Lag You Can Feel: Within-Person Links Between Latency, Performance, and Trust in Mixed Reality

Peng Wu[1], Mingyu Zhu[2], Zifan Zhou[2], Jianan Jiang[2], Haowei Li[2], Juaren Steiger[2], Nasim Ahmed[3], Kaiming Huang[2], Gang Tan[2], Tian Lan[4], Rifatul Islam[3], Jonathan Dodge[2], Mahdi Imani[1]
[1]Northeastern University, Boston, MA, USA [2]Pennsylvania State University, University Park, PA, USA
[3]Kennesaw State University, Kennesaw, GA, USA [4]George Washington University, Washington, DC, USA
wu.p@northeastern.edu, {mintrrey, zqz5454, jnjiang, hql5799, juaren.steiger, kzh529, gtan, dodge}@psu.edu, {nahmed25@students., rislam11@}kennesaw.edu, tlan@email.gwu.edu, m.imani@northeastern.edu

Abstract

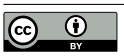
Timely system response is a major determinant of interactive mixed reality (MR) quality, and degraded responsiveness can arise from many sources: network congestion, server or edge/cloud load, contention, or deliberate interference. We study how a degraded responsiveness operating condition and a compensated configuration jointly affect *objective* task performance and *subjective* ratings (engagement and trust), using a within-subject MR shooting study (22 participants × 3 configurations = 66 sessions). The degraded configuration significantly worsens all reported outcomes, and the compensated configuration partially restores them; our focus is the performance–rating association. Using a within–between (Mundlak) decomposition and repeated-measures correlation, we find participant-relative covariation across configurations: sessions with better-than-average task outcomes for the same participant also tend to receive higher ratings (within-person $|r_{\mathrm{rm}}|$=0.32–0.63; all eight associations remain significant after false-discovery-rate correction for multiple comparisons). The between-participant estimates are imprecise, with every 95% confidence interval including zero, so the cross-person relation remains uncertain in this sample. Practically, knowing a participant's own ratings from another session improves prediction of the subjective ratings (R^2 rising from $\approx$ 0 to 0.28–0.37), while full-session behavioral telemetry (gaze, head motion, and firing behavior) predicts the *objective* session outcomes (R^2=0.33–0.55, above a configuration-only baseline). These results support person-relative QoE assessment in MR and motivate measuring task outcomes alongside engagement-related and system-trust ratings.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**; • **Security and privacy** → *Human and societal aspects of security and privacy*.

Keywords

mixed reality, latency, quality of experience, trust, engagement, within-person analysis, behavioral telemetry



XR Security '26, October 26–30, 2026, Austin, TX, USA

ACM ISBN 979-8-4007-2951-5/26/10
https://doi.org/10.1145/3842203.3844589

ACM Reference Format:
Peng Wu[1], Mingyu Zhu[2], Zifan Zhou[2], Jianan Jiang[2], Haowei Li[2], Juaren Steiger[2], Nasim Ahmed[3], Kaiming Huang[2], Gang Tan[2], Tian Lan[4], Rifatul Islam[3], Jonathan Dodge[2], Mahdi Imani[1]. 2026. Lag You Can Feel: Within-Person Links Between Latency, Performance, and Trust in Mixed Reality. In *Second Workshop on Enhancing Security, Privacy, and Trust in Extended Reality (XR Security '26), October 26–30, 2026, Austin, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3842203.3844589

1 Introduction

Mixed and extended reality (XR) systems depend on a tightly timed perception–action loop. A user's movement, gaze, and input must be captured, processed, rendered, and reflected back through the display fast enough for the system to feel responsive. Delay in this loop can reduce task performance and user comfort [16]. In networked, mobile, and edge/cloud-rendered XR, this timing budget is also shaped by computation, transmission, congestion, server load, and resource contention [8, 10]. From the user's perspective, these sources — benign or adversarial — have a common consequence: the system responds slowly or inconsistently, and the user's trust in it is what is ultimately at stake.

A central design question is whether this degradation is fully captured by objective task metrics. If degraded responsiveness only lowers visible task performance, then monitoring performance may be enough. If it also changes how users judge the system, then task metrics alone provide an incomplete view of XR quality. This is especially important for system-trust ratings. Trust in automation is shaped by perceived reliability, predictability, and responsiveness [5, 9]. In an MR shooting task, a user may continue to achieve a reasonable score while still rating the system as less responsive, less dependable, or harder to understand.

