NAVQUESTS
HUMAN SYSTEMS INTELLIGENCE

A RESEARCH FOUNDATION

The Research Foundation of Human Systems Intelligence

What the method measures, the science beneath it, and what we can honestly claim.

A system can meet every requirement set for it and still fail the people it was built to serve. The site loads. The form validates. The policy is satisfied. And a person who was eligible, capable, and trying still does not get through. Human Systems Intelligence™ measures that gap.

ONE LINE DRAWN FIRST

The reasoning behind the method is set out here. The instrument that performs the measurement is not. The scoring rubric, the decision logic, and the calibration record are proprietary. What follows is the thinking they are built on.

01

What the method measures

Most evaluation asks whether a system is usable. Human Systems Intelligence asks a narrower and harder question. Can a defined population accomplish the system's intended purpose, at a level comparable to the broader public? Three commitments follow from that question.

The system is the object of evaluation, not the person. When someone cannot complete a task, the finding is recorded against the system. Human experience is the evidence; the system is what is being judged.

The measure is outcome, not interaction. The question is not whether a button was easy to click. It is whether the person achieved what they came for. A system can be pleasant to use and still leave its user without the thing they needed.

The comparison is a population, not an average user. The method measures a specific, named group against a reference, because a system that serves a default user can quietly fail the people who actually depend on it.

02

Why the population is measured, not assumed

The reference point most design relies on is an average user who, on inspection, rarely exists. National skill surveys show that people spread widely across the very capabilities consequential tasks demand: reading, numeracy, and fluency with digital procedures. The OECD's adult-skills program finds that across member countries, nearly a third of adults score at the lowest level of adaptive problem solving, able to manage only simple problems that hold still while they work (OECD, Survey of Adult Skills, 2023). An average reader assembled from that spread is a statistically uncommon person.

In the United States the spread is wider still: about a quarter of adults score at or below the lowest level in literacy and numeracy — proficiencies that have slipped since the previous survey — and immigrant adults score lower again, which is why the method holds such groups apart rather than averaging them into one (National Center for Education Statistics, 2023).

Human Systems Intelligence measures its reference group rather than assuming it, and is disciplined about what counts as a national comparison — and about what counts as a comparison over time, since the adaptive-problem-solving domain is new to the 2023 cycle and, unlike literacy and numeracy, cannot be read against the earlier one. Survey-methods research shows that probability-based national panels track government benchmarks far more closely than online opt-in samples, and that demographic weighting does not reliably repair the gap (Pew Research Center, 2018, 2023). Each group's performance is read against a measured baseline; where one group falls below another, the gap is treated as a signal rather than noise, because variance points to a barrier that affects some people and not others.

03

The four dimensions of difficulty

The method examines four dimensions, each grounded in established research.

Cognitive load. Working memory holds only a few items at once. When a task demands more, people do not slow down; they fail, guess, or leave. This is basic cognitive architecture, not a finding borrowed from a classroom.

Language. Whether a population can read the words on the path, recognize the terms without looking them up, and tell what to do next. Two decades of

health-literacy and plain-language work place responsibility for comprehension with the institution that writes, not the reader (Nutbeam, 2000; AHRQ, 2024).

Procedural knowledge. What the task assumes a person already knows and never teaches: which document, which order, which step belongs to another institution. Cognitive science treats this know-how as a distinct kind of knowledge, built only through practice and largely unspeakable once acquired (Anderson, 1982; Polanyi, 1966), which is why the people who have it cannot see what it contains.

Time cost. The demands a task places on a person, counted plainly: the trips, forms, calls, and verifications that stand between them and the outcome.

Public-administration research on administrative burden shows these learning and compliance costs decide whether eligible people reach a benefit at all (Herd & Moynihan, 2018; Brodkin & Majmundar, 2010). The same tradition has since been formalized into a practice of structured “sludge audits,” which catalogue the unnecessary frictions a process imposes and sort them by the kind of cost they create — search, evaluation, action, and psychological (Sunstein, 2021; Shahab & Lades, 2021). Human Systems Intelligence shares that concern but parts from it in method: rather than inventorying frictions in the abstract, it measures a named population against a reference row, and lets a single severe barrier cap the result. Every barrier the method finds is then labeled imposed or inherent. Imposed difficulty is created by the design and is not required by the task, so it can be removed. Inherent difficulty belongs to the task itself and can only be restructured or supported. That single distinction decides what each recommendation should be.

