Attention Beyond Activity

The Gap Between Action and Attention.

GrayPass Labs · tools@graypass.org

GPL-TR-2026-03 · 2 August 2026

Abstract. Web systems routinely report time-on-page, cursor activity, clicks, hovers, and similar signals as if they measured a person's attention. This report audits that practice on a public dataset of 2,909 crowdsourced web-search sessions with self-reported attention labels (the Attentive Cursor Dataset), of which 2,775 pass a minimal validity filter. We organize twelve common engagement proxies on a construct ladder from exposure to self-reported attention and measure how they relate to each other, to the participant's own attention report, and to context. The proxies disagree with one another: the median absolute pairwise rank correlation is 0.22, and raw dwell correlates at $\rho = -0.84$ with the fraction of dwell that is active, so 81% of sessions land in discordant quadrants under a median split. Using the dataset's own focus, blur, and gap structure, we find that a median 51% of recorded dwell is idle or blurred, and 66% of sessions have at least 30% of dwell unattributable to observed, focused activity. Every proxy's association with self-reported attention is weak ($|\rho| \leq 0.063$); a regularized multivariate model reaches AUC 0.558. Relationships shift across ad categories, including sign flips, and deterministic counterfactual transformations inflate headline metrics up to 59-fold without adding any evidence of attention. We distill the findings into an Attention-Claims Checklist for engagement reporting.

Keywords: attention measurement; engagement metrics; construct validity; dwell time; mouse cursor tracking; implicit feedback; metric gaming

---

1 Introduction

A large part of the modern web economy is billed, ranked, and evaluated in units of borrowed psychology. Advertising is sold against impressions that were "viewable," content is optimized for "engagement," and product decisions are justified by "time spent." Each of these quantities is computed from interaction logs: timestamps, cursor coordinates, clicks, focus changes. Each is routinely described, in dashboards and in papers, with the vocabulary of attention. The slide from log arithmetic to mental-state language is convenient, because attention is the thing people actually care about: Simon's observation that a wealth of information creates a poverty of attention is the standard motivation for treating user attention as the scarce resource that platforms allocate and advertisers buy [1]. But the slide is an inference, not an identity. A timestamp interval measures a timestamp interval. Whether it also measures attention is an empirical question about construct validity, of exactly the kind psychometrics has asked about test scores for seventy years [2].

The gap between what is recorded and what is claimed has a specific structure, and it is worth stating before any data. An interaction log supports a hierarchy of increasingly ambitious readings. At the bottom, a page was rendered and a logger was attached: exposure. Timestamps span an interval: recorded presence. Input events fired during that interval: activity. The events were continuous and occurred while the page held focus: continuity. The cursor spent time near the object of interest: orientation. The person, asked directly, says they attended: self-reported attention. Above all of these sit the constructs that reporting language most often invokes and that logs least support, comprehension, memory, interest, and intent. Each step up this ladder requires an inference that the step below cannot supply on its own. Industry measurement acknowledges the bottom of the ladder explicitly, as

when viewability standards demand that an ad's pixels be on an in-focus tab for one continuous second before an impression may even be counted as having an opportunity to be seen [3], but reporting practice routinely skips from the middle of the ladder to the top in a single caption. The question this report asks is how much support the middle rungs actually give one another and the rung above them, when all of them are computed on the same sessions.

We ask the question quantitatively, on public data, for the activity proxies most commonly used on the web. The setting is deliberately favorable to the proxies. The Attentive Cursor Dataset [4] records 2,909 crowdsourced sessions of a transactional search task on a search engine results page (SERP) that carries an advertisement, with full client-side event logs (cursor moves, clicks, scrolls, focus and blur events, and per-event cursor distances to the ad) and, crucially, a criterion that most production logs never have: each participant's own report, on a 1 to 5 scale, of whether they paid attention to the ad. Self-report is an imperfect, ordinal criterion, and we treat it as such; it is not ground-truth cognition. But it is an independent measurement of the construct that engagement metrics claim to track, collected from the only witness with direct access. If activity proxies were even rough measures of attention, this is a setting where they should agree with each other and correlate with the report: one page, one task, one ad, desktop mice only, and a dataset published precisely because cursor behavior carries attention signal [4], [5].

What we find is a systematic divergence, which we call the proxy trap. First, the proxies disagree with one another. Rank correlations among twelve standard proxies are mostly modest (median absolute off-diagonal $\rho = 0.22$; 70% of pairs below 0.4), and the two most reported quantities, raw dwell and the active share of dwell, are strongly anti-correlated ($\rho = -0.84$): the longer a session's recorded dwell, the smaller the fraction of it that shows any activity at all. Second, raw time overstates observed engagement by construction. Using only the dataset's own blur and focus events plus inter-event gaps, the median session has 51% of its dwell idle or blurred; two thirds of sessions have at least 30% of dwell unattributable to focused activity; and a single left-open tab accumulates 63 days of "dwell," 97% of the corpus total, while producing 24 seconds of active time. Third, the criterion association is weak everywhere: no single proxy exceeds $|\rho| = 0.063$ against self-reported attention, and a regularized multivariate model on all twelve proxies reaches AUC 0.558 against a high/low split of the report, against 0.792 for predicting the mechanical act of clicking the ad. Fourth, the little structure that exists is context-bound: the strongest proxy-attention association flips sign across ad categories, and out-of-context model performance is erratic even when its average looks stable. Fifth, the headline metrics are arithmetically gameable: deterministic transformations of frozen logs, clearly labeled counterfactual, inflate dwell 13-fold, mousemove counts 59-fold, and cursor path length 3-fold, while every richer signal stays fixed.

None of this says that interaction logs are useless, and none of it says that attention is unmeasurable. It says that activity is not a self-validating measure of attention: the proxies do not converge with each other, associate only weakly and unstably with the person's own report, and can be moved arbitrarily by transformations that add no evidence about a person at all. The correct response, we argue, is the response measurement theory has always prescribed: state which rung of the construct ladder a metric actually occupies, report the cheap invariance checks that would expose inflation, and reserve attention language for measurements that have earned it [2], [6].

This report makes six contributions:

1. A construct ladder that separates exposure, recorded presence, activity, continuity, orientation proxy, and self-reported attention, with comprehension and intent explicitly off the ladder, as a vocabulary for stating what an engagement metric measures (Figure 1).
2. A quantitative disagreement audit among twelve engagement proxies computed on the same 2,775 public sessions, including rank correlations, percentile disagreement, and median-split discordance rates (Figure 2).
3. A direct measurement of raw-versus-active time inflation using the dataset's own focus/blur events and inter-event gaps, with sensitivity to the idle threshold (Figure 3, Figure 4).

4. A criterion-validity table relating each proxy to self-reported attention (Spearman ρ with participant-bootstrap confidence intervals) and to ad clicks (AUC), plus a participant-disjoint multivariate model, reported with its honest, modest performance.
    1. A context-dependence and transfer analysis showing that proxy-attention and proxy-proxy relationships move across ad position, ad type, ad category, and device class, including point-estimate reversals, and that out-of-context predictive transfer is unreliable group by group (Figure 5, Figure 6).
5. Deterministic counterfactual stress tests quantifying how far bounded, attention-free transformations move each headline metric (Figure 7), and an Attention-Claims Checklist that operationalizes the findings for anyone reporting engagement numbers.

What this paper does not claim. This report measures logged activity and one self-report item; it does not measure comprehension, memory, interest, purchase intent, or any private mental state, and no analysis here supports claims about them. Self-reported attention is an ordinal, retrospective, imperfect criterion, not ground truth about cognition. All analyses are observational and correlational; nothing here identifies causal effects. The counterfactual stress tests in Section 5.6 are arithmetic demonstrations on frozen logs and describe no human behavior. The scope is one public desktop SERP dataset with one observation per participant; we quantify, rather than assume away, the limits this places on generality.

2 Related work

2.1 Dwell time and implicit feedback

Reading time entered the interaction literature as an implicit interest indicator: Claypool et al. instrumented a browser, collected explicit page ratings from about 80 users over 2,500 pages, and found that time on page and scrolling correlated with explicit interest while raw mouse click counts did not [7]. The nuance arrived quickly. Kelly and Belkin, in a fourteen-week naturalistic study of seven users' everyday information seeking, found no general, direct relationship between display time and usefulness; display times differed significantly by task and by user, which is precisely a construct-validity failure of dwell as a context-free relevance signal, documented two decades before the present audit [8]. Liu, White, and Dumais modeled page dwell at web scale with Weibull distributions and found dominant "negative aging": most pages are abandoned during an initial screening phase, so any dwell distribution mixes qualitatively different behaviors, screening and gleaning, in one number [9]. Subsequent applied work embraced dwell pragmatically rather than critically: Yi et al. used item-level dwell as a relevance proxy for recommendation, and found they had to engineer device- and context-specific normalizations before the signal was usable, which is applied acknowledgment that raw dwell does not carry one fixed meaning across contexts [10]. What this line of work generally does not do is decompose recorded dwell into attributable and unattributable parts using the browser's own focus events. Our contribution to it is that accounting, on public data, with sensitivity to the idle threshold, and joined to the same sessions' cursor, click, and orientation proxies so that the decomposition and the disagreement can be read together.

