RideCheck Holdout Group

Safety Effectiveness Measurement

2024 H2

NGRV NIGH GOLDENBERG RASO & VAUGHN

Trial Exhibit No.
P-01447

Uber

Meeting goals

01 Align on Problem Definition

02 Refine Proposed Solutions

03 Kick-off Final Solution Design

Project Overview

PROBLEM STATEMENT

We don't know RideCheck's current (baseline) impact on incident rate, and want to track that for future changes.

PROPOSED SOLUTION

A per-trip, always-on holdout group design allowing us to set a baseline and measure product evolution.

KEY LINKS

PRD | ERD [tbd] | DS Doc

- RideCheck SOT (2024)
- RideCheck SOT (2021)

RESPONSIBLE

Data @issamu.okada

Engineering @harish.shanker

Product @bruno.grimaldi

Product Ops @vidya.daryanani

APPROVERS

Data Leadership @sunny.wong

Product Leadership @mike.akamine

REVIEWERS

Legal @dkolta

Product @mariana.esteves

Safety Ops @alex.madsen

01 Problem Definition

RideCheck is evolving

RideCheck today

simple heuristics → passive interventions

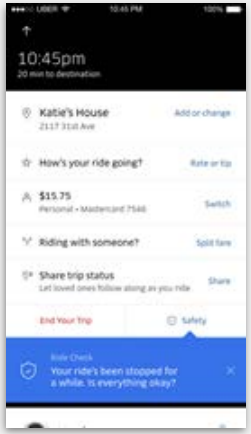
- Car stopped for 5+ minutes on trip
- Trip ends earlier than expected
- Route is too different from our estimate

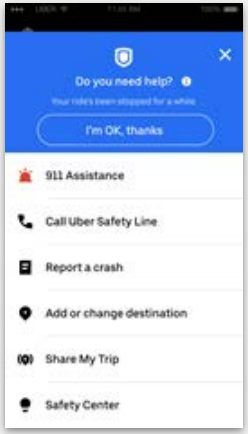
- Push notification
- In-app callout with Safety Tools

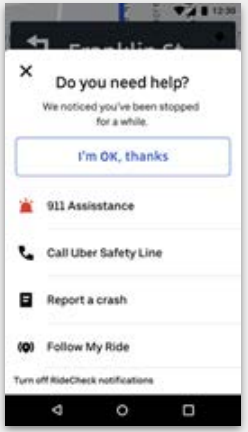
At a glance

Current Heuristics

Heuristic	Description	Trigger Rate	Precision	Recall
Long Stop	Car has been stopped on-trip for a significant amount of time	1.14%	0.36%	7.90%
Midway Drop-Off	Trip completed at significant distance from destination.	0.76%	0.95%	13.70%
Route Deviation	On-trip at significant distance from closest point in "reference route".	0.08%	0.34%	0.50%
Crash Detection	Significant impulse (gyro + accelerometer) followed by long stop.	<0.01%	50.88%	12.25%







At a glance

Current Interventions

RideCheck's interventions acknowledge that the trip is not going according to plan, and surface useful Safety Tools for the Rider and Driver in those moments. At their core they are:

- Low-intrusion
- Educational
- Quality of life
- Empathetic

P-01447.00007

Ops has questions re: RideCheck's effectiveness…

In Serious Incident investigations, we are often asked why RideCheck didn't do more:

- Why it didn't trigger earlier
- Why it didn't escalate actions
- Why it chose to not act at all

…because our understanding of it is changing

On the one hand, RideCheck is a low-intrusion, educational, quality-of-life feature.

On the other hand, it is "the" one platform in our portfolio that is watching trips in real-time and is capable of acting on them.

