Usability and User Experience, Vol. 232, 2026, 143–150
https://doi.org/10.54941/ahfe1008043

Beyond Random Sampling: Behavioral Targeting as a Human Factors Methodology for Uncovering Latent Citizen Needs in Government Web Services

So Nishina

Nishina Office, Tokyo, Japan

ABSTRACT

Government digital services are built on an implicit promise of universal accessibility, yet the methods most commonly used to evaluate them are structurally incapable of capturing the experiences of those who depend on them most. This paper examines a persistent methodological gap in public sector UX evaluation: the reliance on random sampling, which systematically dilutes the feedback of task-motivated citizens with the opinions of users who have never encountered the service's core failure points. Drawing on a case study of a municipal web portal redesign in Japan, we present a behavioral targeting approach in which a real-time analytics platform identifies purposeful visitors — defined operationally by scroll depth — and delivers satisfaction surveys exclusively to this segment. This method is grounded in the principle of ecological validity: measuring users within the context of genuine task performance rather than through decontextualized recall. The targeted survey revealed an overall Net Promoter Score (NPS) of –52.3, a navigation failure rate of approximately 25%, and three prioritized failure dimensions identified through journey map analysis. A subsequent redesign, anchored in smartphone-first architecture and AI-assisted search functionality, produced measurable improvements validated by post-launch citizen feedback. From these findings, we propose a replicable Behavioral Targeting Evaluation Model (BTEM) for public sector usability assessment. The implications extend well beyond the Japanese context: as local governments worldwide accelerate digital transformation, the ability to isolate and address the right users' failures — rather than averaging across the full population — will determine whether digital public services achieve genuine public value.

Keywords: Government UX, Behavioral analytics, Citizen Experience, Human-centered design, Public sector digital transformation, Net promoter score

INTRODUCTION

Digital government services have expanded rapidly over the past two decades, driven by policy commitments to efficiency, citizen empowerment, and inclusive service delivery (Carter & Bélanger, 2005). Yet despite significant investment in portal development and periodic redesign, citizen satisfaction with government websites consistently lags behind equivalent commercial services. The disconnect between institutional intent and user experience is not primarily a technical problem. Infrastructure and design standards in the public sector have improved

Received February 20, 2026; Revised April 3, 2026; Accepted April 18, 2026; Available online July 20, 2026

considerably. The gap is methodological: the tools most commonly used to diagnose service failure are poorly suited to detecting it.

The dominant evaluation model in public administration is the periodic satisfaction survey, typically distributed to a random sample of the citizen population. This approach has well-documented limitations in the commercial sector (Reichheld, 2003), but its shortcomings are particularly acute in public services. Citizens do not interact with government websites in the same way they browse retail or social media platforms. They arrive with specific tasks — filing a form, locating a permit requirement, confirming a deadline — under conditions of time pressure and often without viable alternatives. A random sample drawn from the general population will include individuals who have never attempted these tasks, whose responses introduce systematic noise into any aggregate satisfaction measure.

This paper reports on a study conducted in partnership with a mid-sized Japanese municipality to redesign its official web portal. The methodological contribution lies in its departure from random sampling in favor of behavioral targeting: the use of a real-time analytics platform to identify users demonstrating active, task-oriented engagement with the portal — operationally defined by scroll depth — and to deliver satisfaction instruments exclusively to this segment. The approach draws on the concept of ecological validity in Human Factors research (Schmuckler, 2001), which holds that behavioral measurement is most informative when conducted within the context of actual task performance rather than through retrospective or hypothetical recall.

The results are presented in three layers: an aggregate service quality assessment, a granular navigation failure analysis, and a journey map decomposition of prioritized usability failures. Taken together, these findings produce a diagnostic picture of previously invisible service failures that had persisted undetected through years of conventional survey administration. A targeted redesign intervention, anchored in the failure dimensions identified through behavioral targeting, produced measurable post-launch improvements confirmed by subsequent citizen feedback.

The paper concludes by proposing a replicable framework — the Behavioral Targeting Evaluation Model (BTEM) — as a methodological contribution for public sector Human Factors practice. The framework's applicability extends to any public agency operating digital services where citizen behavior is task-driven, alternatives are limited, and evaluation budgets are constrained.

