

Lab or Field? A Method-Comparison Study of Lighthouse and CrUX Measurements on the Turkish Web

1st.com.tr Research Unit, 1st.com.tr, İstanbul, Türkiye

DOI: <https://doi.org/10.5281/zenodo.20732713> · **Published:** 10 June 2026 · **Series:** TR-WEB Working Papers, 1 (2026) · **License:** CC BY 4.0

ABSTRACT

Objective: To determine how well laboratory (Lighthouse, throttled) performance measurements predict real-user field (Chrome UX Report, CrUX p75) experience. **Method:** For 7,403 Turkish sites with laboratory (Lighthouse) and field (CrUX p75) values from the same PageSpeed Insights call, Largest Contentful Paint (LCP), First Contentful Paint (FCP) and Cumulative Layout Shift (CLS) were matched; agreement was assessed with Pearson and Spearman correlations, Bland-Altman analysis, and Cohen's kappa over the Core Web Vitals 'good/needs-improvement/poor' classification. **Results:** The laboratory systematically and substantially over-reported loading metrics (LCP: lab median 10,876 ms vs field 2,247 ms). The lab-field relationship for LCP was weak ($p = 0.26$) and classification agreement did not exceed chance ($\kappa \approx 0.00$); 85.6% of sites were classified worse in the lab than in the field, and while only 3.1% of sites were 'good' in the lab, 59.6% were 'good' in the field. FCP showed similarly weak agreement ($\kappa = 0.01$); only CLS reached moderate agreement ($\kappa = 0.33$). The overall Lighthouse score correlated only weakly with field LCP ($r = -0.25$). **Conclusion:** On the Turkish web, default laboratory Lighthouse loading metrics are a poor proxy for real-user experience; performance decisions should rest on field data, with the laboratory reserved for diagnosis.

Keywords: Core Web Vitals, Lighthouse, CrUX, lab-field agreement, Bland-Altman, Cohen's kappa, LCP, method comparison

1. Introduction

Web performance is measured from two basic sources: controlled, synthetic laboratory tests (e.g., Lighthouse) and field data collected from real users (e.g., the Chrome UX Report, CrUX). A laboratory test provides a reproducible diagnosis from a single throttled run, whereas field data report the 75th percentile (p75) of the real distribution of devices, networks and geographies. The two are related but not identical quantities.

In practice many teams base decisions on laboratory scores alone; yet how well the laboratory represents real experience is an empirical question. This study asks: on the Turkish web, are laboratory Lighthouse measurements a valid proxy for real-user field experience? We answer it by centring a single feature, Core Web Vitals, and applying method-comparison statistics.

2. Related Work

Core Web Vitals (LCP, FCP, CLS, INP) are established indicators of user experience; Google's guidance positions field data as the 'ground truth' and the laboratory as a diagnostic tool. To test the agreement of two measurement methods, the Bland & Altman (1986) approach (bias plus limits of agreement) from the clinical-measurement literature and, for categorical agreement, Cohen's (1960) kappa are standard; kappa is interpreted with the Landis & Koch (1977) benchmarks. The contribution of this study is to apply this method-comparison framework to lab-field CWV agreement across a large population of Turkish sites.

3. Method

The sample comprises 7,403 Turkish sites audited by the 1st.com.tr analysis engine that had both laboratory and CrUX field data (field data were available for 53.7% of all 13,775 audited sites). Both laboratory and field values were obtained from a single PageSpeed Insights (PSI) API call (strategy=mobile): the laboratory values from the Lighthouse run PSI executes (lighthouseResult; mobile device emulation with network and CPU throttling) and the field values from the CrUX data in the same response (loadingExperience/originLoadingExperience; 28-day p75). Because CrUX data are present only in a successful PSI response, every site in the matched sample takes both its laboratory and field numbers from the same PSI call; the comparison is therefore between the laboratory and field numbers that PageSpeed Insights presents side by side for the same origin. For each site, LCP, FCP (milliseconds) and CLS (unitless) laboratory-field pairs were matched.