04

Why one severe barrier caps the score

A system is a sequence. A person stopped at one point never reaches the rest, however well built the rest may be. For that reason a single severe barrier caps the overall score, rather than being averaged away by everything that works.

Field studies of benefits enrollment show how literally this can hold. In one large county, a single procedural step — a missed eligibility interview — accounted for roughly one in three denials, against about six percent denied for actual ineligibility (Code for America). The wall was not the applicant's capability; it was one gate, and it decided the outcome.

This rule was not adopted for convenience. The same structure appears independently in three fields that do not cite one another: the overload cliff in cognitive-load research, where capacity is exceeded and performance

collapses (Sweller); the sequential threshold model in benefits research, where missing one gate means never arriving (Kerr, 1982); and the single-point-weakness concept in formal risk analysis, where one failure point is escalated on its own regardless of strength elsewhere (the Veterans Health Administration's HFMEA; DeRosier et al., 2002). Three independent discoveries of the same shape are strong grounds for building the rule in.

Severity is assigned by rule, not by judgment, and the reason is evidence. Usability evaluation has measured its own severity ratings and found that independent evaluators rarely agree on which problems are severe (Nielsen, 1994; Hertzum & Jacobsen, 2003). Where the same classification task has been run both ways, rule-based ratings reach near-perfect agreement while expert judgment reaches far less, a reliability of 0.83 to 0.90 by rule against 0.63 to 0.67 by judgment (Snyder et al., 2007, adapting the NCC MERP index). A method meant to be trusted by others cannot rest on a single rater's feel.

05

Why the failure is usually invisible

The hardest part of this problem is that the failure leaves no record. A person who cannot get through rarely complains. They stop, and the system logs nothing: no error, no denial, no trace except absence. Public-administration research calls this administrative exclusion, where a departure produced by process is recorded in the system's own ledger as a voluntary one (Brodkin & Majmundar, 2010).

The scale is documented. Across means-tested benefits, the share of eligible people who never claim what they are entitled to has been measured between thirty and fifty percent (van Oorschot, 1991). Field experiments have moved real claiming behavior with presentation and prompting alone, evidence that the barrier sits in the process rather than in the person (Bhargava & Manoli, 2015).

The same pattern recurs in present-day service research. When a benefits application asked for income in a way that invited guessing, applicants understated it and were screened out until the question was rebuilt; and submitting verification documents earlier, or redesigning the interview step, measurably raised approvals (Code for America). Each barrier sat in the process rather than the person — and each stayed invisible until someone went looking for the people who had already left.

This is why the method does not rely on a client's existing analytics to find these people. They are not in the data. Analytics can only measure those who arrived and stayed long enough to be recorded, and the people the method exists to find did neither. Human Systems Intelligence works as a prediction

instrument. It identifies where a defined population will fail by walking the system against the measured baseline, and the client's behavioral data, where it exists, then tests the prediction rather than generating it.

06

What we can honestly claim

The strength of the evidence is not uniform, and the method says so. Every claim it makes carries one of three labels.

TIER	WHAT IT MEANS
Direct	Measured in the method's own territory, real interfaces and real institutional systems, or grounded in domain-general cognitive architecture.
Mechanistic extension	Established in instructional research and applied here through the shared working-memory mechanism. The laboratory shows why an effect occurs; the method applies that why to a new setting.
Boundary-conditioned	A real effect with documented limits the auditor must respect when applying it.

Marking these differences is deliberate. A sound prediction is never presented as a proven fact, because a careful reader will eventually ask which is which, and the answer should already be on the page. The space between what is proven and what is predicted is not a weakness to conceal. It is the company's research roadmap, and every engagement tests its predictions against real behavior, building evidence no one else is positioned to gather.