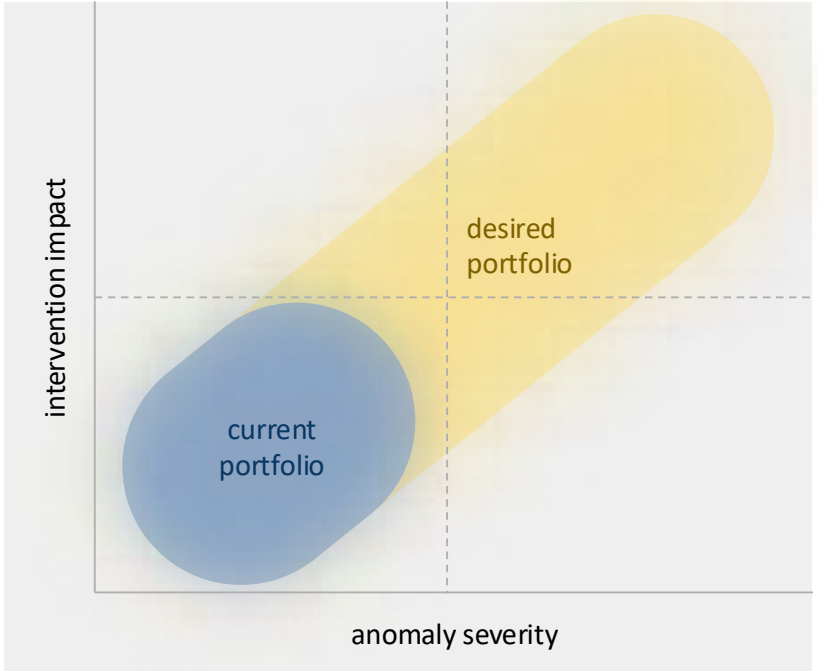
We want to extend RideCheck's portfolio

From quality of life check-ins

The original intent behind RideCheck is not abandoned. We intend to perfect current quality-of-life check-ins, evolving both their heuristics and the check-in experience.

To on-trip Safety policies

Simultaneously, we want to explore how we can use the RideCheck engine to better identify and react to sensitive moments for our users.



To support

simple heuristics → passive interventions

but also

sophisticated heuristics → proactive interventions

Anomaly Progressions

- Allows multiple signals to be considered
- First step towards a live on trip safety model

Destination Lingering

- Suite of anomalies for a post-trip period
- Extends the coverage of RideCheck

Call the User

- Robocalls
- Live Agent-initiated calls*

Escalate

- Internally (eg. CommOps IIT*)
- Externally (eg. Guardian, U4B Operator*)

A "Live Trip Monitoring" mini-program

Baseline Anomaly Improvements

Crash detection ML model

Route Deviation "reference route"

Multiple anomaly Long Stop

High-Severity Anomalies

Anomaly Progressions

Destination Lingering

Robocall for unresolved Long Stops

Robocall for Crash Detection

Escalate to Teen's Guardian

Escalation Paths

Call from Live Agent

Escalate to U4B Operator

Escalate to CommOps IIT

Proactive Escalation


Issamu (Data Science):
But how will we measure the success of this mini-program? We've only ever measured RideCheck as a low-intrusion quality-of-life feature.

We don't know RideCheck's impact on Safety Incidents

On a montly basis, we only check how well RideCheck heuristics model incidents and how often we're sending checks.

In 2021 we did a treatment/control experiment and survey, tracking IPC deterrence potential and a few Safety Sentiment metrics.

Objective	Measured
Trigger Rate Are we sending too many checks?	Monthly
Precision % anomalies that became incidents	Monthly
Recall % incidents that had anomalies	Monthly
Committed to Safety Using our Brand Tracking question	Last in 2021 survey / experiment
Overall IPC Rate Is RideCheck is an IPC deterrent?	Last in 2021 survey / experiment


Issamu (Data Science):
To measure the impact of upcoming product development, we need to set a baseline of how RideCheck performs as an "on-trip safety policy" platform today.

02 Proposed Solution

RideCheck Holdout Group

Holdout Description

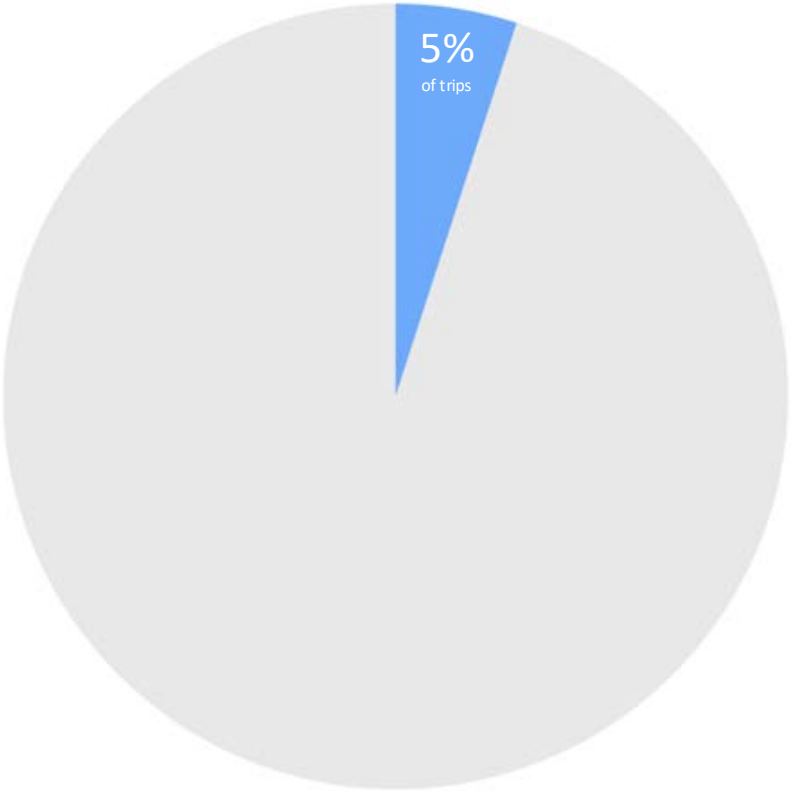
Methodology

At request time, a trip is assigned treatment or control. If it is in the control group, it will receive no RideChecks included in the scope of the Holdout Group. Treatment trips receive all eligible RideChecks as normal. On the backend, we keep track of all anomaly triggers as normal.

