Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring

Wolfram Wingerath & Sebastian Grebasch

Design & Experience
Measure user experience on the web.
Why should you care?

Site speed improvement of just 0.1s leads to an increase in conversion rates & average order value.

50 ms

Google/55/Deloitte, Speed Impact study, EMEA and US, Oct-Nov 2019, n=37 brand sites analysed hourly over a 30 day period totalling 30.5m sessions.
Timing for bringing page experience to Google Search

Tuesday, November 10, 2020

This past May, we announced that page experience signals would be included in Google Search ranking. These signals measure how users perceive the experience of interacting with a web page and contribute to our ongoing work to ensure people get the most helpful and enjoyable experiences from the web. In the past several months, we’ve seen a median 70% increase in the number of users engaging with Lighthouse and PageSpeed Insights, and many site owners using Search Console’s Core Web Vitals report to identify opportunities for improvement.

Today we’re announcing that the page experience signals in ranking will roll out in May 2021. The new page experience signals combine Core Web Vitals with our existing search signals including mobile-friendliness, safe-browsing, HTTPS-security, and intrusive interstitial guidelines.

http://goo.gle/page-experience-timing
Core Web Vitals
Providing unified guidance for quality signals that are essential to delivering a great user experience on the web.
Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring

Performance Metrics

Let us focus on the user-centered core metrics
Prioritising the User Experience

Is it **happening?**

Is it **responsive?**

Is it **stable?**

**LCP**
Largest Contentful Paint

**FID**
First Input Delay

**CLS**
Cumulative Layout Shift
Tool Taxonomy and CRUX
Let’s Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring

How to think about Lab vs. Field speed tools...

Taxonomy of Speed Tooling @ Google

Lab data
(aka metrics from the lab)

Field data
(aka metrics from the wild)
How to think about Lab vs. Field speed tools...

Taxonomy of Speed Tooling @ Google

Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
How does CrUX work?

CrUX provides “field data” for top sites on the web. Field data is performance data collected from real page loads users are experiencing in the wild.

1. Real-world user experience data (real devices, network conditions, etc) is collected from opted-in Chrome users as they browse the web, and uploaded to Google’s URL Keyed Metrics service.

2. UKM data is anonymized, non-public URLs are filtered out and URLs and origins with insufficient number of samples are removed. The end result is consumable, regularly updated data on how real-world users experience sites on the web.

3. CrUX data is made public via a variety of channels: BigQuery, PageSpeed Insights, Search Console, and other 1P & 3P products.
Case Study

**Vodafone A/B Testing AD Landing Pages**

Optimizations for 50% of traffic via different traffic sources

- Image Optimization
- Server-side Rendering of critical HTML & Widget Optimization

![Image of Vodafone mobile app with statistics]

<table>
<thead>
<tr>
<th>Improvement</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCP led to...</td>
<td>31%</td>
</tr>
<tr>
<td>Increase in total sales</td>
<td>+8%</td>
</tr>
<tr>
<td>Uplift in the lead to visit rate</td>
<td>+15%</td>
</tr>
<tr>
<td>Uplift in the cart to visit rate</td>
<td>+11%</td>
</tr>
</tbody>
</table>

Learn more @ [https://web.dev/vodafone/](https://web.dev/vodafone/)

Screenshot Source: [https://www.web.dev/vodafone](https://www.web.dev/vodafone)
Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Real-User Monitoring
We bring performance research to practice

30+ man-years of web performance research at University of Hamburg

7,000 customer websites are already using Speed Kit

$2.6 billion in annual revenue runs with Speed Kit

160 million users per month benefit from Speed Kit

Novel technology for caching dynamic data went into Baqend in 2014

Baqend launched Speed Kit as the all-in-one page speed platform in 2018
The Basic Idea

Real-User Monitoring (RUM)

Let's Agree to Disagree: Why Google's CrUX Results Are Not Reproducible With Your Real-User Monitoring

- Time-to-First-Byte
- First (Contentful) Paint
- DOM Timer
- First Input Delay

Performance

- Session Length
- Time on Site
- First User Interaction
- Bounce Rate

User Engagement

- Page Views & Sessions
- Browser Distribution
- JavaScript Errors
- Caching Insights

QA Metadata

- Timing API
- Service Worker
- Unhandled Errors
Let’s Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Beaconnect: Continuous Processing

Key Topics:
- Continuous Aggregation
- Real-Time Reporting & Analytics
- Extreme Scalability
Beaconnect: Continuous Processing

Key Topics:
- Continuous Aggregation
- Real-Time Reporting & Analytics
- Extreme Scalability

Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring

Analyzing Performance Data
Split Testing for Web Performance

Speed Kit Users vs. Normal Users

- Speed Kit enabled
- Measurable uplift:
  + Performance
  + User engagement
  + ...