An important distinction: the laboratory measures a single synthetic run, whereas the field measures the p75 of the real-user distribution. The comparison is therefore not 'do they measure the same thing' but 'is the laboratory a useful proxy for the field'.

4. Statistical Analysis

Three complementary methods were used. (1) Association: Pearson (p1-p99 winsorized to limit outlier influence) and distribution-free Spearman correlations, each with a 95% confidence interval (Fisher z) and two-tailed p. (2) Agreement: Bland-Altman analysis; bias (mean lab – field difference) and 95% limits of agreement (bias $\pm$ 1.96-SD); the median absolute difference was also reported as an outlier-robust measure. (3) Class agreement: using Core Web Vitals thresholds (LCP 2,500/4,000 ms; FCP 1,800/3,000 ms; CLS 0.10/0.25), each measurement was classified 'good/needs-improvement/poor' and Cohen's kappa computed (interpreted per Landis & Koch 1977). Because significance is near-guaranteed at large n, interpretation rests on effect/agreement magnitude.

5. Results: Laboratory and Field Values

The laboratory reported loading metrics strikingly higher (more pessimistic) than the field (Table 1). For LCP the laboratory median is roughly five times the field median; for FCP the gap is smaller but marked. For CLS the two sources gave the closest values.

METRIC	LAB MEDIAN	FIELD MEDIAN	LAB MEAN	FIELD MEAN
LCP (ms)	10,876	2,247	14,769	2,515
FCP (ms)	3,785	1,649	4,466	1,941
CLS	0.008	0.010	0.130	0.096

Table 1. Central-tendency values in the laboratory and the field for matched sites (n = 7,403). LCP/FCP in ms, CLS unitless.

6. Results: Correlation

The relationship between laboratory and field values is weak for loading metrics: Spearman $\rho = 0.26$ for LCP and $\rho = 0.17$ for FCP. Only layout stability (CLS) showed a moderate relationship ($\rho = 0.41$). The overall Lighthouse performance score correlated with field LCP and INP in the expected direction but weakly ($r \approx -0.25$): a higher laboratory score is only loosely associated with better real-user values (Table 2).

RELATIONSHIP	PEARSON R	95% CI	SPEARMAN P	P
LCP (lab ↔ field)	0.21	[0.19; 0.23]	0.26	<0.0001
FCP (lab ↔ field)	0.15	[0.13; 0.17]	0.17	<0.0001
CLS (lab ↔ field)	0.43	[0.41; 0.45]	0.41	<0.0001
LH score ↔ field LCP	-0.25	[-0.27; -0.23]	-0.28	<0.0001
LH score ↔ field INP	-0.25	[-0.27; -0.22]	-0.27	<0.0001

Table 2. Laboratory-field correlations (Pearson winsorized; Spearman robust). CI: Fisher-z 95% confidence interval.

7. Results: Bland-Altman Agreement

The Bland-Altman analysis reveals wide, unusable disagreement for loading metrics (Table 3). For LCP the laboratory exceeds the field by 12,254 ms on average (bias); the 95% limits of agreement span from -18.4 s to +42.9 s, so the true field LCP cannot, in practice, be inferred from a single laboratory LCP. Even the outlier-robust median absolute difference is 8,384 ms for LCP. For CLS the bias is small (+0.034) but the limits of agreement remain wide.

METRIC	BIAS	95% LIMITS OF AGREEMENT	MEDIAN DIFF
LCP (ms)	+12,254	[-18,377; +42,886]	8,384
FCP (ms)	+2,524	[-3,738; +8,786]	2,050
CLS	+0.034	[-0.46; +0.53]	0.02

Table 3. Bland-Altman agreement analysis (lab - field). Bias: mean difference; limits of agreement: bias ± 1.96·SD.