2.2 Cursor behavior as an attention proxy

A second tradition reads spatial attention from the pointer. Chen, Anderson, and Sohn reported strong gaze-cursor relationships in web browsing [11]; Rodden et al. catalogued eye-mouse coordination patterns on SERPs [12]; Guo and Agichtein inferred whether gaze and cursor were coordinated from cursor features alone [13]. The most careful study in this line, by Huang, White, and Buscher, is explicitly a caution: gaze-cursor alignment varies with time, behavior pattern, user, and task, and the authors "question the maxim that gaze is well approximated by cursor" [14]. Huang, White, and Dumais nonetheless showed that cursor data at scale usefully estimates result relevance and distinguishes good from bad abandonment [15]. The dataset we reanalyze descends from this

tradition: Leiva and Arapakis released it with attention labels [4], and companion papers predict attention to SERP ads from cursor movements with neural models [5], predict engagement with direct displays [16], study how much trajectory is needed for such predictions [17], price attention in ad auctions using cursor-based attention predictors [18], and demonstrate the privacy risks of the same signals [19].

Our position relative to this work is complementary and deliberately different in direction. Prior work largely asks how well engineered representations of cursor behavior can predict an attention label, and reports the achievable ceiling. We ask the converse audit questions that a metric consumer needs answered: how strongly do the simple proxies that dashboards actually report agree with each other on the same sessions, how much criterion validity does each carry on its own, how stable is any of it across contexts, and how far can each be moved by transformations that add no information about the person. To our knowledge no prior work has run this construct-validity audit, disagreement, inflation accounting, context instability, transfer, and counterfactual gaming, on one public dataset with a self-report criterion.

2.3 Engagement measurement and its standards

The engagement literature has long acknowledged that "engagement" is a multi-attribute construct rather than a single scale. O'Brien and Toms, from interview and survey work across search, shopping, webcasting, and gaming, define engagement as a quality of experience with attributes including attention, positive affect, novelty, and perceived control, unfolding through stages of engagement, disengagement, and re-engagement; nothing in that definition reduces to a timestamp interval [20]. Lehmann et al., analyzing behavioral metrics across a large portfolio of online services, found that popularity, activity, and loyalty metrics form several independent models of engagement rather than one, so that services cannot be ranked on a single engagement axis [21]. The synthesis by Lalmas, O'Brien, and Yom-Tov catalogs the measurement families available, self-report questionnaires, observational and physiological methods, and web analytics, and is explicit that the analytic family measures behavior at scale while negating access to motivation and cognition; the families are complements, not substitutes [22]. Industry measurement standards make the gap between exposure and attention explicit at the bottom of the ladder: the Media Rating Council's viewability guideline counts a desktop display impression as viewable when at least half its pixels are on an in-focus browser tab for one continuous second, and its authors are careful to present this as a minimal opportunity-to-see condition, not an attention claim; notably, the guideline also declines to accept a mouse-over as evidence of viewing [3]. Our focus/blur accounting in Section 5.2 measures how much recorded dwell fails even the in-focus precondition of that deliberately minimal standard.

2.4 Measurement theory

The vocabulary for what goes wrong here is seventy years old. Cronbach and Meehl's account of construct validity holds that a measure of an unobservable attribute is validated only by its position in a nomological network: the pattern of its correlations with other measures must match what the theory of the construct predicts, and no single correlation, least of all the measure with itself, can establish validity [2]. Campbell and Fiske operationalized this with the multitrait-multimethod matrix: convergent validation requires that different methods of measuring the same trait agree, and discriminant validation requires that a measure correlate more strongly with other measures of its own trait than with measures of different traits that merely share a method [6]. An engagement dashboard is a multitrait-multimethod matrix waiting to be read: dwell, activity counts, continuity fractions, and orientation proxies are different methods all claimed for the same trait, and Section 5.1 reads the matrix. The rank correlation we use throughout is Spearman's, introduced in the same era for exactly the situation we face, ordinal criteria and non-normal measurements [23]. Finally, Strathern's formulation of Goodhart's law, that when a measure becomes a target it ceases to be a good measure, is the one-sentence theory of Section 5.6: a metric whose value can be manufactured by parties with a stake in it will be manufactured, and the audit's job is to say how cheaply [24].

3 Dataset and rights

All empirical results in this report derive from the Attentive Cursor Dataset, released by Leiva and Arapakis [4] and distributed through a public GitLab repository. The dataset records 2,909 crowdsourced sessions, one per participant, of a transactional web-search task: each participant was shown a Google SERP for a product query on which one advertisement appears, in one of two positions (top-left, above the organic results, or top-right, in the sidebar), of one of two types (direct display or native), drawn from fourteen ad categories ranging from Computers & Electronics (805 sessions) to Real Estate (17). Participants were instructed to complete the search task from a desktop or laptop with a computer mouse, and the task page was the instrumented SERP snapshot, so the recorded interval brackets one page visit rather than a browsing journey. Interaction was captured client-side by the evtrack logger as a space-delimited event log per session: each row carries a millisecond timestamp, cursor coordinates, the browser event name, the XPath of the target element, and, for cursor events, a JSON field with Euclidean distances from the cursor to five reference points of the ad's bounding box plus an inside-the-box flag. A sidecar XML file per session records screen, window, and document sizes and the user-agent string; median screen width is 1,366 px, with deciles at 1,024 and 1,920 px, which is the basis of the device buckets used later. Two tab-delimited files provide the labels and stimuli: `groundtruth.tsv` with the self-reported attention score (1 to 5, where 1 denotes no attention) and whether the participant clicked the ad, and `participants.tsv` with demographics bins and the stimulus fields. After the task, participants answered an exit survey that included the attention item; they consented to the logging and to the sharing of their anonymized data, as described in the dataset paper [4]. The three files join exactly: all 2,909 log identifiers appear in both label files and have both a CSV log and an XML sidecar.

We follow the terms stated by the source. The repository's README requests citation of the dataset paper as the condition of use, and the paper itself is published open access under a CC BY license; we cite it wherever the data are used and have modified nothing in the distributed files, which we treat as read-only. Three properties of the data discipline everything downstream. First, there is exactly one observation per participant, so no within-person analysis is possible and every session-level statistic is also a participant-level statistic; our bootstrap procedures resample sessions, which is therefore a participant bootstrap. Second, the criterion is a single retrospective ordinal self-report about one ad, with all the biases self-report carries; we use it as an imperfect criterion, never as ground truth. Third, the context is a single desktop SERP task with a computer mouse (the logger's cursor field confirms no touch devices among the retained sessions' input), so nothing here generalizes automatically to feeds, mobile, video, or any other surface, a restriction we return to in Section 8. The dataset ships demographic bins (age, gender, education, income, country); consistent with our purpose, we perform no demographic profiling and use only the stimulus fields (ad position, type, category) and the XML screen size as context variables.

4 Construct ladder and methods

4.1 The ladder

Figure 1 shows the conceptual organization of the paper. Reading upward: **exposure** (content was rendered somewhere), **recorded presence** (a page was open and produced timestamps spanning an interval), **activity** (input events fired), **continuity** (activity while the page held focus, without long idle gaps), **orientation proxy** (the cursor was near or inside the target region), and **self-reported attention** (the person, asked directly, says they attended). Comprehension, memory, and intent sit above the ladder and are not measured here. The ladder is not a causal model; it is a declaration of measurement distance. Each rung is computed from the same log with strictly more assumptions than the rung below it, and each arrow is an inferential step that can fail silently. The empirical sections quantify three failure modes: adjacent rungs disagree (Section 5.1), lower rungs overstate

higher ones by construction (Section 5.2), and even the highest instrumented rungs barely track the self-report criterion (Section 5.3).

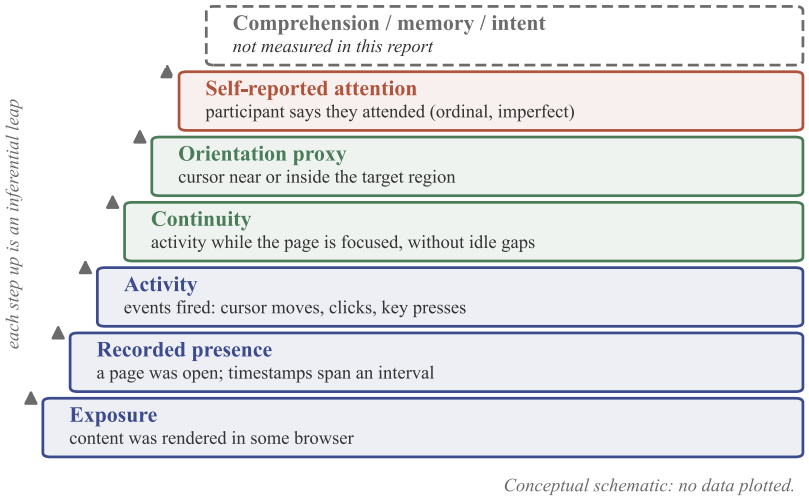

Figure 1: The construct ladder. Each rung is computable from interaction logs except the top two: self-reported attention requires asking the person, and comprehension, memory, and intent are not measured in this report at all. Every arrow is an inferential step that the audited proxies do not certify. Evidence class: conceptual schematic; no data plotted.

