

A White Box Framework for Estimating Long-Term Experimental Impact via Short-Term Metrics

Wen Qiang Toh
wth@roblox.com

Roblox
San Mateo, California, USA

Yidi Wang
yidiwang@roblox.com
Roblox

San Mateo, California, USA

Wenjing Zheng
wzheng@roblox.com

Roblox
San Mateo, California, USA

Joao Dimas
jdimas@roblox.com
Roblox

San Mateo, California, USA

ABSTRACT

Experiment decisions often require not only long-term treatment effects, but also insights into the underlying causal mechanisms. We present a white box framework that estimates long-term treatment effects by linking short-term key metrics to a long-term north star and its driving mechanisms. We also introduce validation methods using pseudo-experiment data and actual experiment data. We further outline the estimation and deployment considerations on an experimentation platform.

KEYWORDS

A/B testing, Experimentation, Surrogates, Long-term treatment effect, Metric hierarchy

ACM Reference Format:

Wen Qiang Toh, Wenjing Zheng, Yidi Wang, and Joao Dimas. 2025. A White Box Framework for Estimating Long-Term Experimental Impact via Short-Term Metrics. In *Proceedings of ACM KDD 4th End to End Customer Journey Optimization Workshop*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

1.1 Context

Companies would often like to optimize toward a long-term north star metric, but experiments are rarely run long enough to observe it directly. Instead, they measure short-term key metrics that target different components of the north star. Therefore, we need methods to predict how experimental interventions will affect the long-term north star, which the literature on long-term treatment effects has been focusing on.

However, beyond predicting the overall impact, stakeholders would also like to understand **how** their interventions impact the north

star metric. Our work provides a framework targeting this objective by linking short-term key metrics teams often monitor in their experiments to the long-term north star and its driving mechanisms.

1.2 Relationship with other approaches to predict long-term experimental effects

One approach to measuring long-term experimental effects leverages existing customer long-term value models. The average difference in predicted long-term value between users in the treatment and control variants is taken as the long-term effect of the treatment intervention. One key downside of this method is that it lacks insights on the joint causal relationships of short-term key drivers on the long-term north star and its driving mechanisms.

Our work tightly links with the surrogate index literature [1, 2, 3, 4], where surrogate metrics are used to predict the long-term treatment effect of the experiment. These surrogates lie on the causal path from the experimental intervention to the long-term outcome. A key assumption of the method is that conditional on the value of the surrogate index, treatment allocation and the long-term outcome are independent. The key difference between this literature and our work is that the surrogate index often uses shorter-term equivalents of the long-term north star to predict long-term treatment effects. However, when the goal is to illuminate how experimental interventions affect the long-term north star, we should **exclude** shorter-term versions of the long-term north star in the surrogate index as while these may have the highest predictive power, they are themselves results of the pathways we wish to understand. Instead, we have to provide a causal structure of how short-term metrics drive the long-term north star.

2 FRAMEWORK

2.1 Setup

To illustrate this framework, let the user-level long-term north star metric ($NorthStar_{LT}$) be 1 year engagement hours, where $NorthStar_{LT} = ReturnFreq_{LT} \times EngagementperReturn_{LT}$, i.e. the driving mechanisms of interest are return frequency (number of visit days in 1 year period) and engagement intensity (average engagement hours per return in 1 year period)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM KDD 4th End to End Customer Journey Optimization Workshop, August 2025, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Suppose experiments measure a set of short-term key metrics $\{KeyMetric_{j,ST} : j \in J\}$, where some of these could be short-term values of the driving mechanisms (eg. short-term return frequency, and short-term engagement intensity). We measure the relationships among relative lifts of the short-term key metrics and the relative lifts of the long-term driver mechanisms.

$$\Delta ReturnFreq_{LT} = \sum_j \gamma_j \Delta KeyMetric_{j,ST}$$

$$\Delta EngagementperReturn_{LT} = \sum_j \xi_j \Delta KeyMetric_{j,ST},$$

where Δ refers to the **relative change** in the metric between treatment and control variants in an experiment (i.e. $\Delta X = \frac{X_{treat}}{X_{ctrl}} - 1$).

Next, we link short-term metrics to the north star metric. For small relative changes,

$$\Delta NorthStar_{LT} \approx \Delta ReturnFreq_{LT} + \Delta EngagementperReturn_{LT}.$$

Therefore, we have

$$\begin{aligned} \Delta NorthStar_{LT} &= \sum_j \beta_j \Delta KeyMetric_{j,ST} \\ &= \sum_j (\gamma_j + \xi_j) \Delta KeyMetric_{j,ST} \end{aligned} \quad (1)$$

, where $\beta_j = \gamma_j + \xi_j$.