This paper studies the relationship between human–system task outcomes and post-session ratings under three MR system configurations: Baseline, Degraded, and Compensated, treating the manipulation as a system operating condition rather than a millisecond-calibrated latency threshold. Our goal is to understand how task outcomes and ratings change across these configurations, and whether their relationship is best understood within a participant or across participants.

The level of analysis matters. A pooled correlation between performance and ratings can mix three sources with different meanings: the configuration moving both variables, stable differences between participants (users who perform better on average also trust more), and session-level variation within the same participant (a user rates the system more favorably when their own performance is better

than usual). We therefore analyze the data with repeated-measures methods and a hybrid within–between model.

We report a within-subject MR shooting study with 22 participants and three configurations, for 66 sessions in total. The study measures four human–system task outcomes, engagement-related ratings, direct system-trust ratings, and supporting behavioral telemetry. Our contributions are:

- A controlled within-subject characterization of four human–system task outcomes and construct-specific ratings across three MR system configurations.
- A hybrid repeated-measures analysis that separates within-participant and between-participant performance–rating associations, with configuration adjustment used to distinguish shared condition effects from participant-relative covariation.
- A participant-held-out test of whether a person-specific baseline improves prediction of subjective ratings, and an exploratory analysis of whether full-session behavioral telemetry estimates objective task outcomes for held-out participants.

In short, Degraded worsens both task outcomes and ratings and Compensated partially recovers them; sessions in which a participant beats their own average receive higher ratings, whereas between-participant estimates remain imprecise; and a person-specific baseline improves rating prediction. These findings support person-relative QoE assessment in MR and motivate measuring task outcomes alongside engagement and system-trust ratings.

2 Background and Related Work

Latency, performance, and experience. Prior work has established that timing degradation can affect interaction quality, task performance, and discomfort in immersive systems [16]. Systems work has also studied how mobile, networked, and untethered XR systems manage tight rendering and latency budgets [8, 10]. This paper addresses a level-of-analysis question: when task outcomes and ratings change together, does the association reflect the shared system configuration, participant-relative variation, stable differences among participants, or a combination of these sources?

Predicting XR user state. A growing line of work infers a user's internal state in XR from behavioral and physiological telemetry, most prominently cybersickness, typically quantified with the Simulator Sickness Questionnaire [7] and reviewed broadly across displays and applications [1, 15]. Recent Bayesian-network models show that personal attributes (age, gender, prior VR experience) improve cybersickness prediction for previously unseen users [19], and that the same machinery can *verify* and quantify cybersickness rather than merely classify it [20]. Both target cybersickness, a discomfort outcome. We study engagement-related and system-trust ratings under degraded MR responsiveness and evaluate whether same-participant baseline information improves rating prediction.

Engagement and system-trust ratings. Trust in automation is shaped by perceived reliability, predictability, and responsiveness [5, 9]. Engagement and presence also have established measurement traditions in HCI and XR [14, 18]. Because our post-session questionnaire uses compact task-specific items rather than full validated instruments, we treat the ratings as study-specific and separate task engagement, perceived responsiveness, visual control, direct system trust, expected fallibility, and understandability.

Security and privacy in XR. The XR security and privacy literature has mapped authentication for head-mounted displays [6, 17] and shown that motion telemetry can support user identification [11, 13]. This creates a tension for person-specific monitoring: the same behavioral streams that may help estimate quality or performance also require privacy-aware handling. In this paper, degraded responsiveness is treated as a source-agnostic operating condition; attack attribution and detector evaluation are outside our scope.

Within- vs. between-person analysis. Repeated-measures designs can mix stable differences between participants, shared condition effects, and session-level variation within the same participant. A pooled correlation cannot separate these sources. The within–between, or Mundlak, device separates participant-centered observations from participant means by entering the person mean as its own predictor [3, 4, 12]. Repeated-measures correlation reports a common within-participant association as an unadjusted descriptive summary [2]. Because our experimental configuration shifts both performance and ratings, our primary interpretation must also account for configuration.