4.2 Event inventory and inclusion

Because proxies must be built from events that actually exist, the pipeline's first step enumerates the corpus (script `00_enumerate.py`; all counts cached in `results/event_inventory.json`). The 2,909 logs parse to 106,341 events after dropping one malformed row. The event vocabulary and coverage are: `mousemove` (47,020 events; present in 100% of logs), `scroll` (16,378; present in only 48.5% of logs), `mouseover` (14,804), `focus` (3,735; 66.1% of logs), `mousedown`/`mouseup` (3,726/3,710), `click` (2,892; 90.4% of logs), `blur` (2,878; 54.8% of logs), the lifecycle events `load`, `unload`, and `beforeunload`, and small counts of `keydown`, `keyup`, `select`, `contextmenu`, `resize`, `copy`, and touch events (19 events at most). Every log contains at least one parseable ad-distance JSON payload, so orientation proxies are computable corpus-wide.

Two consequences follow. First, we build no scroll proxy: scroll events appear in under half the logs, and a zero is uninterpretable because it confounds "did not scroll" with "page fit the viewport," which depends on window size, not the person. We state this explicitly rather than reporting a degenerate metric. Second, blur and focus events do exist at scale, so time out of focus is directly measurable; a session with no blur event has zero observed blurred time by definition, which is the correct reading because evtrack registers those events when they occur.

A session is included if it has at least 10 parsed events and strictly positive dwell (last timestamp after first). Of 2,909 sessions, 134 fail the event minimum, none fail positive dwell, and **N = 2,775** (95.4%) remain. All numbers below refer to these sessions unless stated otherwise. Among them, 66.4% report attention 4 or 5 (distribution 1: 263, 2: 488, 3: 182, 4: 1,039, 5: 803) and 25.0% clicked the ad (695 of 2,775).

4.3 Proxy definitions

Twelve proxies are computed per session, each deliberately simple and each in wide use; the point of the audit is precisely the summary statistics that reports and dashboards quote, not the best representation a model could learn. Table 1 gives the exact definitions; three constructions need commentary.

Active time follows the standard capped-gap construction: with inter-event gaps g_i in seconds and an idle threshold θ,

$$\text{active}(\theta) = \sum_i \min(g_i, \theta), \quad \text{idle} = \text{dwell} - \text{active}(\theta), \tag{1}$$

with $\theta = 2$ s primary and $\theta \in \{1, 5\}$ s reported as sensitivity. The construction charges each gap to activity up to θ and treats the remainder as idle; it makes no claim about what the person was doing, only about what the log shows. Its one parameter is honest and visible, unlike the session-timeout constants buried in most analytics pipelines, and Section 5.2 reports how conclusions move as it varies. Blurred time comes from a two-state machine: the session starts focused, a `blur` event switches the state to blurred, a `focus` event switches it back, repeated same-state events are no-ops, and the time between consecutive events accrues to the state in force. Because evtrack writes these events when the browser fires them, a session with no blur event has zero observed blurred time, which is the correct, conservative reading; 56.5% of included sessions contain at least one blur event, so out-of-focus time is not an exotic corner case but the majority experience. Active-focused time applies Equation 1 only to gaps that begin in the focused state; the unattributable share of a session is $1 - \text{active-focused} \,/\, \text{dwell}$, the fraction of recorded presence that observed, in-focus activity cannot account for. The orientation proxies use the logger's own geometry: the per-event JSON reports the Euclidean distance from the cursor to the ad box center ("middle"), and ad-near time accumulates capped gaps that begin at an event within 250 px of the center, a band chosen near the 75th percentile of the per-session minimum distances (median minimum distance 138 px) so that "near" is neither trivial nor empty. Time to first mouse event is measured from the first logged event to the first mouse-positioned event of any kind.

Proxy	Definition (per session)	Ladder rung
dwell	last minus first event timestamp, s	recorded presence
active time	$\sum_i \min(g_i, \theta)$ over inter-event gaps, $\theta = 2$ s (Equation 1)	activity
active fraction	active time / dwell	continuity
blurred fraction	time in blurred state (focus/blur machine) / dwell	continuity
mousemoves	count of `mousemove` events	activity
path length	$\sum$ Euclidean steps between consecutive `mousemove` positions, px	activity
moving fraction	time in inter-`mousemove` gaps ≤ 0.5 s / dwell	continuity
clicks	count of `click` events	activity
hovered elements	distinct XPath targets over `mousemove` + `mouseover`	activity
time to first move	first mouse-positioned event minus first event, s	activity
ad min distance	minimum cursor distance to ad-box center, px	orientation
ad-near time	capped gaps beginning within 250 px of ad center, s	orientation

Table 1: Proxy definitions. All proxies are computed from event types verified present in the corpus inventory; no scroll proxy is built because scroll events appear in only 48.5% of logs. Reanalysis of public data [4], N = 2,775.

Medians locate the corpus: dwell 24.0 s, active time 11.1 s, active fraction 0.57, blurred fraction 0.20, 12 mouse-moves, 1,064 px of path, 9 hovered elements, 1 click, 1.2 s to first mouse event, 138 px minimum ad distance, 1.3 s of ad-near time. Click count is nearly degenerate by design of the task (the 10th and 90th percentiles are both 1), which is itself informative: the most billed-against event in the industry has almost no variance to carry information here. 63.7% of sessions ever place the cursor inside the ad's bounding box.

4.4 Audit protocols

Disagreement (A2). We compute the full Spearman matrix among the twelve proxies, the mean absolute percentile-rank gap for each pair, and, for named key pairs, the median-split discordance rate: the share of sessions above the median on one proxy and at or below it on the other. Independent proxies would give 50% discordance; a pair measuring the same thing would give 0%. Rank statistics are used throughout because several proxies are heavy-tailed by orders of magnitude (dwell spans four decades) and because the criterion is ordinal; Pearson correlations on these variables would be dominated by a handful of extreme sessions. Median splits are crude by design: they mimic the "high engagement" segmentations that reporting actually performs, and any subtler segmentation inherits at least their disagreement.

Inflation (A3). Using Equation 1 and the focus/blur machine, we report distributions of the active/raw ratio, the blurred fraction, and the unattributable share, plus corpus-level person-time accounting, raw and with dwell winsorized at its 99th percentile (419 s). This analysis uses the dataset's own recorded events; it involves no model and no counterfactual. We report per-session distributions first and corpus sums second, because sums are exactly where unbounded dwell does its damage.

Criterion association (A4). Each proxy is related to self-reported attention by Spearman ρ (attention treated as ordinal) with 95% percentile bootstrap intervals over 2,000 resamples of sessions, which is a participant bootstrap because each participant contributes one session, and to the click by AUC with the same bootstrap. AUC is used for the binary criterion because it is rank-based and base-rate free, so the two columns of the criterion table are on comparable footing; an AUC below 0.5 is reported as is, since the direction of a proxy is part of what is under audit. A multivariate audit then asks how much the proxies carry jointly: L2-regularized logistic regression ($C = 1$) on all twelve proxies, heavy-tailed proxies log-transformed via $\log(1 + x)$ and all features standardized, under stratified 5-fold cross-validation, with performance computed on pooled out-of-fold predictions and its interval by participant bootstrap over sessions. With one observation per participant every row split is participant-disjoint, so no leakage across folds is possible by construction. Attention is dichotomized at 4 or above ("high," 1,842 sessions) versus 3 or below (933), the natural break in the label distribution; the cut is stated wherever the model is. We use a simple regularized linear model deliberately: the audit's question is what the standard summary statistics support, and a model expressive enough to re-derive trajectory information from them would answer a different question.

Context dependence (A5). The strongest proxy-attention association from A4 and one benchmark proxy-proxy relationship (dwell versus path length) are re-estimated within strata of ad position (2 levels), ad type (2), ad category (levels with at least 100 sessions, 8, plus a pooled "Other," 283 sessions), and a device bucket from XML screen width ($\{\leq 1280, 1281\text{–}1440, > 1440\}$ px; 781, 1,224, and 770 sessions). Each stratum gets its own bootstrap interval; dispersion across strata is summarized by the standard deviation and range of the stratum estimates. The design exploits the dataset's randomized stimulus: participants do not differ systematically across positions or types, so movement of a coefficient across those strata reflects the measurement's context dependence, not population differences.

Transfer (A6). For each context family we run leave-one-group-out transfer of the A4 models: train on all other groups, test on the held-out group, and compare against a baseline of 20 random participant-disjoint splits with test sets of exactly the held-out group's size, so any drop is attributable to the context shift rather than to test-set size or training-set reduction. We report per-group results, the mean drop with a percentile bootstrap interval over groups, and, because averages can hide asymmetric failures, the full spread of out-of-context performance including the count of groups at or below chance.

Counterfactual stress tests (A7). Three deterministic transformations are applied to frozen logs; they are labeled COUNTERFACTUAL throughout, are not human behavior, and add no observations. (i) Idle injection appends a duplicate of the final event $\Delta \in \{15, 30, 60, 120, 300\}$ s after the last timestamp: dwell grows by exactly Δ while active time can grow by at most $\min(\Delta, \theta) = 2$ s. (ii) Mousemove densification inserts $k \in$

$\{1, 3, 7, 15, 63\}$ points per consecutive mousemove pair, placed on the segment and deflected alternately ± 2 px perpendicular to it, so the trajectory stays within a 2 px envelope of the original polyline with endpoints and bounding box unchanged; the zero-deflection variant (pure duplication) inflates the count identically while provably leaving path length unchanged. (iii) Click splitting replaces each click with $k \in \{2, 3, 5, 10\}$ clicks inside the 150 ms window after it, never beyond the next event. The arithmetic identities in (i) were additionally validated by re-running the full feature extractor on 100 actually transformed logs (maximum discrepancy below 10^{-13} s). In this corpus no included session ends on a click, so click splitting changes dwell by exactly zero seconds.