THEORETICAL BACKGROUND

Measuring Usability: Between Laboratory Precision and Field Validity

The measurement of usability has been extensively theorized in the Human Factors and HCI literature. Nielsen (1994) established foundational usability dimensions — learnability, efficiency, memorability, error rates, and satisfaction — that continue to structure evaluation practice. Brooke's (1996) System Usability Scale (SUS) operationalized subjective satisfaction in a standardized ten-item instrument now applied across industrial, commercial, and public sector contexts. Both frameworks, however, presuppose measurement conditions that are rarely replicated in real-world public service environments.

Hornbæk's (2006) systematic review of usability measurement practices identified a persistent tension between laboratory control and ecological validity. Laboratory studies achieve precision at the cost of representativeness: task scenarios are constructed, user samples are recruited, and the emotional and cognitive context of authentic use is stripped away. Field studies preserve context but introduce confounds that make causal inference difficult. The challenge for public sector UX evaluation is to resolve this tension at scale, with limited resources and heterogeneous user populations.

Frøkjær, Hertzum, and Hornbæk (2000) demonstrated empirically that the three core usability dimensions — effectiveness, efficiency, and satisfaction — are weakly correlated in practice. A user who completes a task may nonetheless report low satisfaction; a user who reports satisfaction may have completed their task through a highly inefficient path. This finding has significant implications for survey-based evaluation: self-reported satisfaction alone provides an incomplete and potentially misleading picture of actual service quality. Behavioral data — what users did, not what they say they felt — must anchor any credible diagnostic evaluation.

Ecological Validity and the Behavioral Turn in UX Research

The concept of ecological validity, originating in Brunswik, (1956) probabilistic functionalism, entered cognitive science and later Human Factors research as a criterion for evaluating whether experimental conditions produce data generalizable to real-world performance. In UX research, ecological validity implies that measurement should occur in the context of actual task performance, with genuine stakes, under the time and cognitive constraints that users actually face (Schmuckler, 2001).

Behavioral analytics — the systematic tracking of user interactions with digital interfaces — provides a natural methodological bridge between laboratory precision and field validity. Rather than constructing artificial task scenarios or relying on retrospective recall, behavioral analytics captures what users actually do in situ. Kohavi and Thomke (2017) have argued that online behavioral data, properly collected, constitutes the highest-fidelity source of user insight available to practitioners. Their work on controlled online experiments in commercial settings has direct parallels in the public sector, where the same principles of behavioral measurement can be applied to diagnose service failure without the resource demands of formal usability testing.

The present study extends this logic through a specific application: using scroll depth as a behavioral proxy for purposeful, task-motivated engagement. This operationalization is grounded in the well-established finding that users who scroll through a significant portion of a page are disproportionately likely to be pursuing a specific informational goal (Nielsen, 1994). By restricting survey delivery to this segment, the study effectively filters out the ambient noise that contaminates random-sample evaluation, producing a dataset anchored in the experiences of users whose failure to find what they need constitutes a genuine service failure.

E-Government UX and Citizen Satisfaction

The literature on e-government adoption and citizen satisfaction has historically prioritized adoption barriers over diagnostic usability analysis. Carter and Bélanger (2005) identified trust, compatibility with existing values, and perceived ease of use as primary determinants of e-government service utilization. Verdegem and Verleye (2009) argued for a user-centered approach to e-government evaluation that foregrounds subjective user experience alongside objective transaction completion. Welch, Hinnant, and Moon (2005) established empirical linkages between e-government satisfaction and broader trust in government institutions — a finding that elevates website UX from a technical concern to a governance issue.

Despite this theoretical grounding, practical UX evaluation in local government remains underdeveloped. Most municipalities lack the technical capacity for continuous usability monitoring and rely instead on periodic, resource-intensive redesign projects informed by practitioner intuition or generic usability heuristics. The result is a pattern in which service failures persist across successive redesign cycles, not because they are intractable but because the evaluation instruments in use are not designed to detect them.

The Net Promoter Score (NPS), originally developed by Reichheld (2003) as a single-question loyalty metric for commercial services, has been adapted for public sector satisfaction measurement in a number of contexts. Its simplicity facilitates large-scale deployment; its single-question format reduces survey fatigue; and its –100 to +100 scale produces an intuitive aggregate measure of citizen sentiment. Crucially, NPS is sensitive to the emotional quality of the service encounter rather than merely to task completion, making it particularly informative when applied to the segment of users most likely to have experienced the service's failure modes directly.

