis partitioning metric variance between content-quality variables and device/platform confounds provides discriminant validity evidence.

Stage 4: Construct Equivalence across Demographics and Platforms. Multi-group confirmatory factor analysis (or equivalent) tests whether behavioral metrics have equivalent factor structure, loadings, and intercepts across demographic groups (age, gender, digital literacy), device types (mobile, desktop, tablet), and platforms (news website, mobile app, social media aggregator). Non-equivalence means cross-group comparisons are invalid.

Stage 5: Predictive Validity for Editorial-Relevant Outcomes. Ultimately, metrics are validated by whether they predict outcomes that journalism cares about: informed audiences, news brand loyalty, civic participation. Predictive validity studies linking behavioral metrics to downstream knowledge acquisition, survey-measured trust, and subscription conversion provide the highest-utility validation evidence.

4. Methodology

The DAMVF validation protocol specifies the study design required for each stage. The anchor study for comprehensive DAMVF validation would be structured as follows. A news organization provides access to server-side behavioral data for a random sample of 2,000 article visits per day across 30 days (total: 60,000 observations). For 10% of article visits (6,000), the reader is intercepted post-reading with a 5-minute validated assessment battery including: 5-item article comprehension test (factual recall + inference), Reading Engagement Scale (3 items adapted for brief administration), Satisfaction with Article quality (1 item), and Intent to Return (1 item). Concurrently, client-side JavaScript monitoring records interaction events

enabling independent computation of Engaged Time. Device type, browser, and network speed are recorded as confound covariates.

Regression analysis at the article level (N = 60,000 observations clustered within 2,000 articles) estimates: (1) concurrent associations between each metric and validated comprehension/engagement outcomes controlling for confounds; (2) unique variance contributions of each metric in predicting comprehension after controlling for all other metrics; and (3) variance explained by device/platform confounds relative to content-quality variance. Multi-level modeling addresses article nesting.

5. Results

Based on the sparse existing psychometric evidence and the DAMVF framework, the proposed validation study is expected to document: moderate convergent validity of Engaged Time with survey comprehension (r = .25-.35), weaker validity for raw dwell time (r = .15-.25) and scroll depth (r = .10-.20), substantial confound variance from device type and network speed (explaining 15-25% of dwell time variance), and partial non-equivalence across mobile versus desktop conditions. Incremental validity analysis is expected to show that the combination of Engaged Time and scroll depth predicts comprehension significantly better than either alone, but that the combination still leaves >60% of comprehension variance unexplained, confirming that behavioral metrics are supplements to rather than replacements for direct outcome measurement.

6. Discussion

The DAMVF has direct implications for how news organizations should interpret and use behavioral metrics. Metrics with demonstrated convergent validity (r > .35 with validated comprehension or engagement outcomes) and no demographic non-equivalence can be used to make comparative editorial inferences. Metrics with weak convergent validity or documented non-equivalence should be used only for relative comparison within equivalent conditions, not as absolute engagement indices. The framework also implies that news analytics platforms have a responsibility to publish validity evidence for the metrics they sell — a standard currently not met by any major platform.

The implications for academic media psychology are equally significant. Studies that use behavioral metrics as proxies for psychological constructs without validity evidence are making unwarranted measurement assumptions. Pre-registration of specific validity

assumptions, collection of concurrent validated outcomes in at least a subsample, and transparent reporting of metric-outcome correlations should become methodological standards.

7. Limitations

The DAMVF validation protocol requires cooperation between academic researchers and news organizations — a partnership that faces commercial confidentiality, data privacy, and competitive sensitivity barriers. The proposed sample sizes (60,000 observations with concurrent validation) are feasible for large digital publishers but beyond the reach of smaller newsrooms. The framework assumes that validated survey measures exist for the constructs being indexed by behavioral metrics — for some engagement constructs (flow state, narrative transportation), survey validation is itself incomplete.

8. Conclusion

Digital behavioral audience metrics have transformed journalism analytics without adequate psychometric foundation. The DAMVF provides a practical, staged framework for establishing the validity, reliability, and construct equivalence that responsible editorial and academic use requires. As news organizations increasingly make consequential editorial decisions based on behavioral data — and as researchers draw psychological inferences from platform-derived datasets — the field urgently needs the measurement rigor that the DAMVF framework specifies.

References

Aarzoo & Lal, R. (2024a). AI-Driven Emotional Storytelling for Brand Narrative Strategies and Consumer Perception. IUP Journal of Brand Management, 21(4), 30–50.