Everything is deterministic under seed 42; every number quoted in this report is written by the analysis scripts to `results/results.json`, and `analysis/run_all.sh` regenerates all numbers and figures from the raw public data.

5 Results

5.1 Proxies disagree with each other

Figure 2 shows the Spearman matrix. The twelve proxies do not behave like readings of one underlying quantity. The median absolute off-diagonal correlation is 0.22; 70% of pairs sit below $|\rho| = 0.4$ and 80% below 0.6. Under the convergent-validation logic of the multitrait-multimethod matrix [6], numbers of this size between purported measures of the same construct would provide weak evidence of convergence: two tests whose rankings of the same people agree at $\rho = 0.22$ are not two tests of one trait. The structure that does exist is instrumental rather than psychological: mousemove count and path length correlate at 0.80 because both count the same log rows, and mousemove count and hovered elements at 0.79 for the same reason; these are method factors, agreement induced by shared instrumentation, not convergent validity. Click count is close to orthogonal to everything ($|\rho| \leq 0.16$ against every other proxy), partly because the task makes it nearly constant, and the orientation proxies form their own block, strongly coupled to each other ($\rho = -0.79$ between minimum ad distance and ad-near time) but nearly uncoupled from dwell ($\rho = 0.23$ between dwell and ad-near time) and from activity volume. A composite "engagement score" built on these inputs does not aggregate one signal; it launders several unrelated ones into a single number whose meaning is set by arbitrary weights.

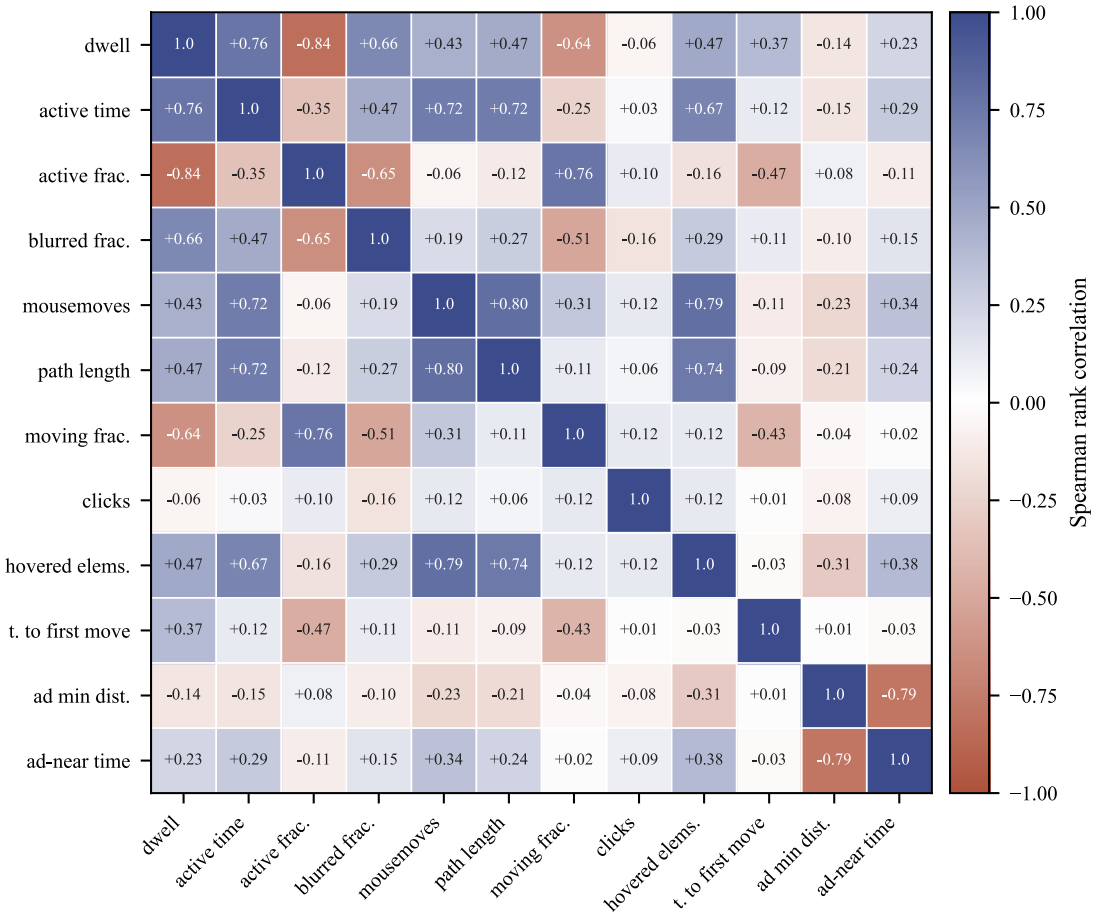

Figure 2: Spearman rank correlations among the twelve proxies of Table 1. Reanalysis of public data [4], N = 2,775 sessions (one per participant); coefficients are printed in each cell. Uncertainty is not shown per cell; at this N the standard error of a Spearman coefficient is about 0.02, so the displayed structure is not sampling noise. Evidence class: reanalysis of public data.

The pair that matters most for reporting is dwell against the active fraction of dwell: $\rho = -0.84$. Longer recorded sessions are systematically emptier sessions, in relative terms; dwell and continuity are not two views of one construct but near-opposites in rank. The mechanism is visible in the inflation analysis of Section 5.2, where the pair's quadrant geometry is also plotted: sessions become long mostly by accumulating idle and blurred time, not by accumulating activity, so the numerator of the active fraction saturates while the denominator grows. Under a median split, 40.6% of sessions combine long dwell with a below-median active fraction and another 40.6% combine short dwell with an above-median active fraction, for a discordance rate of 81.3%. The mean absolute percentile gap between a session's dwell rank and its active-fraction rank is 47 percentile points, close to the 50 expected under rank independence even though the two are computed from the same timestamps; the smallest rank gap among all audited pairs, mousemove count against path length, is still about 14 percentile points. Other named pairs disagree less dramatically but still substantially: dwell versus path length has 32.6% discordance ($\rho = 0.47$), dwell versus mousemove count 34.1% ($\rho = 0.43$), and dwell versus ad-near time 42.6% ($\rho = 0.23$). Even the tightest pair audited for discordance, mousemove count against hovered elements, leaves 19.5% of sessions in discordant quadrants (Table 2). The concrete consequence: a dashboard that segments these same 2,775 sessions into "engaged" halves by dwell, by activity, and by ad proximity produces three materially different populations, and any downstream statistic computed on "engaged users" depends on which proxy happened to define them.

Proxy pair	Spearman ρ	Discordant	High-A / low-B
dwell vs. active fraction	-0.84	81.3%	40.6%
dwell vs. ad-near time	$+0.23$	42.6%	21.3%
path length vs. ad-near time	$+0.24$	43.7%	21.9%
dwell vs. mousemoves	$+0.43$	34.1%	18.2%
dwell vs. path length	$+0.47$	32.6%	16.3%
active time vs. blurred fraction	$+0.47$	29.3%	14.6%
mousemoves vs. hovered elements	$+0.79$	19.5%	11.7%

Table 2: Median-split discordance for key proxy pairs, sorted by discordance. "Discordant" is the share of sessions above the median on exactly one member of the pair (50% under independence, 0% under perfect rank agreement); the last column is the share high on the first proxy and low on the second. Reanalysis of public data, N = 2,775.

5.2 Raw time overstates observed engagement

Figure 3 quantifies the gap between recorded presence and observed activity, using only recorded events. At the primary threshold $\theta = 2$ s, the median session's active/raw ratio is 0.57; a quarter of sessions fall below 0.32, and a tenth below 0.15. Equivalently, the median inflation factor of dwell over active time is 1.76, the 75th percentile 3.17, and the 90th percentile 6.52. Folding in focus: the median unattributable share (idle or blurred) is 51.4% of dwell across sessions, with a median blurred fraction of 0.20; 66.1% of sessions have at least 30% of their dwell unattributable to focused activity, and 51.0% have at least half. Blur alone is heavy: 47.7% of sessions spend at least a quarter of dwell out of focus and 32.1% at least half, and these are lower bounds on non-engagement since a focused-but-idle tab counts as attributable only up to the gap cap. The threshold matters in level but not in kind: at $\theta = 1$ s the median active fraction is 0.44 and 72.6% of sessions have at least 30% idle; at $\theta = 5$ s the figures are 0.80 and 42.8%. Every claim in this section is a direct measurement on recorded focus, blur, and timestamps, not a model.

Figure 4 shows the per-session geometry of the headline consequence: ranking sessions by dwell and by active fraction places 81% of them in discordant quadrants, so the inflation documented here is not a uniform background that cancels in comparisons; it reorders sessions.

The tail makes the point vividly. Twelve sessions exceed ten minutes of dwell and two exceed an hour; the maximum is 5,443,292 s, 63.0 days, a tab evidently left open, which by itself is 97.4% of the corpus's raw dwell. Summed naively, the corpus contains 1,552 hours of "person-time" but only 11.0 hours of active time (0.7%); winsorizing dwell at its 99th percentile still leaves the corpus only 30.6% active. There are 167 sessions with over two minutes of dwell and under 30 seconds of activity, and the top decile of blurred sessions has a median dwell of 121 s, five times the corpus median, confirming that long dwell and absent participants are largely the same phenomenon. None of this is pathological data; it is what presence logging records when nothing constrains it. A person-time total quoted without an activity complement is not conservative reporting with some noise on top. It is an upper bound produced by a measurement that cannot saturate, summed over sessions that can only add to it.