METHODOLOGY

Study Context

The study was conducted in partnership with the digital transformation office of a mid-sized Japanese municipality responsible for serving a diverse citizen population across a broad range of administrative functions. The municipality had received repeated citizen feedback indicating difficulties with its official web portal — specifically, complaints about poor information findability, confusing navigation, and inadequate mobile performance. Prior attempts to address these complaints had relied on post-event satisfaction surveys administered to random samples of portal visitors, producing aggregate satisfaction data insufficient to identify specific failure points or prioritize design interventions.

Behavioral Targeting Protocol

A real-time behavioral analytics platform was integrated into the web portal's infrastructure to track visitor interactions at the session level. Scroll depth was selected as the primary behavioral trigger for survey eligibility on the grounds that users who scroll beyond a defined threshold within a page are

demonstrably engaged in purposeful content-seeking behavior. This criterion operationalizes the concept of ecological validity by restricting measurement to users whose interaction with the portal reflects genuine task motivation, rather than incidental browsing.

Upon reaching the scroll depth threshold, eligible users received an NPS survey instrument embedded within the portal interface. The survey was designed to minimize cognitive interruption while capturing three dimensions of the service encounter: overall satisfaction (NPS), task outcome (whether the user found what they were looking for), and qualitative elaboration via open-ended comment. Survey delivery was calibrated to achieve a statistically robust response volume without over-deployment.

Journey Map Analysis

In parallel with the NPS survey, session-level behavioral data was aggregated to construct journey maps representing the modal paths taken by users with different stated visit purposes. Journey map analysis followed the framework proposed by Rosenbaum, Losada Otalora, and Contreras Ramírez (2017), decomposing the service encounter into arrival, navigation, content engagement, and exit phases. Failure points within each phase were identified by cross-referencing behavioral data (page exits, back-navigation events, search query patterns) with qualitative feedback from the NPS open-comment field.

Post-Redesign Validation

Following the identification of priority failure dimensions, a targeted redesign was implemented and launched. Post-launch validation was conducted using the same behavioral targeting protocol, allowing direct comparison of pre- and post-redesign satisfaction indicators and navigation success rates.

RESULTS

Overall Service Quality Assessment

The behavioral targeting methodology produced a dataset anchored in the experiences of task-motivated portal users. The overall NPS across this segment was –52.3, indicating a strongly net-negative citizen experience. This figure is consistent with benchmark data for government web services (Verdegem & Verleye, 2009) but represents a significantly more adverse outcome than prior random-sample surveys had suggested — illustrating the suppression effect that non-task users exert on aggregate satisfaction measures when included in evaluation samples.

Approximately 25% of users who demonstrated purposeful engagement with the portal reported that they had failed to reach their intended destination page. This navigation failure rate, invisible to random-sample evaluation by definition, constitutes the most significant single finding of the study. A user without a destination cannot fail to find it; by restricting the evaluation sample to users with destinations, the behavioral targeting protocol surfaces a failure mode that conventional surveys structurally preclude.

The NPS gap between smartphone users and PC users was measured at -3.8 points, with smartphone users reporting systematically lower satisfaction. Given the proportion of mobile visitors to the portal, this gap carries material implications for redesign priorities. Users accessing public service portals via smartphone are, by the nature of the device and the context of use, disproportionately likely to be in time-constrained, task-urgent situations — precisely the conditions under which poor information architecture imposes the greatest experiential cost.

Journey Map Findings

Journey map analysis identified three dimensions as the primary targets for redesign intervention, ranked by the frequency and intensity of failure signals in the combined behavioral and qualitative dataset.

Information findability emerged as the dominant failure dimension, accounting for the largest share of negative qualitative feedback and correlating with the highest rates of back-navigation and portal exit without task completion. Users consistently reported difficulty locating specific procedural information — permit requirements, application deadlines, welfare eligibility criteria — within the portal's categorical navigation structure.

Visual clarity of page-level design was identified as the second priority dimension, with users reporting cognitive effort in parsing dense, text-heavy pages with inconsistent visual hierarchies. Norman's (2013) concept of discoverability — the degree to which design affords immediate comprehension of available options — was consistently violated on the pages generating the highest negative feedback density.

Content comprehension, encompassing the linguistic accessibility and structural clarity of informational content, emerged as the third priority dimension. Users attempting to complete administrative tasks reported difficulty understanding procedural instructions written in bureaucratic register and presented without navigational scaffolding.