3 Study and Method

Task and conditions. Each of 22 participants played a ∼6 min MR shooting task (mean duration ≈ 361 s) on a Meta Quest Pro head-mounted display with integrated eye tracking: enemies spawn in the user's physical surroundings and advance toward the player, who finds them by looking around and shoots them with a hand-held controller; an enemy that reaches the player counts as a breakthrough. The three randomized system configurations — Baseline, Degraded, and Compensated — used internal degradation settings of 0/100/40 on the tracking-and-rendering update path (the 40 is the setting *before* compensation). These are implementation-specific values rather than calibrated millisecond measurements, so we describe the manipulation as degraded system responsiveness. We did not instrument motion-to-photon delay; the configuration's measurable footprint in our logs is the effective eye-tracking stream rate (∼71/9/19 Hz), which drops because the throttled update path also delays tracking-sample delivery — a manipulation check, not a latency measurement. Compensated combined the lower setting with reprojection/timewarp, late input sampling, predictive state compensation, and timestamped hit registration; the contrast therefore estimates the complete compensated configuration, not any individual component.

Measures. Four *objective* metrics are parsed from the end-of-session log. Let h be the number of hits, s the number of shots, D the set of defeated enemies, and $t_e^{\text{find}}, t_e^{\text{def}}$ the find and defeat timestamps of enemy e. Then

$$\text{accuracy} = \frac{h}{s}, \qquad \text{avg\_ttd} = \frac{1}{|D|} \sum_{e \in D} \left(t_e^{\text{def}} - t_e^{\text{find}} \right), \tag{1}$$

with defeated = $|D|$ (enemies destroyed) and reached_user the count of enemies that breached the player. Two *subjective* outcomes aggregate four post-session 1–7 Likert items each (Table 1); trust

Table 1: The eight post-session survey items (1–7 Likert). Engagement items are positively worded; trust item T3 is reverse-worded and recoded as $8 - \text{T3}$ in the aggregate.

Engagement (1–7 Likert; higher = more)	
E1	How much were you engaged in the mission?
E2	How responsive was the environment to actions that you initiated?
E3	How completely were you able to actively survey or search the environment using vision?
E4	How focused and mentally involved were you with the mission?
Trust (1–7 Likert)	
T1	I trust the MR system.
T2	I can rely on the MR system.
T3	The system might make occasional errors. *(reverse-worded)*
T4	I was able to understand why things happened.

reverse-codes its negatively worded item T3:

$$\text{engagement} = \tfrac{1}{4}\sum_{k=1}^{4} \text{E}k, \qquad \text{trust} = \tfrac{1}{4}\big(\text{T1}+\text{T2}+(8-\text{T3})+\text{T4}\big). \quad (2)$$

The items follow the style of established presence/engagement questionnaires [14, 18] (E1–E4) and trust-in-automation scales [5, 9] (T1–T4), in a compact one-item-per-facet form (a limitation, §7). Internal consistency over the 66 sessions is acceptable (Cronbach's α=0.85 for engagement and 0.80 for trust). Per-session behavioral telemetry (gaze, head motion, tracking validity, and firing behavior; listed in §6) is used by the decoder.

Coupling analysis. A correlation pooled over the 66 sessions confounds the three sources of covariation noted above, which can even differ in sign. For each objective metric x and subjective outcome y (participant i, session j) we therefore fit a within–between (Mundlak) regression [3, 12] that separates the within- and between-person components by entering the person mean $\bar{x}_i$ as its own predictor:

$$y_{ij} = \beta_w\,(x_{ij} - \bar{x}_i) + \beta_b\,\bar{x}_i + u_i + \varepsilon_{ij}, \quad (3)$$

where β_w is the within-person (state) slope, β_b the between-person (trait) slope, and u_i a participant intercept; all variables are z-scored so the two slopes are comparable. We complement Eq. (3) with: (i) the intraclass correlation $\text{ICC} = \sigma_b^2/(\sigma_b^2+\sigma_w^2)$, the trait share of each outcome's variance; (ii) the repeated-measures correlation $r_{\text{rm}} = \text{sign}(b)\sqrt{\text{SS}_x/(\text{SS}_x+\text{SS}_{\text{err}})}$, the common within-subject slope b from $y \sim x + \text{factor}(i)$ [2]; (iii) a within-person multiple regression of each y on all four metrics jointly (partial standardized coefficients) to identify its dominant driver; (iv) a within-person mediation splitting latency's total effect c on y into a direct effect c' and an indirect effect $c - c'$ routed through performance; and (v) a calibration test contrasting cross-person (leave-one-participant-out) with per-user (leave-one-session-out, own baseline known) predictability. The participant is the unit of generalization, so every estimate carries a 95% participant-bootstrap interval (B=2000).

Ethics. Participation followed informed consent. The human-subject data collection was approved by the Pennsylvania State University Institutional Review Board under Protocol STUDY00028749; data are de-identified.