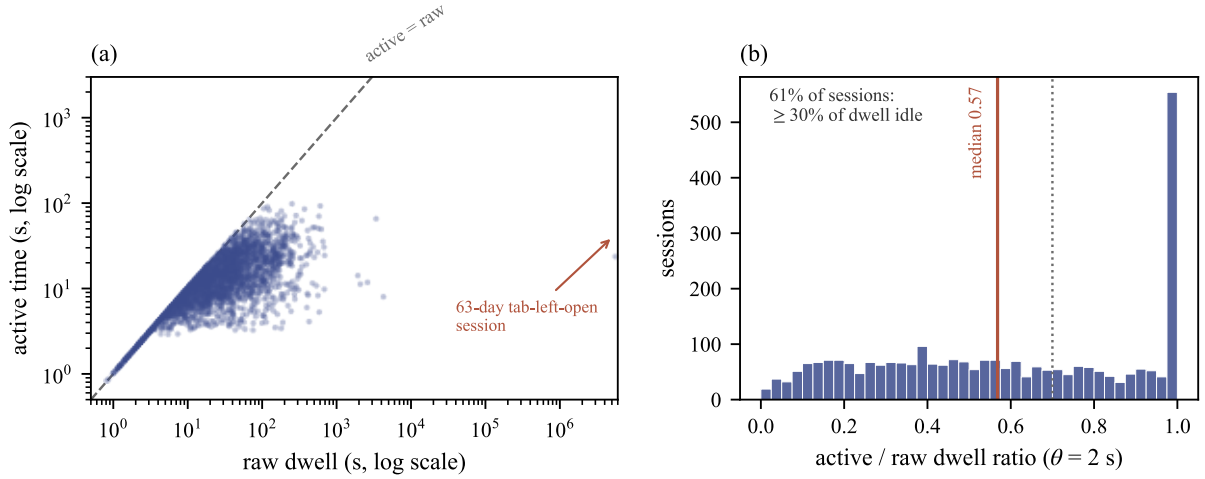

Figure 3: Raw dwell versus active time. (a) Per-session scatter on log axes; the dashed line is active = raw, and the labeled point is the 63-day left-open tab (5.44 million seconds of dwell, 24 s of active time). (b) Distribution of the active/raw ratio at $\theta = 2$ s; the solid line marks the median (0.57), and 61% of sessions have at least 30% of dwell idle (dotted line at ratio 0.7). Reanalysis of public data [4], N = 2,775. Evidence class: reanalysis of public data.

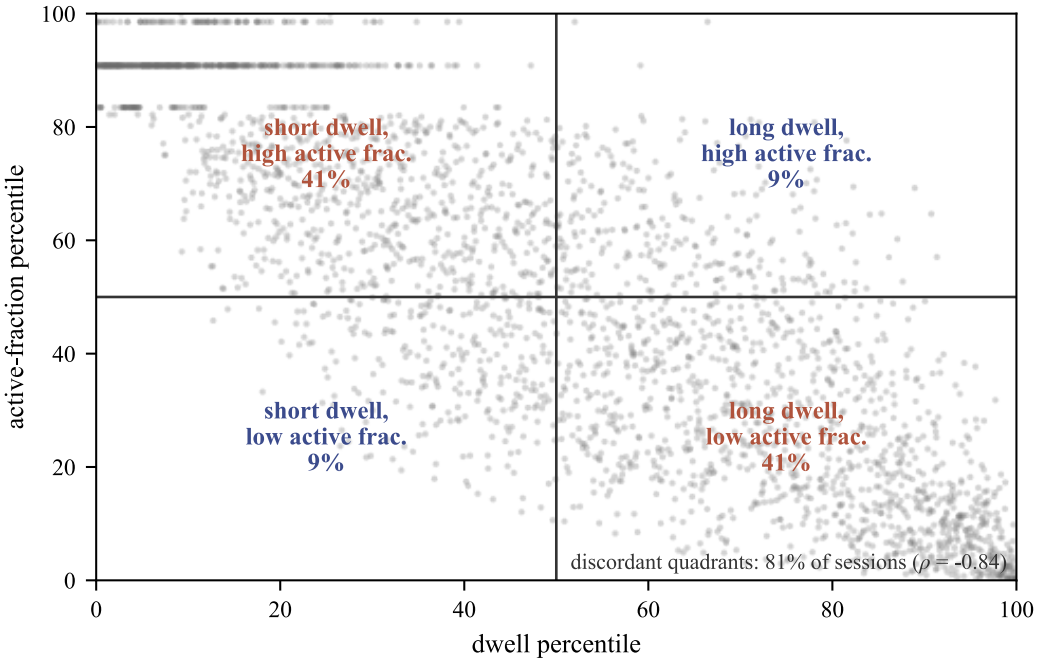

Figure 4: Discordance between dwell and active fraction under a median split. Each point is one session, placed by its percentile rank on each proxy; quadrant labels give the share of sessions in each quadrant. 81% of sessions are in the discordant quadrants (long dwell with low active fraction, or short dwell with high active fraction); horizontal banding reflects ties in active fraction (for example, fully active short sessions). Reanalysis of public data, N = 2,775.

5.3 Association with the criterion is weak everywhere

Table 3 is the criterion-validity table. Against self-reported attention, no proxy exceeds $|\rho| = 0.063$. The strongest association belongs to an orientation proxy, ad-near time ($\rho = +0.063$, 95% CI $[+0.025, +0.098]$), closely

followed by its mirror, minimum ad distance ($\rho = -0.062$ $[-0.098, -0.024]$); the weakest is hovered elements ($\rho = +0.004$ $[-0.032, +0.041]$). Dwell, the most reported engagement number on the web, associates with the person's own attention report at $\rho = +0.017$ $[-0.019, +0.054]$: indistinguishable from zero in 2,775 sessions collected under ideal single-page conditions. Seven of twelve intervals cross zero, and several point estimates run opposite to the folk reading: more total activity aligns weakly with less reported attention (active time -0.026, path length -0.037, clicks -0.030), while the blurred fraction, nominally an inattention signal, is weakly positive ($+0.058$), consistent with participants who left and returned still reporting that they noticed the ad.

Proxy	ρ (att.)	95% CI	AUC (click)	95% CI
ad-near time	$+0.063$	$[+0.025, +0.098]$	0.690	$[0.669, 0.710]$
ad min distance	-0.062	$[-0.098, -0.024]$	0.275	$[0.256, 0.294]$
blurred fraction	$+0.058$	$[+0.020, +0.091]$	0.466	$[0.442, 0.491]$
active fraction	-0.050	$[-0.085, -0.012]$	0.478	$[0.452, 0.506]$
path length	-0.037	$[-0.074, -0.001]$	0.412	$[0.387, 0.437]$
clicks	-0.030	$[-0.065, +0.007]$	0.500	$[0.483, 0.515]$
active time	-0.026	$[-0.063, +0.009]$	0.396	$[0.369, 0.421]$
moving fraction	-0.024	$[-0.063, +0.012]$	0.507	$[0.480, 0.532]$
dwell	$+0.017$	$[-0.019, +0.054]$	0.471	$[0.444, 0.498]$
time to first move	$+0.013$	$[-0.023, +0.051]$	0.525	$[0.500, 0.550]$
mousemoves	-0.010	$[-0.048, +0.027]$	0.447	$[0.422, 0.472]$
hovered elements	$+0.004$	$[-0.032, +0.041]$	0.456	$[0.432, 0.480]$

Table 3: Criterion associations, sorted by $|\rho|$ against self-reported attention (ordinal 1-5; Spearman with 95% participant-bootstrap CIs, 2,000 resamples) and AUC against ad click. AUC below 0.5 means the proxy runs opposite to the folk direction (for example, clickers accumulate less active time because clicking ends the session). Reanalysis of public data, N = 2,775.

The click column tells a different and instructive story. Orientation proxies predict the click well (ad-near time AUC 0.690; minimum ad distance 0.275, equivalently 0.725 when folded), for the mechanical reason that clicking the ad requires approaching it: the proxy and the outcome share geometry, not psychology. Activity volume runs backward: active time has AUC 0.396 and path length 0.412, meaning clickers produced less activity, because a click typically ends the session early. This sign reversal deserves emphasis, because "more activity" is the folk direction of every engagement composite, and here more activity is evidence against the one behavioral conversion the dataset records. The two criteria themselves agree only weakly: self-reported attention and the click correlate at $\rho = 0.157$ $[0.122, 0.190]$, and attention predicts the click at AUC 0.600 $[0.578, 0.621]$. Even the two most defensible "engagement outcomes" in the dataset, the person's own report and the person's own conversion, are far from interchangeable, so a metric validated against one has not been validated against the other.

The multivariate audit sharpens the contrast. The same twelve proxies, the same regularized model, the same participant-disjoint five-fold protocol, yield out-of-fold AUC 0.792 $[0.774, 0.811]$ for predicting the click but 0.558 $[0.535, 0.579]$ for predicting high (4-5) versus low (1-3) self-reported attention. Interaction logs carry real signal about what the hand did; they carry very little, at least through these standard proxies, about what the participant says their attention did. This is not a claim that attention is unpredictable from richer representations; trajectory models on this dataset report better figures [5]. It is a measurement of how little of that signal survives in the summary statistics that reporting actually uses.