Post-Redesign Validation

The targeted redesign prioritized three structural changes: migration to a smartphone-first interface architecture, integration of AI-assisted search functionality to supplement categorical navigation, and simplification of content structure on high-traffic procedural pages. Post-launch behavioral data and NPS survey results confirmed improvements across all three identified failure dimensions, with citizen feedback specifically referencing improvements in visual clarity and search performance.

DISCUSSION

The Behavioral Targeting Evaluation Model (BTEM)

The findings support the formalization of a Behavioral Targeting Evaluation Model (BTEM) as an alternative to random-sample surveys for public sector UX evaluation. The model rests on three operational principles. First, survey eligibility should be determined by behavioral criteria that proxy for task

motivation, ensuring that evaluation data reflects the experiences of users whose service outcomes matter most to institutional performance. Second, behavioral and attitudinal data should be integrated, with session-level behavioral logs providing the contextual scaffolding for interpreting NPS responses and qualitative feedback. Third, findings should be translated directly into prioritized design interventions, with post-launch validation conducted using the same behavioral targeting protocol to enable meaningful before-and-after comparison.

Methodological Implications

The navigation failure rate identified in this study — approximately 25% of purposeful visitors failing to reach their intended destination — would not have been detectable through random-sample evaluation. This is not a marginal point. It implies that a significant class of service failures is systematically invisible to the evaluation instruments currently standard in public administration. If 25% of task-motivated users of a commercial service failed to complete their intended transactions, the operator would be expected to identify and address the cause within a short operational window. The structural opacity of random-sample evaluation allows equivalent failure rates to persist indefinitely in public services.

The commercial UX literature has addressed this issue through conversion rate optimization and funnel analysis, but these tools are rarely applied in public sector contexts, where success metrics are poorly defined and behavioral data is infrequently collected at the session level. The BTEM framework represents an adaptation of behavioral measurement principles to the specific constraints — limited technical capacity, heterogeneous user populations, absence of conversion events — of local government digital operations.

Limitations

This study is based on a single municipal case in Japan, and the generalizability of specific findings — NPS values, failure rates, priority dimensions — to other national or administrative contexts cannot be assumed. The scroll depth trigger is a proxy for task motivation rather than a direct measure and may misclassify some non-purposeful users as task-motivated. Future research should investigate alternative behavioral proxies and their comparative validity across service types and cultural contexts.

CONCLUSION

This study presents empirical evidence that conventional random-sample satisfaction surveys systematically fail to detect the most consequential usability failures in government web portals — specifically, the navigation failures experienced by citizens with active task goals. A behavioral targeting methodology, grounded in ecological validity and implemented through a real-time analytics platform, produced diagnostic findings that had remained invisible through years of conventional evaluation. A subsequent targeted redesign confirmed the actionability of these findings.

The proposed BTEM framework provides a replicable methodological alternative for public sector UX practitioners operating under resource constraints, with direct applicability to any digital government service in which citizen interactions are task-driven. As governments worldwide accelerate the migration of public services to digital channels, the ability to identify and address genuine service failures — not the averaged perceptions of heterogeneous populations — becomes a precondition for democratic public service delivery. Behavioral targeting offers a methodologically grounded path toward that standard.

REFERENCES

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), Usability evaluation in industry (pp. 189–194). Taylor & Francis.

Brunswik, E. (1956). Perception and the representative design of psychological experiments (2nd ed.). University of California Press.

Carter, L., & Bélanger, F. (2005). The utilization of e-government services: Citizen trust, innovation and acceptance factors. Information Systems Journal, 15(1), 5–25.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 345–352). ACM.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies, 64(2), 79–102.

Kohavi, R., & Thomke, S. (2017). The surprising power of online experiments. Harvard Business Review, 95(5), 74–82.

Nielsen, J. (1994). Usability engineering. Academic Press.

Norman, D. A. (2013). The design of everyday things (Revised and expanded ed.). Basic Books.

Reichheld, F. F. (2003). The one number you need to grow. Harvard Business Review, 81(12), 46–54.

Rosenbaum, M. S., Losada Otalora, M., & Contreras Ramírez, G. (2017). How to create a realistic customer journey map. Business Horizons, 60(1), 143–150.

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. Infancy, 2(4), 419–436.

Verdegem, P., & Verleye, G. (2009). User-centered e-government in practice: A comprehensive model for measuring user satisfaction. Government Information Quarterly, 26(3), 487–497.

Welch, E. W., Hinnant, C. C., & Moon, M. J. (2005). Linking citizen satisfaction with e-government and trust in government. Journal of Public Administration Research and Theory, 15(3), 371–391.