8. Results: Class Agreement (Cohen's Kappa)

For the Core Web Vitals 'good/needs-improvement/poor' classification, agreement between laboratory and field does not exceed chance for loading metrics (Table 4). For LCP, $\kappa = 0.00$: the laboratory classification predicts the field classification no better than chance. The most striking result is the asymmetry: 85.6% of sites fell into a worse class in the laboratory than in the field; while only 3.1% of sites appeared 'good' for LCP in the laboratory, 59.6% were 'good' in the field. FCP is similar ($\kappa = 0.01$). Only CLS shows moderate agreement ($\kappa = 0.33$; 72.8% of sites in the same class).

METRIC	K	LAB WORSE %	SAME CLASS %	LAB 'GOOD' %	FIELD 'GOOD' %
LCP	0.00	85.6	13.1	3.1	59.6
FCP	0.01	75.4	20.5	7.7	58.3
CLS	0.33	16.1	72.8	73.3	76.5

Table 4. CWV class agreement (Cohen's kappa). 'Lab worse %': share of sites the laboratory placed in a worse class than the field.

9. Discussion

The findings paint a consistent picture: on the Turkish web, default laboratory Lighthouse loading metrics are a poor proxy for real-user experience. The weak correlation (LCP $\rho = 0.26$), the unusable Bland-Altman limits of agreement, and the chance-level kappa (LCP $\kappa = 0.00$) all point the same way. The laboratory's systematic pessimism is explained by the default mobile-throttled run, cold cache, and the fact that a single run is not the p75.

Layout stability (CLS) is the exception: laboratory and field agree moderately ($\kappa = 0.33$), because CLS is largely a structural property of the document, independent of network conditions. The practical implication is clear: performance decisions and rankings should rest on field (CrUX) data; the laboratory should be used to diagnose regressions locally. Ranking the field by the laboratory's overall score ($r \approx -0.25$) would mis-position most sites.

10. Limitations

(i) The laboratory measures a single synthetic run while the field measures the p75; this difference in estimand naturally explains part of the observed disagreement, but from a practitioner's standpoint the question remains whether the laboratory is a useful proxy, so the finding stands. (ii) Field data are available only for the higher-traffic 53.7% subset (selection bias). (iii) The laboratory used the default configuration; different throttling settings may yield different bias. (iv) The study is observational and cross-sectional. (v) Because of outliers, the primary interpretation rests on robust measures (Spearman, median difference, kappa).

11. Ethics and Data Use

All measurements rest on automated audits of public pages and on public CrUX aggregates; no personal data are collected and results are aggregated. The study is published open-access under CC BY 4.0; data are shared free for academic purposes on request.

12. Conclusion

Across 7,403 Turkish sites, agreement between laboratory and field Core Web Vitals declines from weak to negligible for loading metrics and rises only to moderate for layout stability. Laboratory loading scores are not a reliable proxy for real-user experience and are systematically pessimistic. We recommend practitioners leave the decision to the field and the diagnosis to the laboratory. Future work will examine how different throttling configurations and device classes affect agreement.

Declarations

DATA AVAILABILITY

The anonymised, matched laboratory-field dataset and the reproducible analysis script are shared free for academic purposes on request: akademi@1st.com.tr.

FUNDING

This research received no external funding; it was conducted within 1st.com.tr.

CONFLICT OF INTEREST

Laboratory data were collected with the authors' 1st.com.tr engine (disclosed for transparency); field data come from an independent source (Google CrUX). No other conflicts of interest are declared.

References

- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307-310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8)
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46. <https://doi.org/10.1177/001316446002000104>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. <https://doi.org/10.2307/2529310>
- Google. Chrome UX Report (CrUX): origin-level field data. <https://developer.chrome.com/docs/crux>
- Google. Lighthouse: automated tool for improving web page quality. <https://developer.chrome.com/docs/lighthouse/overview>
- web.dev. Core Web Vitals. <https://web.dev/articles/vitals>

HOW TO CITE

1st.com.tr Research Unit (2026). Lab or Field? A Method-Comparison Study of Lighthouse and CrUX Measurements on the Turkish Web. 1st.com.tr Academy: Open Data Working Papers, TR-WEB-2026-04. <https://doi.org/10.5281/zenodo.20732713>