With this, we can predict the impact of a treatment intervention on the north star with insights into how it does so.

2.2 Interpretation

We use hypothetical numbers to illustrate this framework can help inform short-term key metric tradeoffs and experiment decisions. Suppose we learn the following estimated relationships from our data:

$$\Delta ReturnFreq_{LT} = 0.5 \Delta KeyMetric_{1,ST} + 3 \Delta KeyMetric_{2,ST}$$

$$\Delta EngagementperReturn_{LT} = 1.5 \Delta KeyMetric_{1,ST} + 1 \Delta KeyMetric_{2,ST}$$

Then, from (1) we have

$$\Delta NorthStar_{LT} = 2.0 \Delta KeyMetric_{1,ST} + 4.0 \Delta KeyMetric_{2,ST}$$

It provides stakeholders with a rule of thumb in experiment decisions:

- For 1% decrease in $KeyMetric_{2,ST}$, we need at least 2% increase in $KeyMetric_{1,ST}$ to net positive on long-term north star.
- $KeyMetric_{2,ST}$ is our most efficient lever to move the north star, as it has sizable impact on both of its drivers. By contrast, $KeyMetric_{1,ST}$ predominantly impacts through engagement intensity.

3 ESTIMATION OF TRADEOFF COEFFICIENTS

3.1 Implementation

We estimate tradeoff coefficients using user-level observational data (denoted by i) as we do not have sufficient experiments that last the duration of the long-term metrics to reliably estimate its relationships with short-term metrics. We use log-log regressions

to approximate the percentage change relationships. We use the Double Machine Learning (DML) partially linear model, where X_i is a list of covariates that includes past user behaviour and demographics. LightGBM models were used for the estimation of $l(X_i)$, $g(X_i)$ and $m(X_i)$. Residuals of the metrics were computed and a regression on these residuals retrieved the coefficients of interest.

$$\begin{aligned} \log(ReturnFreq_{i,LT}) &= \sum_j \gamma_j \log(KeyMetric_{i,j,ST}) + l(X_i) + u_i \\ \log(EngagementperReturn_{i,LT}) &= \sum_j \xi_j \log(KeyMetric_{i,j,ST}) + g(X_i) + \epsilon_i \\ \log(KeyMetric_{i,j,ST}) &= m_j(X_i) + e_i \end{aligned}$$

3.2 Evaluation of model predictions

Out-of-fold predictions showed an improvement of 31-55% of RMSE across metrics over a baseline RMSE with the mean metric as the prediction.

3.3 Limitations

Estimation of log models face an issue with zero-valued outcomes. Adding a small constant is a common way to resolve this issue, with the tradeoff that estimated trade-off coefficients are scale-variant to the value of the constant [5].

4 APPLICATION TO EXPERIMENTS

4.1 Inference of predicted long-term experimental effect

For each experiment, we get the relative lifts of short-term metrics. The variance of the relative lifts in the short-term metrics ($\hat{Var}(\Delta KeyMetric_{j,ST})$) can be computed using the Delta method.

We get the point estimate of the predicted lift in the long-term north star metric by combining the lifts of the short-term metrics. However, getting the variance of the combined metric in production is challenging (see section 7). For simplicity, we assume the β 's to be constant and the covariance terms to be zero, although the latter is likely to matter for the quality of the inference due to metric co-movements.

$$\begin{aligned} \Delta NorthStar_{LT} &= \sum_j \hat{\beta}_j \Delta KeyMetric_{j,ST} \\ \hat{Var}(\Delta NorthStar_{LT}) &\approx \sum_j \hat{Var}(\beta_j \Delta KeyMetric_{j,ST}) \end{aligned}$$

4.2 Determination of metric horizons

Appropriate time horizons must be chosen for both the long-term north star metric goals and short-term key metrics. For the long-term north star, optimizing towards a shorter horizon may yield different outcomes than a longer one depending on the business. However, targeting longer horizons involves tradeoffs in operational feasibility and model freshness. For short-term metrics, shorter horizons enable quicker experiments.