Table 2: Per-condition means and within-subject contrasts (n=22). Harm = Degraded−Baseline; Recov. = Compensated−Degraded. $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$, $^{\text{NS}}$ not sig.

Outcome	Base.	Degr.	Comp.	Harm	Recov.
accuracy	0.37	0.25	0.29	-0.11^{***}	$+0.04^{\text{NS}}$
defeated	64.9	44.2	58.3	-20.7^{***}	$+14.1^{***}$
avg_ttd (s)	2.14	2.76	2.36	$+0.62^{***}$	-0.40^{***}
reached_user	0.50	6.18	2.45	$+5.68^{***}$	-3.73^{**}
engagement	5.57	4.40	5.00	-1.17^{***}	$+0.60^{*}$
trust	4.52	3.44	4.24	-1.08^{***}	$+0.80^{**}$

4 Latency Degrades Performance and Experience; Compensation Partially Recovers Both

Within-subject paired tests (Wilcoxon signed-rank, confirmed with paired t; n=22) show the Degraded condition significantly harms *every* outcome ($p<.001$ throughout) and Compensated significantly recovers most of them, while a significant residual gap from Baseline remains on all but one (Table 2). The effects are large. Throughput (defeated) falls 32% under latency ($64.9 \rightarrow 44.2$) and compensation returns it about two-thirds of the way (+14.1 of the 20.7 lost, ≈68% recovered); find-to-defeat time lengthens 29% and recovers similarly; and breakthroughs (reached_user) rise more than ten-fold ($0.5 \rightarrow 6.2$) before compensation roughly halves them. Crucially, the *subjective* outcomes move in lockstep with the objective ones: engagement and trust each drop ∼1.1–1.2 points on the 1–7 scale and recover from about half (engagement, $0.60/1.17 \approx 51\%$) to three-quarters (trust, $0.80/1.08 \approx 74\%$) of that loss.

This recovery is a genuine resolution gain, not an artifact of the fixed session timer. Decomposing each enemy's fate (spawned = defeated + reached-user + still-alive-at-end), spawn counts and session duration are essentially constant across conditions, yet enemies left stranded in transit at session end jump from 5.9 (Baseline) to 17.3 (Degraded) and fall back to 8.7 (Compensated): compensation lets players resolve targets again rather than merely running out the clock. That every objective *and* subjective outcome degrades and partially recovers *together* is the first sign of the coupling we quantify next: is the subjective drop merely a re-encoding of the visible performance loss, or something more?

5 The Objective–Subjective Coupling Is a Within-Person Effect

Trait vs. state. The subjective outcomes are about half trait: the between-person ICC is 0.49 (engagement) and 0.44 (trust), versus much lower ICCs for the objective metrics (e.g. defeated 0.07, avg_ttd 0.10) which are dominated by the condition. Roughly half of each rating's variance is thus stable across a person's sessions; the other, state half is what can couple to performance.

The coupling is within-person. Separating the three sources with Eq. (3), across all four objective metrics the within-person slope β_w is sizable and its 95% interval excludes zero (Figure 1, Table 3): defeating more enemies than one's own norm predicts higher engagement (β_w=0.47 [0.27, 0.71]) and trust (0.45 [0.27, 0.72]). Repeated-measures correlation agrees — defeated couples within-person with engagement at r=0.63 [0.43, 0.81] and trust at 0.58 [0.42, 0.73], with

$|r|$=0.45–0.63 across pairs — and **all eight within-person couplings survive Benjamini–Hochberg FDR correction** (q<.05). The between-person slopes β_b, by contrast, are *not* detectable: every interval spans zero (e.g. 0.34 $[-0.22, 0.82]$). With 22 participants the between-person estimates are imprecise, and a formal within–between contrast is inconclusive for seven of eight pairs, so we claim a within-person coupling that the cross-person view *cannot recover*, not a proven trait/state difference. Either way the asymmetry explains the cross-person null (below and §6): when a player beats their own baseline they feel more engaged and trusting, but that signal is invisible to a model that has never seen the person.

What drives which feeling. A within-person multiple regression (all four metrics jointly, standardized) shows a *suggestive* differential: **engagement** loads most on defeated (partial β=+0.27, accomplishment) and **trust** most on avg_ttd (−0.32, responsiveness — how quickly the system lets the user resolve a target). We flag this as a hypothesis rather than a tested contrast: the competing partial coefficients are close.