IN CLOSING

Human Systems Intelligence rests on this foundation: general cognitive architecture at the base, field outcomes from administrative-burden research at the top, and the mechanisms that connect them in between, each marked for how firmly it stands. The instrument that turns the reasoning into a score is proprietary and is deliberately not described here. What is described is the question underneath all of it, the one worth asking of any system built to serve people.

Can the people it was built for actually use it?

SELECTED RESEARCH

COGNITIVE LOAD AND PROCEDURAL KNOWLEDGE

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review,* 89(4).

Polanyi, M. (1966). *The Tacit Dimension.* University of Chicago Press.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist,* 38(1).

Mayer, R. E. (2009). *Multimedia Learning* (2nd ed.). Cambridge University Press.

LANGUAGE AND PLAIN COMMUNICATION

Nutbeam, D. (2000). Health literacy as a public health goal: A challenge for contemporary health education and communication strategies into the 21st century. *Health Promotion International,* 15(3).

Brach, C. (Ed.) (2024). *AHRQ Health Literacy Universal Precautions Toolkit* (3rd ed.). Agency for Healthcare Research and Quality.

ADMINISTRATIVE BURDEN, SLUDGE, AND NON-TAKE-UP

Herd, P., & Moynihan, D. P. (2018). *Administrative Burden: Policymaking by Other Means.* Russell Sage Foundation.

Brodkin, E. Z., & Majmundar, M. (2010). Administrative exclusion: Organizations and the hidden costs of welfare claiming. *Journal of Public Administration Research and Theory,* 20(4).

van Oorschot, W. (1991). Non-take-up of social security benefits in Europe. *Journal of European Social Policy,* 1(1).

Bhargava, S., & Manoli, D. (2015). Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. *American Economic Review,* 105(11).

Bhargava, S., Loewenstein, G., & Sydnor, J. (2017). Choose to lose: Health plan choices from a menu with dominated options. *Quarterly Journal of Economics,* 132(3).

Sunstein, C. R. (2021). *Sludge: What Stops Us from Getting Things Done and What to Do about It.* MIT Press.

Shahab, S., & Lades, L. K. (2021). Sludge and transaction costs. *Behavioural Public Policy.*

FIELD EVIDENCE IN TRANSACTIONAL SYSTEMS

Code for America (2019). *Bringing Social Safety Net Benefits Online* (Benefits Enrollment Field Guide).

Code for America (2023). Think Big, Start Small: How Implementing Flexible Interviews Improves Benefit Delivery.

SEVERITY, RELIABILITY, AND THE CAP

Kerr, S. (1982). Deciding about supplementary pensions: A provisional model. *Journal of Social Policy,* 11(4).

Nielsen, J. (1994). Severity ratings for usability problems. In *Usability Inspection Methods.* John Wiley & Sons.

Hertzum, M., & Jacobsen, N. E. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction,* 15(1).

Snyder, R. A., et al. (2007). Reliability evaluation of the adapted NCC MERP index. *Pharmacoepidemiology and Drug Safety,* 16(9).

DeRosier, J., Stalhandske, E., Bagian, J. P., & Nudell, T. (2002). Using Health Care Failure Mode and Effect Analysis. *Joint Commission Journal on Quality Improvement,* 28(5).

POPULATION MEASUREMENT

OECD (2024). *Do Adults Have the Skills They Need to Thrive in a Changing World? Survey of Adult Skills 2023.*

National Center for Education Statistics (2023). *PIAAC Cycle 2: Highlights of U.S. National Results.*

Pew Research Center (2023). *Comparing Two Types of Online Survey Samples;* and (2018) *For Weighting Online Opt-In Samples, What Matters Most?*

Agresti, A., & Coull, B. (1998). Approximate is better than ‘exact’ for interval estimation of binomial proportions. *The American Statistician,* 52.

Evidence labeled by tier in the manner described in Section 6. A fuller research series, document by dimension, stands behind this summary. The scoring instrument and calibration record are proprietary and are not part of this foundation.

Because working systems should work for people.