→ Per trip = All users benefit from RideCheck most of the time

In Scope

All Active Safety checks for Mobility trips. (See what's out of scope here.) Exclusions on exception basis only, such as regulatory requirements — and always accompanied by a DS analysis on exclusion impact to the Holdout.



Out of scope

✗ Delivery Workflows

Courier deliveries look very different from mobility trips, which would confound our analysis. They are also covered by a single RideCheck as of 2024 H2.

✗ Crash Detection

Road Safety's Crash Detection RideCheck is optimized for its own set of goals. It already has a more complex heuristic and higher-level interventions compared to other Mobility RideChecks.

Known risks

Scale / Impact

An effective Holdout Group will exclude somewhere between 5–10% of trips from receiving any RideCheck intervention.

This is a significant volume of trips that will have one less Safety product covering them, which is compounded by our ambition to make this product more important to the overall on-trip Safety journey. There is a chance that an incident will occur that we did not have interventions for.

Inconsistent Experience

Because our design is per-trip, users (esp. Drivers) may realize their trip-to-trip experience is inconsistent, and attribute that to faulty or random behavior, eroding some of their trust in our Safety product portfolio.

Discussion

What metric are we measuring here?

Safety Incident Rate ← Our main candidate

Pros

A no-brainer

As "the" top-line Safety KPI, it requires no explanation. It is everywhere, allowing for comparison between RideCheck and other Safety features, as well as more systematic analyses.

Alignment

This metric corresponds with our intent to extend the RideCheck portfolio into higher-impact anomalies and interventions.

Cons

Confusion

A new RideCheck intervention may at once both decrease real-life incidents and also increase the rate of reporting. Because this metric encompasses the two effects, it becomes hard to interpret.

Whether this number goes up, down or stays the same is neither good nor bad if we can't separate those two effects.

Proposed solution

Safety Incident Rate

Holdout Setup

A 5% holdout group (1 in 20 trips) is sufficient to measure effects as small as 3% in Safety Incident Rate in 1.5 months, and effects as small as 1% within a year.

	5% holdout	10%	15%	20%	25%
1% effect	52 wks	26	17	13	10
2%	13	6	4	3	3
3%	6	3	2	1	1
4%	3	2	1	1	1
5%	2	1	1	1	0
6%	1	1	<1	<1	<1
7%	1	1	<1	<1	<1
8%	1	<1	<1	<1	<1
9%	1	<1	<1	<1	<1
10%	<1	<1	<1	<1	<1

Uber | RideCheck Holdout Group Check Point | Source

L3/L4 Incident Rate ← An alternative

Pros

De-Escalation

Sophisticated heuristics with progressive interventions should help us de-escalate and prevent Serious Incidents.

While the overall Safety Incident Rate may go up by educating users on Safety tools, Serious Incidents should hopefully always trend down.

Cons

Exceptionality

Uber is already a exceptionally safe platform. L3/L4 incidents are rare enough that measuring effects in this rate is a finicky process subject to long lead times and susceptible to random variations.

Proposed solution

L3/L4 Incident Rate

Holdout Setup

A 10% holdout group could measure medium effects (5%) on L3/L4 Incident Rate within a year, and large effects (10%+) within 3 months.

	5% holdout	10%	15%	20%	25%
1% effect	2,316 wks	1,158	772	579	463
2%	576	288	192	144	115
3%	255	127	85	64	51
4%	143	71	48	36	29
5%	91	45	30	23	18
6%	63	31	21	16	13
7%	46	23	15	11	9
8%	35	17	12	9	7
9%	27	14	9	7	5
10%	22	11	7	6	4

Discussion: Should we have a metric?

Diff-in-Diff Enabler

Having historic holdout data allows us to monitor the evolution of RideCheck's performance at an aggregate level (all heuristics and all interventions considered). This can help us steer efforts towards incremental gains, and give insight into how to balance user annoyance with being there when we're needed.

Research and Brand Tracking

If we opt for a per-user holdout design, we can use the historical data to support Brand Tracking and Research on RideCheck's evolving impact on Safety Sentiment, Safety Feature Awareness and other user behavior.

Thank You | Questions?