A hidden reflection. If the ratings merely re-encoded the scoreboard, latency's effect on them would run *through* measured performance. A within-person mediation (latency → performance → rating; indirect effect by the difference method $c - c'$, with percentile participant-bootstrap CIs) instead finds a *sizable direct path*: the point estimate routes only ≈ 0.40 (engagement) and ≈ 0.36 (trust) of latency's effect through the four objective metrics (Table 4, Figure 2). We do *not* claim the direct path dominates — at n=22 the proportion-mediated interval spans 0 to 1, so the split is not estimable. What is robust is that a direct effect is *present* and that the ratings fall by more than the four measured metrics alone predict. Because only latency was randomized (performance is an observational mediator), this decomposition is associational, not causal. A residual analysis (rating minus the part explained by objective performance) does not robustly track our measured eye-tracker timing proxies, so the direct signal is better read as the degraded *experience* as a whole than as any single instrument channel.

From unpredictable to calibrated. The dissociation has a practical reading. A cross-person predictor of the subjective ratings from objective performance, evaluated leave-one-participant-out, is essentially at chance ($R^2 \approx -0.05$ for both) — exactly because the between-person link is null. But once a participant's own baseline is available (leave-one-session-out, with the baseline computed from the participant's *other* sessions so the held-out rating never enters its own predictor), the same predictor reaches a *modest* R^2=0.37 (engagement) and 0.28 (trust) (Table 4). The two regimes differ in both training population and access to a baseline, so this is not a clean ablation; given the ratings' trait share (ICC ≈ 0.44–0.49), much of the gain is simply knowing a person's level. The takeaway is deliberately limited: subjective state is *within-person* reachable, given a lightweight per-user baseline.

Item level and shape. The coupling holds item by item; the most latency-sensitive item is the engagement *responsiveness* item (E2; mean drop 2.1 Likert points), followed by the two direct trust items (T1/T2, ∼1.3 points) — consistent with latency being felt foremost as lost responsiveness. A within-person quadratic test finds no significant inverted-U ("flow") in performance (smallest p=.08 of eight tests; underpowered): worse performance simply meant lower engagement, monotonically.

6 Objective Degradation Is Decodable from Telemetry

We now ask whether latency's *objective* cost can be read off behavioral telemetry alone — the basis for a runtime monitor. We fit a single multi-output random forest (400 trees, depth 4) that predicts all six outcomes jointly: each leaf stores the full outcome vector, so one fitted model shares structure across the correlated metrics, and targets are standardized per training fold so no single scale dominates the split criterion. Inputs are the per-session behavioral telemetry of §3 (eye-tracker gaze fractions and dwell, head angular velocity and path length, per-eye validity, fire rate, time-to-first-kill) plus the known latency level — a legitimate operational input, since a system applying compensation knows its own configuration. We evaluate leave-one-participant-out (a participant's three sessions held out together), the hard cold-start regime in which the model has never seen the test user.

The objective degradation decodes with modest accuracy (Table 5). Because the latency level could let the forest merely read back the experimenter-set condition, we ablate it. **Behavioral telemetry alone**, with no latency input, already attains out-of-fold R^2=0.33–0.55 across the four metrics, well above a *latency-level-only* baseline (0.12–0.36); adding the latency level lifts this only to 0.39–0.59. The telemetry thus carries most of the signal (the decoder reads what the player *does* under latency), and block-permutation importance localizes it to *firing behavior* (fire rate, time-to-first-kill) and gaze dwell rather than gaze-data quality. The same model assigns near-zero out-of-fold R^2 to engagement and trust (0.10, 0.09), mirroring the between-person null of §5: their state is not cross-person decodable and needs the calibration above. Behavioral telemetry could therefore support QoE monitoring of the *objective* cost of latency, with the *subjective* cost requiring a per-user baseline — subject to the one-task, one-cohort, one-compensation-stack caveat of §7.

7 Discussion

Trust carries what the score hides. The headline is a dissociation: objective performance and subjective experience are tightly coupled *within* a person yet not recoverable *across* people, and a sizable share of latency's effect on engagement and trust is direct, beyond the measured score. For XR systems this means a latency-compensation layer that restores throughput can still leave a trust deficit — in Table 2 Compensated returns defeated most of the way to Baseline, yet engagement and trust remain visibly below it. Monitoring only objective QoE may therefore *overestimate* how well users are being served.