- Speed Kit disabled
  (no acceleration)

Split Testing for Web Performance

Speed Kit Users vs. Normal Users

- Speed Kit enabled
- Measurable uplift:
  + Performance
  + User engagement
  + ...

- Speed Kit disabled (no acceleration)

3 Levels of Aggregation

**Partial Page Impressions (PPIs)**
Enhanced Data Beacons

**1-Min. Time Windows**
Immediate Aggregates (Storage)

**Arbitrary Time Windows**
Real-Time Reporting (Dashboard Queries)

<table>
<thead>
<tr>
<th>Time</th>
<th>Browser</th>
<th>Device</th>
<th>Test Group</th>
<th>First Contentful Paint (FCP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:05:04.578</td>
<td>Firefox</td>
<td>Mobile</td>
<td>Speed Kit</td>
<td>127ms</td>
</tr>
<tr>
<td>11:06:48.139</td>
<td>Chrome</td>
<td>Mobile</td>
<td>Original</td>
<td>958ms</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Browser</th>
<th>Device</th>
<th>Test Group</th>
<th>First Contentful Paint (FCP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:05</td>
<td>Firefox</td>
<td>Mobile</td>
<td>Speed Kit</td>
<td>{200ms: 1, 500ms: 2}</td>
</tr>
<tr>
<td></td>
<td>Firefox</td>
<td>Mobile</td>
<td>Original</td>
<td>{600ms: 2, 800ms: 5}</td>
</tr>
<tr>
<td></td>
<td>Safari</td>
<td>Desktop</td>
<td>Original</td>
<td>{1100ms: 1}</td>
</tr>
<tr>
<td>11:06</td>
<td>Firefox</td>
<td>Mobile</td>
<td>Speed Kit</td>
<td>{200ms: 3}</td>
</tr>
<tr>
<td></td>
<td>Chrome</td>
<td>Mobile</td>
<td>Speed Kit</td>
<td>{400ms: 2}</td>
</tr>
<tr>
<td></td>
<td>Opera</td>
<td>Tablet</td>
<td>Original</td>
<td>{700ms: 1, 1300ms: 2}</td>
</tr>
<tr>
<td></td>
<td>Safari</td>
<td>Desktop</td>
<td>Original</td>
<td>{600ms: 4, 900ms}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Browser</th>
<th>Device</th>
<th>Test Group</th>
<th>First Contentful Paint (FCP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:05</td>
<td>Firefox</td>
<td>Mobile</td>
<td>Speed Kit</td>
<td>{200ms: 4, 500ms: 2}</td>
</tr>
<tr>
<td>11:06</td>
<td>Firefox</td>
<td>Mobile</td>
<td>Speed Kit</td>
<td></td>
</tr>
</tbody>
</table>
Applying Dimension Filters: All Users

Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Applying Dimension Filters: Chrome

Analyzing Performance Data

Let's Agree to Disagree: Why Google's CrUX Results Are Not Reproducible With Your Real-User Monitoring

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Speed Kit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop</td>
<td>1510</td>
<td>1123</td>
</tr>
<tr>
<td>Mobile</td>
<td>1720</td>
<td>1381</td>
</tr>
<tr>
<td>Overall</td>
<td>1671</td>
<td>1324</td>
</tr>
</tbody>
</table>

- **Good (≤ 1.8s)**
  - Original: 54%
  - Speed Kit: 63%
- **Needs improvement (1.8s < 3.0s)**
  - Original: 25%
  - Speed Kit: 19%
- **Poor (> 3.0s)**
  - Original: 21%
  - Speed Kit: 18%

![Page Impressions Chart](chart.png)
Analyzing Performance Data

Applying Dimension Filters: Chrome, Product Pages

Let’s Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Applying Dimension Filters: Chrome

Let’s Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
CrUX vs. RUM
RUM vs. CrUX

Why Google’s CrUX Data is a Black Box

CrUX Tracking Funnel

Filter 1
Only Chrome User (Excluding WebView Experiences)
reproducible with your RUM