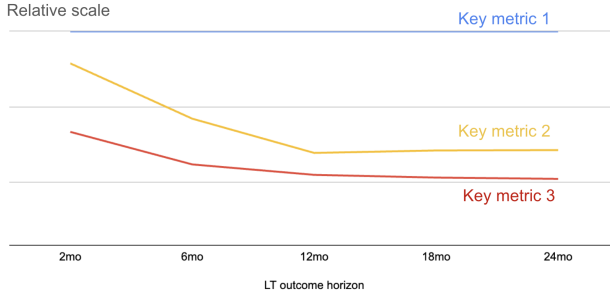


Figure 1: Relative tradeoffs across long-term north star horizons

Note: Relative scale has been modified for publication

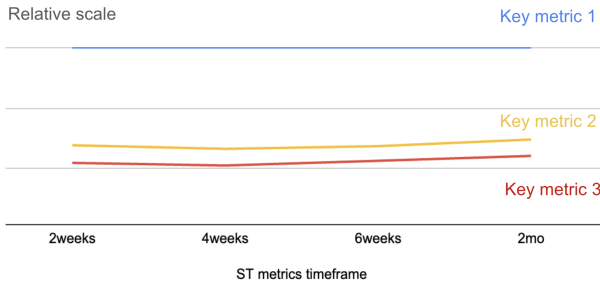


Figure 2: Relative tradeoff across short-term metric horizons

Note: Relative scale has been modified for publication

We estimated tradeoff coefficients for varying long-term north star and short-term key metric horizons. In our case, relative tradeoffs of short-term key metrics with the long-term north star stabilized around a 12 months horizon (Fig. 1), while similar relative tradeoffs were observed across various short-term metric horizons (Fig. 2).

5 VALIDATION WITH PSEUDO-EXPERIMENT DATA

5.1 Methodology

We evaluated whether the combined long-term north star metric behaves as a valid experimental metric. We select a small sample of experiments ($N=18$), selecting one treatment variant and one control variant for each experiment.

False positive rate

We perform 1000 placebo tests for each experiment by re-randomizing treatment allocation. We apply the methodology to get the test decision of the combined metric. As these are placebo tests, with a 5% significance level, a false positive rate should be 5%. We use a chi-square test to test whether the observed number of nonsignificant and significant test results differs from the expected (950, 50).

Uniform distribution of p-values

The distribution of p-values from the placebo tests should be uniformly distributed. An experiment is deemed to have passed the p-value distribution test if using a Kolmogorov–Smirnov test, we cannot reject the null that the distribution of p-values come from a uniform distribution.

Statistical power

Using the original sample sizes of the experiments, we calculate the implied statistical power for an assumed minimum detectable effect (MDE), e.g. a 1% lift in the long-term north star metric.

Type S error

A statistically significant result in an underpowered experiment may turn out to be a false positive. However, the chance of a sign reversal could be low. We quantify given an imputed true effect, how likely we observe a statistically significant result of the opposite direction given the uncertainty of the experiment (Type S error) [6]. In this study, we take the size of the imputed true effect to be the MDE used for computing statistical power and assume symmetry, but the choice of the imputed true effect does not have to be symmetric.

5.2 Results

The results are in Table 1. First, the performance across experiments is heterogeneous. The quality across evaluation criteria is correlated — a severely underpowered experiment is also likely to have issues with false positive rate, non-uniform p-values and high Type S error. Second, all experiments are significantly underpowered for this north star metric. However, for some experiments, the chances of making a Type S error are reasonably low. We further found that the power of the combined metric is caused by a constituent metric that most experiments often do not power for, suggesting that power improvements can come from improving the power of this metric.

6 VALIDATION WITH EXPERIMENTAL DATA

6.1 Methodology

To validate our predicted long-term north star, we ideally would use real experiments as ground truth. Our key interest is in investigating whether **test conclusions** of the actual realized long-term north star from experiments and the predicted long-term north star are aligned, i.e. whether the treatment effect is statistically significant positive, not statistically significant or statistically significant negative. However, we lack experiments that span the full long-term horizon of interest. As a result, we assume that validation results from shorter-term horizons generalize to the long-term.

We collected 48 experiments lasting at least one month, yielding 81 experiment-variant combinations for one-month evaluation. For each combination, we computed the actual north star lift. We then estimated tradeoff coefficients for each horizon and used them to predict the north star lift. For both the actual and predicted north star lifts, we conducted two-sided tests at the 95% confidence level. We report accuracy, false positive rate (Type I error), Type II error, precision and recall with exact 95% binomial confidence intervals.

We also explored extending the analysis to longer horizons, but only 8 combinations lasted 1.5 months, limiting interpretability.

6.2 Results

For a one-month north star horizon, we observed an accuracy of 86% (95% CI: [77%, 93%]). The false positive rate was 11% (95% CI: [5%, 20%]), and the Type II error rate was 4% (95% CI: [1%, 12%]). Estimates of precision and recall of each class are in Table 2. If the launch decision were to be contingent on a positive (negative) long-term effect, precision and recall of the class *stats sig positive* (*negative*) will be the most relevant. However, these statistics have wide confidence intervals, indicating the need to include more experiments for a larger sample to achieve more precise estimates. However, improving precision for these statistics will remain difficult due to the limited number of experiments with statistically significant long-term results.