What trust may be tracking. The within-person multiple regression offers a tentative handle on *why* the felt signal is partly independent of the score: engagement appears to load most on defeated (accomplishment) and trust most on avg_ttd (responsiveness; not a significant differential, §5). Latency lengthens avg_ttd directly, giving trust a route to fall without the final score moving, consistent with the direct path in the mediation. If it holds, trust is the more responsiveness-sensitive — and thus security-relevant — of the two constructs.

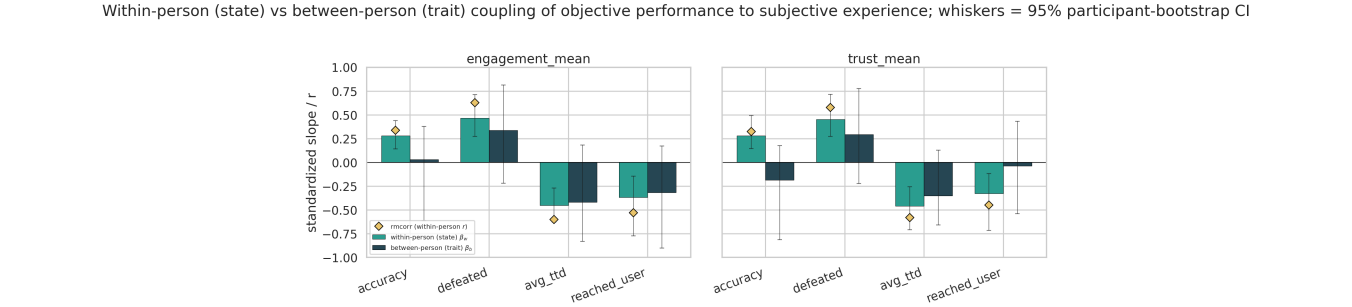


Figure 1: Within-person (state, β_w) vs. between-person (trait, β_b) standardized coupling of each objective metric to engagement (left) and trust (right); gold diamonds are the repeated-measures correlation. Within-person coupling is strong and excludes zero; between-person coupling is not detectable (its CIs span zero, but the within–between contrast is inconclusive at this n). Whiskers: 95% participant-bootstrap CI.

Table 3: Objective–subjective coupling, per pair: within-person slope β_w and between-person slope β_b (standardized, Mundlak), and repeated-measures correlation (rmcorr), each with a 95% participant-bootstrap CI. Every β_w and rmcorr excludes zero (all rmcorr survive Benjamini–Hochberg FDR, q<.05); every β_b includes zero; the within–between contrast is inconclusive at n=22 (CIs span zero for 7/8 pairs).

Subjective	Objective	β_w (within) [95% CI]	β_b (between) [95% CI]	rmcorr [95% CI]
engagement	defeated	+0.47 [+0.27, +0.71]	+0.34 [−0.22, +0.82]	+0.63 [+0.43, +0.81]
	avg_ttd	−0.45 [−0.63, −0.27]	−0.42 [−0.83, +0.18]	−0.60 [−0.77, −0.38]
	reached_user	−0.37 [−0.77, −0.15]	−0.32 [−0.90, +0.17]	−0.53 [−0.78, −0.28]
	accuracy	+0.28 [+0.14, +0.44]	+0.03 [−0.61, +0.38]	+0.34 [+0.17, +0.50]
trust	defeated	+0.45 [+0.27, +0.72]	+0.29 [−0.22, +0.78]	+0.58 [+0.42, +0.73]
	avg_ttd	−0.46 [−0.71, −0.26]	−0.35 [−0.66, +0.13]	−0.58 [−0.73, −0.39]
	reached_user	−0.33 [−0.72, −0.12]	−0.04 [−0.54, +0.43]	−0.45 [−0.68, −0.23]
	accuracy	+0.28 [+0.15, +0.49]	−0.19 [−0.81, +0.18]	+0.32 [+0.15, +0.55]

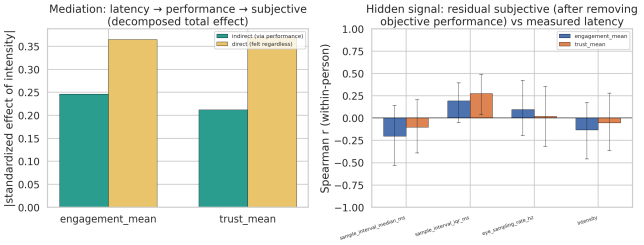


Figure 2: Left: latency's effect on each rating decomposes into a smaller indirect path (via objective performance) and a larger *direct* path (felt regardless of score). Right: the residual rating, after removing objective performance within-person, does not robustly track the measured eye-tracker timing channels. Whiskers: 95% participant-bootstrap CI.