5.4 Relationships move with context

Nothing about the participants changes across ad positions, ad types, or ad categories; the stimulus does. If a proxy measured a stable property of people, its relationships would not move materially across such strata. Figure 5 shows the strongest proxy-attention association, ad-near time, re-estimated within sixteen strata. Across ad categories the coefficient spans -0.095 (Toys, $[-0.268, +0.084]$) to $+0.148$ (Computers & Electronics, $[+0.078, +0.217]$), a range of 0.243 with a sign flip; the stratum standard deviation (0.075) is larger than the pooled coefficient itself (0.063). Individual stratum intervals are wide, and several cross zero, which is itself the finding in operational terms: within most single contexts, 100 to 800 sessions cannot even establish the direction of the best proxy-attention relationship, yet production decisions routinely ride on smaller slices.

The position strata expose a compositional subtlety worth separating from noise. Within top-left pages the association is $+0.107$ $[+0.061, +0.153]$ and within top-right pages $+0.092$ $[+0.027, +0.160]$, both clearly above the pooled $+0.063$. Pooling mixes two different distance geometries, sidebar ads live at systematically different cursor distances than top-of-page ads, so the pooled coefficient is attenuated below both of its parts. A proxy can thus look weaker, or stronger, than it is in any actual context, purely as a function of the traffic mix, and any longitudinal engagement series computed over a shifting mix inherits drift that has nothing to do with users. The same instability afflicts proxy-proxy structure: dwell versus path length is $\rho = 0.468$ pooled but ranges from 0.340 to 0.556 across categories (range 0.216) and from 0.438 to 0.514 across device buckets. A proxy whose meaning shifts with the stimulus is a context measurement wearing the vocabulary of a person measurement.

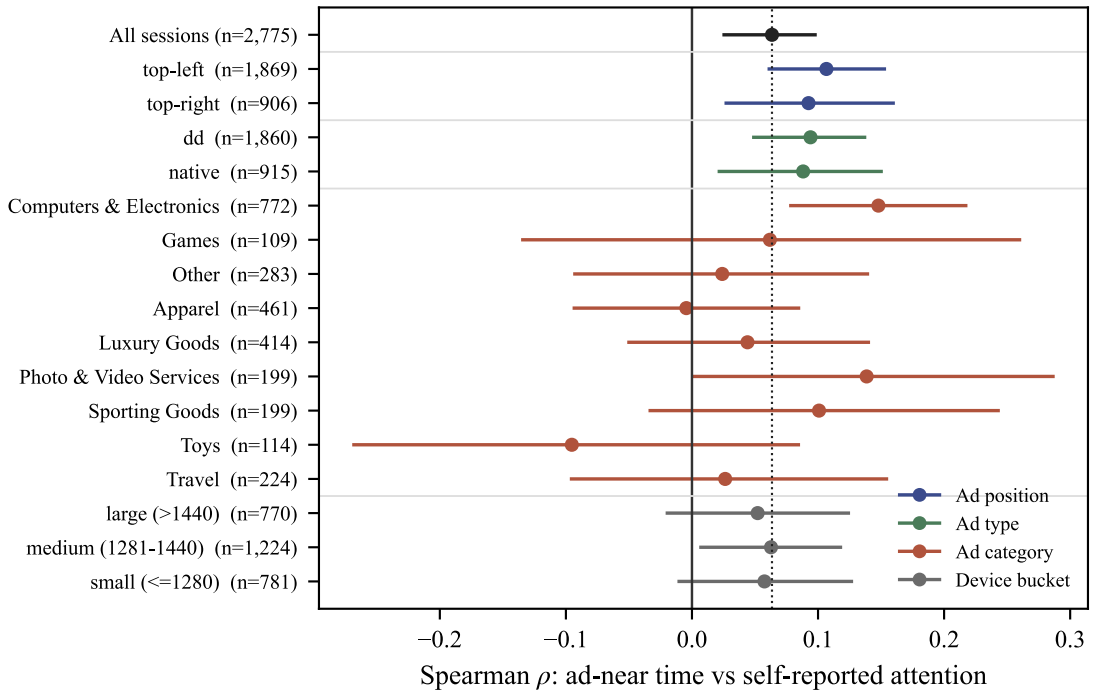

Figure 5: The strongest proxy-attention association (ad-near time vs. self-reported attention, Spearman ρ) re-estimated within context strata; horizontal bars are 95% participant-bootstrap CIs and the dotted line is the pooled estimate. Stratum sizes are printed in the labels (total N = 2,775). Across ad categories the estimate spans -0.095 to $+0.148$, including a sign flip. Reanalysis of public data [4].

5.5 Transfer is unreliable group by group

Figure 6 reports the leave-one-context-out audit. Averaged over held-out groups, transfer looks benign: for the click model, leave-one-ad-category-out AUC averages 0.787 against a size-matched random-split baseline of 0.793, a mean drop of $+0.006$ with a bootstrap interval over groups of $[-0.015, +0.030]$; leave-one-device-bucket-out shows no drop at all (-0.001 $[-0.015, +0.016]$, out-of-context range 0.776 to 0.810). We report these near-null averages exactly as they are: on this dataset, the pooled proxy-to-click mapping does not collapse on average when a category or a device class is held out, and an audit that expected otherwise must say so.

The instability lives one level down, in the group-by-group results that an average erases. For the click model, out-of-context AUC ranges from 0.721 ("Other," a drop of 0.075) to 0.829 across held-out categories, a spread of 0.108 that the baseline's sampling variation (SD 0.006 across groups) cannot begin to explain; some categories transfer better than in-context validation predicts, others substantially worse, and nothing in the pooled number says which. For the attention model the same pattern is starker because there is less signal to begin with: mean drop $+0.007$ $[-0.022, +0.035]$, but out-of-context AUC spans 0.474 to 0.627 across categories (spread 0.153), falling to or below chance in two of nine held-out categories (Toys 0.474, Travel 0.496) while the size-matched random baseline never approaches chance for any group. The device-bucket version is milder (0.538 to 0.588) but the model was near chance everywhere to begin with. The operational reading: a practitioner who validated the attention model in-context at 0.567 and shipped it to a new category could receive anything from a modest improvement to worse-than-chance behavior, and the mean transfer drop, the number a summary table would show, conceals exactly this. For weak-signal models, "transfers on average" and "transfers" are different claims, and only the second one matters to whoever inherits the model in one specific context.

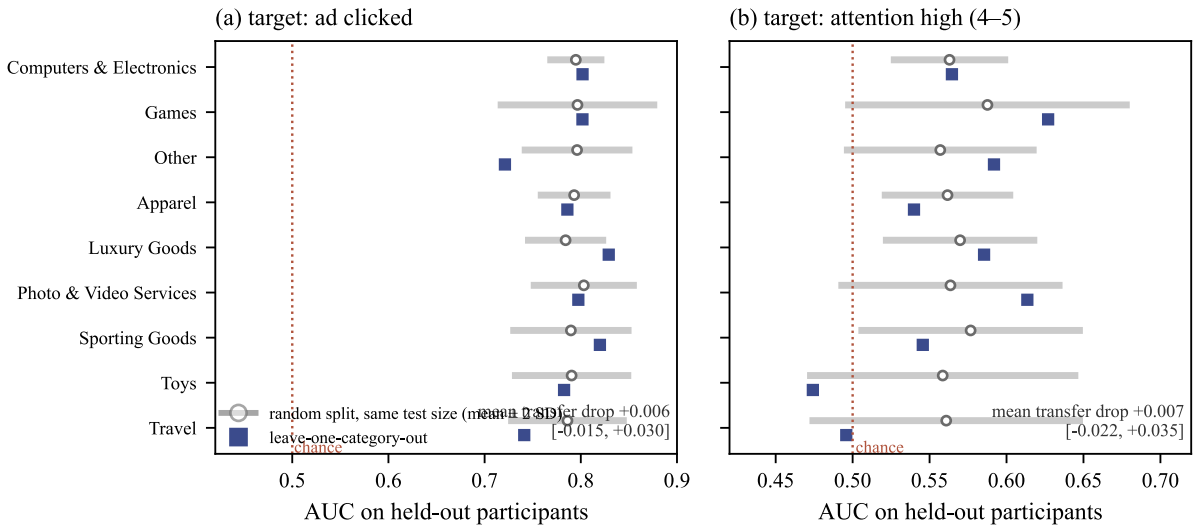

Figure 6: In-context versus out-of-context performance, leave-one-ad-category-out. For each held-out category: gray open circles show the mean (with ± 2 SD bars) of 20 random participant-disjoint splits with a test set of the same size; indigo squares show the AUC when that category is held out entirely. (a) Click model; (b) attention-high model (cut at 4). The dotted line is chance. Mean drops with bootstrap 95% CIs over groups are printed per panel. Reanalysis of public data, N = 2,775.

5.6 Counterfactual stress tests: the metrics move without the person

The final audit asks how far each headline metric can be pushed by transformations that are deterministic, bounded, and add no evidence about any person. These are metric-arithmetic demonstrations on frozen logs, labeled COUNTERFACTUAL in Figure 7; they are not observations of behavior, and they change nothing about any participant. Each transformation is chosen to be boring: no fabricated sessions, no synthetic users, just small

perturbations of event streams of the kind that a motivated party, or a careless instrumentation change, could produce.