Filter 2
Only Logged In Users (via Google Chrome Profile)

Filter 3
Only Users with Active Browser History Sync & no Passphrase

Filter 4
Only Users with Enabled Usage Statistic Reporting


Let’s Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Let's Agree to Disagree: Why Google's CrUX Results Are Not Reproducible With Your Real-User Monitoring

**CrUX vs. RUM**

**Wrapup**

<table>
<thead>
<tr>
<th>CrUX</th>
<th>vs.</th>
<th>RUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on field data</td>
<td></td>
<td>Based on field data</td>
</tr>
<tr>
<td>Public data</td>
<td></td>
<td>Custom deployment</td>
</tr>
<tr>
<td>=&gt; zero-effort</td>
<td>=&gt; complex to operate</td>
<td></td>
</tr>
<tr>
<td>=&gt; competitor data available</td>
<td>=&gt; Just your own website</td>
<td></td>
</tr>
<tr>
<td>Fixed granularity</td>
<td></td>
<td>Complete freedom</td>
</tr>
<tr>
<td>=&gt; only by month / last 28 days</td>
<td>=&gt; real-time / full detail</td>
<td></td>
</tr>
<tr>
<td>=&gt; only fixed dimensions</td>
<td>=&gt; custom dimensions</td>
<td></td>
</tr>
<tr>
<td>=&gt; only (part of) Chrome users</td>
<td>=&gt; all browsers</td>
<td></td>
</tr>
</tbody>
</table>

Neither gives you the full picture!
CrUX + RUM!
CrUX 28 Days Rolling Window: After 2 Days (7%)

After 2 Days of Optimization

Two days after the 100% Rollout of an optimization, the Google CrUX 28 Days Report still includes 26 days (93%) of not optimized performance.

The 2 Days (7%) of optimized performance can only impact the overall result slightly.

What Google Reports That Day

- Poor: 13%
- Needs Improvement: 73%
- Good: 14%

What Google Aggregates In Its Report That Day (But doesn't report in that detail)

- 93% Not Optimized (26 of 28 Days)
- 7% Optimized (2 of 28 Days)

Still looks bad!

Looks good already!
After 7 Days of Optimization

One week after the 100% Rollout of an optimization, the Google CrUX 28 Days Report still includes 21 days (75%) of not optimized performance.

The 7 Days (25%) of optimized performance are still not dominant in the overall result.

What Google Reports That Day

- Poor: 11%
- Needs Improvement: 63%
- Good: 26%

What Google Aggregates In Its Report That Day (But doesn’t report in that detail)

- 75% Not Optimized (21 of 28 Days)
- 25% Optimized (7 of 28 Days)
CrUX + RUM!

CrUX 28 Days Rolling Window: After 14 Days (50%)

After 14 Days of Optimization

Two weeks after the 100% Rollout of an optimization, the Google CrUX 28 Days Report contains 14 days (50%) of not optimized and 14 Days (50%) of optimized performance.

From now on the impact of the optimization will be at least dominant in the overall result.

What Google Reports That Day

- Poor: 9%
- Needs Improvement: 46%
- Good: 46%

What Google Aggregates In Its Report That Day (But doesn’t report in that detail)

- 50% Not Optimized (14 of 28 Days)
- 50% Optimized (14 of 28 Days)
CrUX + RUM!

CrUX 28 Days Rolling Window: After 28 Days (100%)

**After 28 Days of Optimization**

Only when the page optimization is live for **28 days (100%)** the effect will be fully reflected in the Google CrUX 28 Days Report.

**What Google Reports That Day**

- **Poor**: 5%
- **Needs Improvement**: 10%
- **Good**: 85%

**What Google Aggregates In Its Report That Day**

- **100% Optimized (28 of 28 Days)**

**Optimization Rollout to 100% Page Impressions**

Let's Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Before Optimization
Before the rollout of the optimization, the First Contentful Point (FCP) of the last 28 Days is obtained from the CrUX API.

28 Days Rolling Window: FCP CrUX Uplift Example

<table>
<thead>
<tr>
<th>Sep 1 2022 CrUX API</th>
<th>CrUX FCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 Days Not Optimized</td>
<td>2,301 ms</td>
</tr>
</tbody>
</table>

7 Days after Optimization
After 7 Days of optimizations, the 28 Days FCP reported by the CrUX API is already 202 ms faster.