Table 4: Mediation (standardized effect of latency on the rating; participant-bootstrap CI) and calibration (R^2 predicting the rating cross-person vs. with a per-user baseline).

	Engagement	Trust
Total effect of latency	−0.61	−0.58
direct (felt)	−0.37	−0.37
indirect (via perf.)	−0.25	−0.21
Prop. mediated [95%]	0.40 [−.50, 1.37]	0.36 [−.47, 1.28]
R^2 cross-person (cold)	−0.05	−0.05
R^2 per-user calibrated	**0.37**	**0.28**

Table 5: Decoding the four objective metrics (out-of-fold R^2, leave-one-participant-out, single multi-output RF). Behavioral telemetry alone already beats a latency-only baseline; adding the latency level helps only marginally. The subjective ratings are near-zero either way (trust 0.10, engagement 0.09; cross-person).

Target	Latency-only	**Telemetry only**	Telemetry + latency
defeated	0.31	**0.55**	0.59
avg_ttd	0.36	**0.50**	0.57
accuracy	0.12	**0.39**	0.44
reached_user	0.27	**0.33**	0.39

A methodological caution for XR experience studies. Pooling sessions and correlating a performance metric with a satisfaction or trust rating mixes a within-person (state) effect with a between-person (trait) effect that here is null — and the two need not even share a sign. A single pooled correlation would have understated the genuine within-person coupling (Table 3) while implying a cross-person relationship we cannot support. For repeated-measures XR studies, within–between decomposition or rmcorr should be the default.

Per-user calibration as a lightweight mechanism. The calibration result (Table 4) turns the "unpredictable" finding into a design pattern: a single per-participant baseline session lifts subjective-state prediction from chance to R^2=0.28–0.37 — mirroring personalization gains reported for cybersickness prediction [19]. A deployed XR system could thus keep an on-device estimate of

a user's engagement/trust state and flag when responsiveness has degraded enough to move it, without storing survey responses. We stress that this telemetry is *not* benign: head, gaze, and hand motion are biometric-grade and can uniquely identify XR users [11, 13], so any such monitor carries a privacy cost and would need to keep raw streams on-device.

Relevance to trust and security. We frame latency as a source-agnostic operating condition, not an attack, but the trust finding has a security reading. A party that inflates the latency budget (e.g. via congestion or contention) would degrade the experience and, by our results, erode user trust *beyond* the measurable performance loss — an effect a throughput-only monitor misses and a compensation layer only partly masks. Trust is thus both an asset worth protecting and a candidate indicator: a per-user trust estimate that falls without a matching objective drop would signal degraded responsiveness that performance monitoring alone cannot see. We evaluate no adversary and make no detection claim here; formalizing a threat model and a detector is future work.

Limitations. The sample is 22 participants on a single shooting task, with demographics not analyzed and condition order randomized but not tested for practice/fatigue; generalization across tasks and populations is open. Engagement and trust are short post-session Likert scales, not fully validated instruments (internal consistency is acceptable but not evidence of construct validity); "trust" here plausibly indexes perceived responsiveness, as the avg_ttd coupling suggests. The latency condition is given as an internal index, not a calibrated millisecond figure, and its measured footprint is the eye-tracking sample rate — a specific mechanism, not a full motion-to-photon delay; Compensated is also a lower-latency condition, so "compensation recovery" is not separated from "less latency." Statistically, the mediation split is unestimable at this n (CI spans 0–1), the between-person nulls reflect limited power rather than proven absence, and no cybersickness screening was administered. Ms-calibrated latency, validated scales, an explicit adversary/detector, and cross-system generalization are the priorities for future work.

8 Conclusion

Induced latency degrades both what users *do* and how they *feel* in mixed reality, and a compensation layer recovers both only partially. The coupling between the two is a within-person state effect: beating one's own baseline raises engagement and trust, while across people the link is not recoverable — which is why a cross-person model cannot read subjective state, yet a per-user baseline can. A sizable part of latency's effect on feeling is direct, beyond what the objective score captures. Trust and engagement are not redundant with performance metrics; they are a distinct signal worth measuring.