Idle injection appends Δ seconds of event-free time to a log. At $\Delta = 300$ s this multiplies the median session's dwell by 13.49, and by 65.2 at the 90th percentile, since short sessions inflate more; at a mere $\Delta = 30$ s the median inflation is already 2.25-fold. Active time, by contrast, moves by a median factor of 1.18 regardless of Δ, because its increase is capped at the 2 s gap threshold: the capped-gap construction is, by design, bounded against exactly this manipulation, and the validation run on actually transformed logs confirms the arithmetic to floating-point precision. Densification inserts k points per consecutive mousemove step, deflected alternately ± 2 px perpendicular to the segment. At $k = 63$ this multiplies the median session's mousemove count by 58.8 and its path length by 3.14 (7.88 at the 90th percentile), while the trajectory never leaves a 2 px envelope of the original, invisible at screen resolution, with endpoints, bounding box, and dwell untouched. The pure-duplication variant (k copies, no deflection) inflates the count identically while changing path length by a factor of exactly 1.00: the two headline "activity" numbers can be decoupled at will, which no reading of either as "amount of behavior" survives. Click splitting multiplies click count exactly k-fold by placing the extra clicks inside the 150 ms window after each real click; in this corpus no session ends on a click, so dwell changes by exactly zero seconds, and active time is unchanged because the inserted events subdivide gaps already below the threshold.

In each case, every signal above the manipulated rung of the ladder, focused time, hover diversity, ad proximity, and of course the self-report, is unchanged, which is the point: a metric that a bounded, information-free transformation can multiply 59-fold is not measuring a person. This is Goodhart's law in its most literal form [24], and it applies symmetrically to fraud (a bad actor inflating billable engagement), to incentive drift (teams optimizing a target metric with UI changes that add idle presence rather than value), and to innocent instrumentation change (an SDK that raises its mousemove sampling rate silently multiplies "activity" corpus-wide, exactly as the duplication variant does). The defense is equally mechanical: report the invariants alongside the headline. Active time bounds idle injection; path-per-event and spatial envelopes expose densification; deduplication windows expose click splitting. Each check costs a few lines of arithmetic on data already collected.

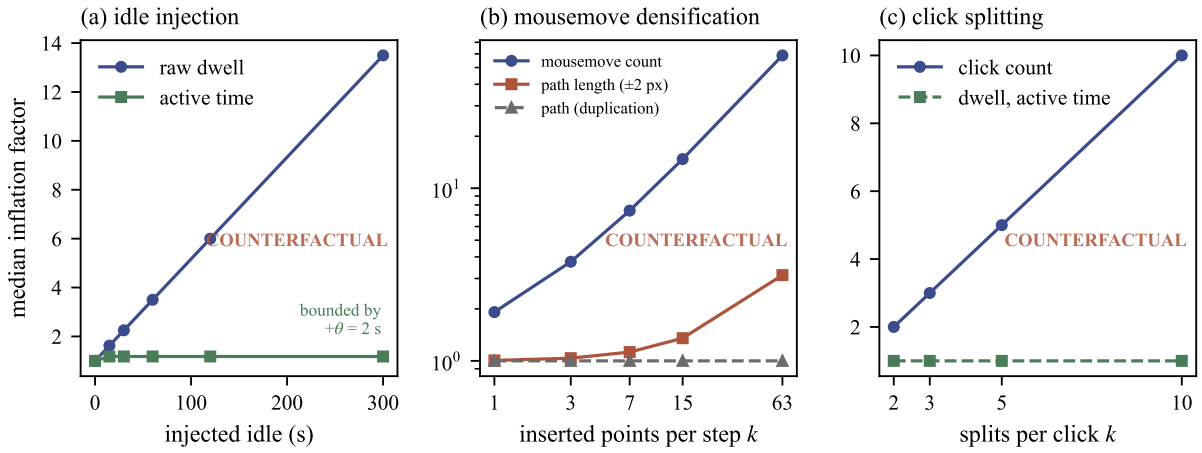

Figure 7: COUNTERFACTUAL stress tests on frozen logs (N = 2,775; densification panel N = 2,741 sessions with at least two mousemoves and positive path; click panel N = 2,544 sessions with a click). (a) Appending idle time inflates median dwell linearly while active time is capped at +2 s. (b) Inserting k points per mousemove step within a ± 2 px envelope inflates the count (k = 63: 58.8-fold) and path length (3.14-fold); pure duplication moves path length by exactly 1.00. (c) Splitting each click into k clicks inside 150 ms inflates click count exactly k-fold with dwell and active time unchanged. Evidence class: counterfactual stress test on frozen logs; these curves describe metric arithmetic, not human behavior.

6 Discussion

The five results are one result seen from five sides. A set of proxies that measured one construct would correlate with each other, would track an independent measurement of the construct, would keep its internal structure when the stimulus changes, and could not be moved by information-free transformations. The audited engagement proxies fail all four tests on data collected under conditions more controlled than many production settings: one page, one task, one ad, one device class, and a population that agreed to be measured.

The practical reading is about claims, not about logs. Interaction logs are excellent measurements of what they actually record. Dwell is a real measurement of recorded presence; it supports capacity planning, abandonment analysis, and session reconstruction. Active and focused time are real measurements of continuity; the focus/blur accounting in Section 5.2 is precisely the in-focus precondition that the viewability standard already demands [3], and reporting it would cost nothing. Cursor proximity is a real measurement of orientation, and it is the only rung in this audit with a consistently nonzero, if small, criterion association, consistent with the gaze-cursor literature it descends from [12], [14]. What the data do not license is the substitution of any of these for attention. The proxies' mutual disagreement means "engagement" composites inherit their weights arbitrarily: a dashboard that averages dwell with active fraction is averaging two numbers that rank the same sessions in nearly opposite order ($\rho = -0.84$). The weak criterion association means that even the best single proxy explains a trivial share of rank variance in the person's own report. The context instability means a threshold tuned in one category silently changes meaning in the next. And the counterfactuals mean that any of these numbers, used as a target or a billing basis, can be manufactured.

For measurement practice, the ladder suggests a discipline that is easy to adopt. Name the rung: report "recorded presence," "active time," or "cursor-near-target time," not "attention," unless an attention measurement was actually taken. Report the complement: any dwell number should travel with its active and focused shares, which this dataset shows would reclassify roughly half of recorded time. Report stability: a proxy offered as meaningful should come with its cross-context dispersion, as in Figure 5, not only a pooled coefficient. And report gameability: the counterfactual curves of Section 5.6 are cheap to compute for any metric and reveal which numbers an adversary, or an SDK update, can move. None of this requires new instrumentation; all of it is contained in logs that systems already keep.

The findings also bear on how engagement research validates itself. A common pattern in the literature, and in this dataset's own companion papers, is to demonstrate that a rich representation of behavior predicts an attention label above chance, and to conclude that behavior carries attention signal. That conclusion is correct, and our multivariate click model shows the logs carry plenty of signal about behavior. But the practice it licenses in the wild is different: dashboards do not ship trajectory models, they ship the twelve numbers in Table 1, and the validity of the model does not transfer to the summary statistics. The audit gap, between what the best model can extract and what the reported statistic retains, is where the proxy trap lives. Related, our results give a concrete reason to distrust cross-study comparisons of "engagement" effect sizes: two studies that both report "dwell" may be reporting quantities that correlate at -0.84 with each other's active-fraction variant, computed under different idle conventions, on different context mixes. The measurement, not just the population, differs.

The audit approach itself is portable. Nothing in A2 through A7 is specific to SERPs, ads, or cursors beyond the proxy definitions: any surface that reports engagement, feed scrolling, video players, reading apps, in-game telemetry, has its own ladder of presence, activity, continuity, and orientation, its own focus events, and its own gameable headline numbers. The battery, disagreement matrix, inflation accounting, criterion table if any criterion exists, context stratification, group-wise transfer, and counterfactual curves, applies unchanged, and the scripts released with this report implement it end to end on public data.

7 Attention-Claims Checklist

Table 4 condenses the audit into a reporting checklist. It is deliberately phrased as questions a reader can ask of any engagement claim; the third column names the failure the question guards against, as measured in this report. The checklist is intended to be used the way limitations sections are supposed to be used, before publication rather than after: an engagement number that passes C1 through C3 is at least honestly labeled and bounded; one that passes C4 through C7 has demonstrated, rather than assumed, that it measures something stable about people; C8 and C9 are the adversarial and rhetorical guards. Most items cost one line of arithmetic or one sentence of prose. None requires new data collection, which is deliberate: the audit above was performed entirely on logs of the kind every instrumented product already has.