7 CHALLENGES

Tension between understanding mechanisms and prediction accuracy

When understanding mechanisms becomes the primary focus of an approach, there are likely to be tradeoffs with prediction accuracy. Short-term north star outcome realizations are likely to be strong predictors of the long-term north star, which is why these are often incorporated into surrogate indices [3, 7]. They are also likely mediators of other short-term drivers not included in our framework above. However, in order to be useful to stakeholders and easily communicated, the selection of metrics must be parsimonious.

Surrogacy Assumption

The surrogacy assumption will be violated if the short-term key metrics measurement window does not capture the stable trend of short-term treatment effect, for example, due to novelty effects or delayed treatment effects. The set of key metrics may also not fully capture all the treatment’s mediating pathways towards the north star.

Productization

Accurate inference requires computing covariances of relative lifts of metrics. If metrics are processed in parallel, this may require substantial platform refactoring.

8 CONCLUSION AND FUTURE DIRECTIONS

Our key contribution is to build a transparent metric decomposition framework that links short-term metrics to a long-term north star and its driving mechanisms. This provides visibility into the impact on causal mechanisms of the north star, helping experimenters make short-term experiment tradeoffs for long-term impact. In addition, we formulate an evaluation framework that uses synthetic data generated from experiments to evaluate the performance of the predicted north star arising from this framework as an experimental metric and also illustrated the use of actual experimental data from shorter horizons for evaluation.

Future directions include deepening the causal structure of driver metrics to relate product domain-level driver metrics to the short-term key metrics mentioned in this work. The main challenge here is a systematic way to select a sufficiently parsimonious set of driver metrics and evaluate the mediating relationship among these driver metrics to assign them to the correct layer of the metric hierarchy. We will expand the experimental sample size for our validation to gain more precision in our validation estimates. In addition, we will improve our treatment effect estimation method to mitigate the impact of surrogacy violations and achieve better validation performance.

REFERENCES

- [1] C. J. Weir and R. J. Walley. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in medicine*, 25(2):183–203, 2006.
- [2] M. M. Joffe and T. Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538, 2009.
- [3] S. Athey, R. Chetty, G. W. Imbens, and H. Kang. The surrogate index: combining short-term proxies to estimate long-term treatment effects more rapidly and precisely, 2019.
- [4] J. Yang, D. Eckles, P. Dhillon, and S. Aral. Targeting for long-term outcomes. *Management Science*, 70(6):3841–3855, 2024.
- [5] J. Chen and J. Roth. Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, 139(2):891–936, 2024.
- [6] A. Gelman and J. Carlin. Beyond power calculations: assessing type s (sign) and type m (magnitude) errors. *Perspectives on psychological science*, 9(6):641–651, 2014.
- [7] S. Huang, C. Wang, Y. Yuan, J. Zhao, and J. Zhang. Estimating effects of long-term treatments. *arXiv preprint arXiv:2308.08152*, 2023.

Table 1: Validation tests of long-term north star metric as an experimental metric

Experiment	FPR (%)	Passed FPR test	Passed p-value test	Power (%)	Type S error (%)
1	21.6%	No	No	5%	48.8%
2	7.2%	No	No	5%	21.0%
3	4.7%	Yes	Yes	6%	11.1%
4	6.2%	Yes	Yes	7%	4.6%
5	5.3%	Yes	Yes	7%	5.4%
6	4.9%	Yes	Yes	10%	1.1%
7	4.8%	Yes	Yes	9%	1.7%
8	3.8%	Yes	Yes	10%	1.1%
9	4.7%	Yes	Yes	9%	2.5%
10	4.4%	Yes	Yes	14%	0.3%
11	4.9%	Yes	Yes	10%	1.1%
12	4.4%	Yes	Yes	16%	0.2%
13	5.9%	Yes	No	13%	0.4%
14	5.4%	Yes	Yes	19%	0.1%
15	5.8%	Yes	No	21%	0.0%
16	5.5%	Yes	Yes	16%	0.2%
17	4.1%	Yes	Yes	28%	0.0%
18	4.5%	Yes	Yes	33%	0.0%
Average	6.0%	89%	78%	13%	5.5%

Table 2: Precision and recall for 1 month long-term north star horizon

	Precision	Recall
Class		
Stats sig positive	50%	80%
95% CI	[16%, 84%]	[28%, 99%]
Non-stats sig	96%	89%
95% CI	[88%, 99%]	[80%, 95%]
Stats sig negative	20%	33%
95% CI	[1%, 72%]	[1%, 91%]