<table>
<thead>
<tr>
<th>Sep 7 2022 CrUX API</th>
<th>CrUX FCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 Days Not Optimized</td>
<td>2,100 ms</td>
</tr>
<tr>
<td>7 Days Optimized</td>
<td>-202 ms</td>
</tr>
</tbody>
</table>

28 Days after Optimization
Extrapolated to 28 Days of Optimization (7 Days Uplift * 4) the FCP is expected to become 806 ms faster compared to the FCP before the optimization.

<table>
<thead>
<tr>
<th>Sep 28 2022 FORECAST</th>
<th>CrUX FCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 Days Optimized</td>
<td>1,495 ms</td>
</tr>
<tr>
<td></td>
<td>-806 ms</td>
</tr>
</tbody>
</table>
## Let’s Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring

<table>
<thead>
<tr>
<th>Website</th>
<th>CrUX Score</th>
<th>RUM Score</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>mindfactory.de</td>
<td>96%</td>
<td>2%</td>
<td>-3%</td>
</tr>
<tr>
<td>amazon.de</td>
<td>93%</td>
<td>5%</td>
<td>-2%</td>
</tr>
<tr>
<td>lapstore.de</td>
<td>92%</td>
<td>6%</td>
<td>-2%</td>
</tr>
<tr>
<td>alternate.de</td>
<td>90%</td>
<td>7%</td>
<td>-3%</td>
</tr>
<tr>
<td>pearl.de</td>
<td>84%</td>
<td>11%</td>
<td>-4%</td>
</tr>
<tr>
<td>mediarkt.de</td>
<td>73%</td>
<td>17%</td>
<td>-10%</td>
</tr>
<tr>
<td>euronics.de</td>
<td>70%</td>
<td>20%</td>
<td>-10%</td>
</tr>
<tr>
<td>caseking.de</td>
<td>69%</td>
<td>23%</td>
<td>-8%</td>
</tr>
<tr>
<td>reichelt.de</td>
<td>56%</td>
<td>25%</td>
<td>-9%</td>
</tr>
<tr>
<td>notebooksbilliger.de</td>
<td>55%</td>
<td>23%</td>
<td>-22%</td>
</tr>
<tr>
<td>cyberport.de</td>
<td>49%</td>
<td>32%</td>
<td>-29%</td>
</tr>
<tr>
<td>conrad.de</td>
<td>45%</td>
<td>26%</td>
<td>-29%</td>
</tr>
</tbody>
</table>
Roadmap Features
**Dashboard KPI Overview**

**Purpose**
Simple overview of the technical Speed Kit KPIs

**Approach**
The dashboard will provide a view to display the technical KPIs:
- See your current traffic usage and how many cache hits you have
- See how the performance of your site behaves at a single glance
Speed Analytics Dashboard

**Purpose**
A simple overview of the main performance metrics

**Approach**
An easy-to-use performance dashboard based on monthly RUM data:
- Status quo and over time development of core web vitals
- Drill-down by origin, device and page types
- Can be extended by SQL workbench for drill-downs

Let’s Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Roadmap Features

Speed Guard

Purpose
Detect performance issues with exact cause and solution

Approach
1. Monitor performance and detect anomalies
2. Find cause by automated dicing and combining RUM and lab data
3. Suggest solution based on caused and web performance best practices

Find responsible dimension

Find responsible point in time in page load process

Suggest Solutions

Find responsible artifacts

+ combine with lab data
Purpose
Suggested speed improvements beyond Speed Kit

Approach
Speed Kit’s real-user monitoring and synthetic testing collect detailed speed data. The advisor analyzes various dimensions (e.g. browsers) and metrics (e.g. CLS) and suggest applicable optimizations.

Website with Speed Kit
- Real-user speed data
- Lab data (WPT & Lighthouse)

Best Practice Database
- Optimizations learned on 200M+ monthly users

Speed Advisor
1. Identify bottlenecks across website
2. Assess impact of potential measures
3. Suggest concrete solutions

Actionable Insights
Example: Remove JavaScript lazy loading of banner image for users coming from Google on page type “PDP” to reduce bad CLS by 23% overall

Let’s Agree to Disagree: Why Google’s CrUX Results Are Not Reproducible With Your Real-User Monitoring
Let's Agree to Disagree: Why Google's CrUX Results Are Not Reproducible With Your Real-User Monitoring
Thanks!

Questions?

research@baqend.com
baqend.com/publications