#	Question to ask of the claim	Failure it guards against (measured here)
C1	Which ladder rung is actually measured: exposure, presence, activity, continuity, orientation, or attention?	Mental-state vocabulary applied to timestamp arithmetic (Figure 1).
C2	Is raw time reported with its active and focused shares, under a stated idle threshold and threshold sensitivity?	Median session: 51% of dwell idle or blurred; 66% of sessions $\geq$ 30% unattributable (Section 5.2).
C3	Are totals robust to left-open tabs (caps, winsorization, per-session distributions rather than sums)?	One session = 97.4% of raw corpus dwell; corpus active share 0.7% raw vs. 30.6% winsorized.
C4	If several proxies are reported or composited, is their mutual agreement shown?	Median pairwise $\lvert\rho\rvert$ = 0.22; dwell vs. active fraction $\rho = -0.84$, 81% discordant (Section 5.1).
C5	Is any attention/interest claim backed by an independent criterion, with effect size and CI, not by the proxy itself?	Best single proxy: ρ = 0.063; multivariate AUC 0.558 vs. self-report (Section 5.3).
C6	Are associations reported per context (position, category, device), with dispersion, not only pooled?	Sign flips across categories; stratum SD larger than the pooled effect (Section 5.4).
C7	Was validation participant-disjoint, and was out-of-context transfer tested group by group?	Attention model at or below chance in 2 of 9 held-out categories despite a near-zero mean drop (Section 5.5).
C8	How far can bounded, information-free transformations move the metric (idle injection, event duplication, click splitting)?	Dwell $\times$13.5, mousemoves $\times$58.8, path $\times$3.1, clicks $\times\ k$, with no attention-relevant signal changed (Section 5.6).
C9	Does the claim avoid comprehension, memory, interest, or intent language unless those were separately measured?	No log signal here measures them; see "What this paper does not claim."

Table 4: The Attention-Claims Checklist. Each row is a question to ask of any engagement or attention claim derived from interaction logs, with the corresponding failure mode quantified in this report.

8 Limitations

The criterion is a single retrospective self-report item per participant, ordinal and coarse; it is subject to demand effects, recall error, and scale-use differences, and it measures noticing an ad, not attention in any richer sense. A weak proxy-criterion association therefore bounds the evidence the proxies provide about this criterion; it cannot rule out that both proxy and self-report are noisy views of something they would agree on under better measurement, and attenuation from criterion noise means the true proxy-attention associations could be somewhat larger than measured, though attenuation cannot manufacture the observed sign flips across contexts. The design has one observation per participant, so within-person questions, whether a given person's proxies track that person's attention over time, are unanswerable here, and all context contrasts are between-participant; a longitudinal replication would be the natural next study. The setting is a desktop SERP with a computer mouse and a single ad; touch, mobile, feeds, video, and multi-page journeys are out of scope, and the cursor-gaze coupling that motivates orientation proxies is itself known to vary by user and task [14], so cursor position is a proxy for gaze, not a measurement of it. The crowdsourced population and the paid, instructed task differ from organic

browsing in unknown ways, most plausibly by inflating on-task activity relative to the wild, which would make our inflation estimates conservative. The idle threshold and the 250 px ad-near band are analyst choices; we report sensitivity for the former and chose the latter from the distance distribution, but other choices would shift levels (not, in our checks, the qualitative pattern). The 134 excluded ultra-short logs are a stated filter, and their exclusion can only have removed sessions whose proxies were least informative. The transfer analysis is limited by the granularity of the released context fields; "ad category" is a coarse proxy for context shift, and nine groups support only wide intervals over the mean drop. Finally, the counterfactual results are statements about metric arithmetic under specified transformations, and about nothing else; in particular they do not estimate how often such inflation occurs in the wild, only how cheap it would be.

9 Conclusion

On 2,775 public web-search sessions with self-reported attention labels, the standard engagement proxies disagree with each other (median pairwise $|\rho| = 0.22$; dwell versus active fraction $\rho = -0.84$), overstate observed engagement by construction (median 51% of dwell idle or blurred; a 63-day left-open tab is 97% of raw corpus dwell), associate only weakly with the participant's own attention report (best single proxy $\rho = 0.063$; multivariate AUC 0.558, against 0.792 for the mechanical click), shift and flip sign across contexts, transfer erratically, and inflate under deterministic transformations that add no evidence about anyone. Activity is not a self-validating measure of attention. The remedy is not to abandon interaction logs but to say what they measure: name the ladder rung, publish the active and focused complements, show the disagreement, validate against an independent criterion in and out of context, and stress-test the metric's arithmetic before anyone else does. The checklist in Section 7 is offered as a compact way to hold engagement reporting, our own included, to that standard.

Reproducibility: all numbers in this report are generated into `results/results.json` by the scripts in `analysis/` (`run_all.sh` executes the event inventory, analyses A1-A7, and figure generation deterministically, seed 42) from the public dataset [4]. Figures F1-F7 are produced as SVG and 300-dpi PNG by `analysis/make_figures.py`. GrayPass Labs is the author affiliation; this report analyzes public data only.

References

[1] H. A. Simon, "Designing Organizations for an Information-Rich World," *Computers, Communications, and the Public Interest*. Johns Hopkins Press, Baltimore, MD, pp. 37–72, 1971.

[2] L. J. Cronbach and P. E. Meehl, "Construct Validity in Psychological Tests," *Psychological Bulletin*, vol. 52, no. 4, pp. 281–302, 1955, doi: 10.1037/h0040957.

[3] Media Rating Council, "Viewable Ad Impression Measurement Guidelines, Version 2.0," technical report, 2015. [Online]. Available: https://www.mediaratingcouncil.org/sites/default/files/Standards/081815%20Viewable%20Ad%20Impression%20Guideline_v2.0_Final.pdf

[4] L. A. Leiva and I. Arapakis, "The Attentive Cursor Dataset," *Frontiers in Human Neuroscience*, vol. 14, p. 565664, 2020, doi: 10.3389/fnhum.2020.565664.

[5] I. Arapakis and L. A. Leiva, "Learning Efficient Representations of Mouse Movements to Predict User Attention," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, ACM, 2020, pp. 1309–1318. doi: 10.1145/3397271.3401031.

[6] D. T. Campbell and D. W. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, vol. 56, no. 2, pp. 81–105, 1959, doi: 10.1037/h0046016.

[7] M. Claypool, P. Le, M. Waseda, and D. Brown, "Implicit Interest Indicators," in *Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI '01)*, ACM, 2001, pp. 33–40. doi: 10.1145/359784.359836.

[8] D. Kelly and N. J. Belkin, “Display Time as Implicit Feedback: Understanding Task Effects,” in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*, ACM, 2004, pp. 377–384. doi: 10.1145/1008992.1009057.

[9] C. Liu, R. W. White, and S. Dumais, “Understanding Web Browsing Behaviors through Weibull Analysis of Dwell Time,” in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*, ACM, 2010, pp. 379–386. doi: 10.1145/1835449.1835513.

[10] X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan, “Beyond Clicks: Dwell Time for Personalization,” in *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*, ACM, 2014, pp. 113–120. doi: 10.1145/2645710.2645724.

[11] M.-C. Chen, J. R. Anderson, and M.-H. Sohn, “What Can a Mouse Cursor Tell Us More? Correlation of Eye/Mouse Movements on Web Browsing,” in *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2001, pp. 281–282. doi: 10.1145/634067.634234.

[12] K. Rodden, X. Fu, A. Aula, and I. Spiro, “Eye-Mouse Coordination Patterns on Web Search Results Pages,” in *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2008, pp. 2997–3002. doi: 10.1145/1358628.1358797.

[13] Q. Guo and E. Agichtein, “Towards Predicting Web Searcher Gaze Position from Mouse Movements,” in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2010, pp. 3601–3606. doi: 10.1145/1753846.1754025.

[14] J. Huang, R. W. White, and G. Buscher, “User See, User Point: Gaze and Cursor Alignment in Web Search,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, ACM, 2012, pp. 1341–1350. doi: 10.1145/2207676.2208591.

[15] J. Huang, R. W. White, and S. T. Dumais, “No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, ACM, 2011, pp. 1225–1234. doi: 10.1145/1978942.1979125.

[16] I. Arapakis and L. A. Leiva, “Predicting User Engagement with Direct Displays Using Mouse Cursor Information,” in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*, ACM, 2016, pp. 599–608. doi: 10.1145/2911451.2911505.

[17] L. Brückner, I. Arapakis, and L. A. Leiva, “When Choice Happens: A Systematic Examination of Mouse Movement Length for Decision Making in Web Search,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, ACM, 2021, pp. 2318–2322. doi: 10.1145/3404835.3463055.

[18] I. Arapakis, A. Penta, H. Joho, and L. A. Leiva, “A Price-per-Attention Auction Scheme Using Mouse Cursor Information,” *ACM Transactions on Information Systems*, vol. 38, no. 2, pp. 13:1–13:30, 2020, doi: 10.1145/3374210.

[19] L. A. Leiva, I. Arapakis, and C. Iordanou, “My Mouse, My Rules: Privacy Issues of Behavioral User Profiling via Mouse Tracking,” in *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21)*, ACM, 2021, pp. 51–61. doi: 10.1145/3406522.3446011.

[20] H. L. O'Brien and E. G. Toms, “What Is User Engagement? A Conceptual Framework for Defining User Engagement with Technology,” *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008, doi: 10.1002/asi.20801.

[21] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret, “Models of User Engagement,” in *User Modeling, Adaptation, and Personalization (UMAP 2012), Lecture Notes in Computer Science, vol. 7379*, Springer, 2012, pp. 164–175. doi: 10.1007/978-3-642-31454-4_14.

[22] M. Lalmas, H. O'Brien, and E. Yom-Tov, *Measuring User Engagement*. in Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool, 2014. doi: 10.2200/S00605ED1V01Y201410ICR038.

[23] C. Spearman, “The Proof and Measurement of Association between Two Things,” *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904, doi: 10.2307/1412159.

[24] M. Strathern, “`Improving Ratings': Audit in the British University System,” *European Review*, vol. 5, no. 3, pp. 305–321, 1997, doi: 10.1017/S1062798700002660.