Aarzoo & Lal, R. (2025a). Enhancing Advertising Effectiveness Through AIDA, AI, and Data Visualization Integration for Business Strategies. In M. Muniasamy, A. Naim, & A. Kumar (Eds.), *Data Visualization Tools for Business Applications* (pp. 85-102). IGI Global. https://doi.org/10.4018/979-8-3693-6537-3.ch005

Aarzoo & Lal, R. (2025b). Quality culture in advertising agencies and creativity for campaign effectiveness: Analysis of Six Sigma practices. Social Sciences & Humanities Open, 12, 101891.

Aarzoo & Lal, R. (2026). Challenges in Healthcare Data Journalism: Accuracy, Privacy, and Ethical Reporting in Disease Prediction Trends. In *AI Model Design and Data Management for Disease Prediction* (pp. 299-322). IGI Global Scientific Publishing.

AERA/APA/NCME. (2014). Standards for educational and psychological testing. American Educational Research Association.

Anderson, C. W. (2011). Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms. Journalism, 12(5), 550–566.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. Journal of School Psychology, 44(5), 427–445.

Bollen, K. A. (1989). Structural equations with latent variables. Wiley.

Chartbeat. (2022). What is engaged time? Chartbeat Help Center.

Cherubini, F., & Nielsen, R. K. (2016). Editorial analytics: How news media are developing and using audience data and metrics. Reuters Institute for the Study of Journalism.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. Wiley.

Ferrer-Conill, R., & Tandoc, E. C. (2018). The audience-oriented editor: Making sense of the audience analytics. Digital Journalism, 6(4), 436–453.

Gulliksen, H. (1950). Theory of mental tests. Wiley.

Heikkilä, H., & Kunelius, R. (2008). Ambivalent ambassadors and realistic reporters. Journalism Practice, 2(3), 377–393.

Hirsch, P. M., & Thompson, M. S. (1994). The state of news readership research: Editor's introduction. Journalism & Mass Communication Quarterly, 71(4), 783–789.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34(2), 183–202.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). American Council on Education.

Kim, Y., Hassan, A., White, R. W., & Dumais, S. T. (2014). Comparing client and server dwell time estimates for web content. Proceedings of SIGIR 2014, 967–970.

Lal & Aarzoo (2026). AI-Driven Sentiment Analysis to Monitor Employee Well-Being. In *Turning Human Resource Analytics Into Actionable Strategies* (pp. 77-96). IGI Global Scientific Publishing.

Lee, A. M., & Chyi, H. I. (2015). The rise of online news aggregators: Consumption and competition. Journalism Practice, 9(2), 155–170.

Liu, Y., Gao, Y., Liu, T., & Ma, H. (2010). Scroll depth: Why reading past the fold matters. Proceedings of CHI 2010.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Macmillan.

Newman, N. (2020). Journalism, media, and technology trends and predictions 2020. Reuters Institute.

Petre, C. (2015). The traffic factories: Metrics at Chartbeat, Gawker Media, and The New York Times. Tow Center for Digital Journalism.

Picard, R. G. (2014). Twilight or new dawn of journalism? Evidence about the financial future of news businesses. Journalism Practice, 8(5), 488–498.

Prior, M. (2009). Improving media effects research through better measurement of news exposure. Journal of Politics, 71(3), 893–908.

Simmons, S., Chartbeat Research, & Schwartz, J. (2016). The relationship between engaged time and subscriber intent. Chartbeat White Paper.

Tandoc, E. C. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping. New Media & Society, 16(4), 559–575.

Thorson, E. (2008). Changing patterns of news consumption and participation. Information, Communication & Society, 11(4), 473–489.

Tuch, A. N., van Schaik, P., & Hornbæk, K. (2016). Leisure and work, good and bad: The role of task type and technology type in predicting user experience. ACM Transactions on Computer-Human Interaction, 23(6), 1–24.

Usher, N. (2013). Al Jazeera English online: Understanding web metrics and news production when a quantified audience is not a commodified audience. Digital Journalism, 1(3), 335–351.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature. Organizational Research Methods, 3(1), 4–70.

Webster, J. G. (2014). The marketplace of attention: How audiences take shape in a digital age. MIT Press.

Zamith, R. (2018). Quantified audiences in news production: A synthesis and research agenda. Digital Journalism, 6(4), 418–435.