References

[1] Nasim Ahmed, Peng Wu, Kaiming Huang, Sungchul Jung, Hansol Rheem, Gang Tan, Mahdi Imani, and Rifatul Islam. 2025. Human Task Performance and Associated Internal States in Extended Reality: A Systematic Review of Cognitive, Psychophysiological, and Physiological Dimensions. *Frontiers in Virtual Reality* 6 (2025), 1589256. https://doi.org/10.3389/frvir.2025.1589256

[2] Jonathan Z. Bakdash and Laura R. Marusich. 2017. Repeated Measures Correlation. *Frontiers in Psychology* 8 (2017), 456. https://doi.org/10.3389/fpsyg.2017.00456

[3] Andrew Bell and Kelvyn Jones. 2015. Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods* 3, 1 (2015), 133–153. https://doi.org/10.1017/psrm.2014.7

[4] Patrick J. Curran and Daniel J. Bauer. 2011. The Disaggregation of Within-Person and Between-Person Effects in Longitudinal Models of Change. *Annual Review of Psychology* 62 (2011), 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

[5] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570

[6] Kaiming Huang, Peng Wu, Mahdi Imani, Tian Lan, and Gang Tan. 2025. Validating Safety Guarantees of LSTM Models in MR Context. In *Proceedings of the 26th International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*. https://doi.org/10.1145/3704413.3765300

[7] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lilienthal. 1993. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology* 3, 3 (1993), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

[8] Zeqi Lai, Y. Charlie Hu, Yong Cui, Linhui Sun, and Ningwei Dai. 2017. Furion: Engineering High-Quality Immersive Virtual Reality on Today's Mobile Devices. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (MobiCom)*. https://doi.org/10.1145/3117811.3117815

[9] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

[10] Luyang Liu, Ruiguang Zhong, Wuyang Zhang, Yunxin Liu, Jiansong Zhang, Lintao Zhang, and Marco Gruteser. 2018. Cutting the Cord: Designing a High-Quality Untethered VR System with Low Latency Remote Rendering. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 68–80. https://doi.org/10.1145/3210240.3210313

[11] Mark Roman Miller, Fernanda Herrera, Hanseul Jun, James A. Landay, and Jeremy N. Bailenson. 2020. Personal Identifiability of User Tracking Data During Observation of 360-Degree VR Video. *Scientific Reports* 10 (2020), 17404. https://doi.org/10.1038/s41598-020-74486-y

[12] Yair Mundlak. 1978. On the Pooling of Time Series and Cross Section Data. *Econometrica* 46, 1 (1978), 69–85. https://doi.org/10.2307/1913646

[13] Vivek Nair, Wenbo Guo, Justus Mattern, Rui Wang, James F. O'Brien, Louis Rosenberg, and Dawn Song. 2023. Unique Identification of 50,000+ Virtual Reality Users from Head and Hand Motion Data. In *Proceedings of the 32nd USENIX Security Symposium*. 895–910.

[14] Heather L. O'Brien and Elaine G. Toms. 2010. The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69. https://doi.org/10.1002/asi.21229

[15] Lisa Rebenitsch and Charles Owen. 2016. Review on Cybersickness in Applications and Visual Displays. *Virtual Reality* 20, 2 (2016), 101–125. https://doi.org/10.1007/s10055-016-0285-9

[16] Jan-Philipp Stauffert, Florian Niebling, and Marc Erich Latoschik. 2020. Latency and Cybersickness: Impact, Causes, and Measures. A Review. *Frontiers in Virtual Reality* 1 (2020), 582204. https://doi.org/10.3389/frvir.2020.582204

[17] Sophie Stephenson, Bijeeta Pal, Stephen Fan, Earlence Fernandes, Yuhang Zhao, and Rahul Chatterjee. 2022. SoK: Authentication in Augmented and Virtual Reality. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. https://doi.org/10.1109/SP46214.2022.9833742

[18] Bob G. Witmer and Michael J. Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments* 7, 3 (1998), 225–240. https://doi.org/10.1162/105474698565686

[19] Peng Wu, Nasim Ahmed, Kaiming Huang, Rifatul Islam, Tian Lan, Gang Tan, and Mahdi Imani. 2025. Personalized Bayesian Networks for Cybersickness Prediction in Virtual Reality. In *Proceedings of the 26th International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*. 491–496. https://doi.org/10.1145/3704413.3765310

[20] Peng Wu, Nasim Ahmed, Abhiram Sarma, Kaiming Huang, Rifatul Islam, Bin Li, Tian Lan, Gang Tan, and Mahdi Imani. 2025. Probabilistic Verification of Cybersickness in Virtual Reality Through Bayesian Networks. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*.