

From the research bench

What is CUPED?

CUPED adjusts an A/B test metric with the same metric measured before the test. It cuts noise, not bias, and needs a covariate the test cannot move.

Author	Published	Reading time	Words
Sinan Isoglu, MBA (Quantic) Commercial growth leader, lecturer and doctoral researcher	18 October 2026	9 min	1,885

Management summary

CUPED (Controlled-experiment Using Pre-Experiment Data) is a way to analyse a randomized experiment that adjusts each user's outcome with data from before the experiment, so that the same effect can be detected with fewer users or in less time. This concept guide defines it (Deng et al.), explains why variance decides how long a test must run (Kohavi et al.), shows how much variance the adjustment can remove as a function of one correlation, reports what it achieved at Bing, sets out the conditions under which it biases a result, and gives a one-page record for a CUPED analysis. The sources are one methods paper with results from one search engine, a survey and practical guide, and an experience report; the last two come from the same Microsoft experimentation group.

Contents	§ 1	What is CUPED?	2
	§ 2	Why does variance decide how long a test must run?	2
	§ 3	How does CUPED reduce variance?	3
	Figure 1	Users needed for the same power, by correlation with the pre-experiment metric	4
	§ 4	What did CUPED achieve at Bing?	4
	§ 5	When does CUPED bias a result?	5
	§ 6	What goes into a CUPED analysis record?	6
	Figure 2	A CUPED analysis record	6
	§ 7	Where has CUPED been tested?	7
		References	8

What is CUPED?

§ 1 of 7

CUPED is an adjustment made when an A/B test is analysed, not when it is designed. Deng et al. (2013)¹ introduced it: “we propose a technique, called CUPED (Controlled-experiment Using Pre-Experiment Data), which adjusts metrics using pre-experiment data for users in the control and treatment to reduce metric variability.” The experiment runs as usual; the analysis then corrects each group’s average with what the same users did before the experiment started.

¹ Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 123–132.

The adjustment keeps the estimate centred where it was and makes it less noisy. In the authors’ words, the adjusted estimate “is still an unbiased estimator for the shift in the means” and “has a smaller variance” than the unadjusted difference, so “the corresponding t-statistic would be larger for the same expected effect size”.

In this guide, CUPED means estimating a randomized experiment’s effect after adjusting each user’s outcome by a covariate measured before assignment, usually the same metric in a pre-experiment period, with one adjustment coefficient for both groups. It changes how precise the estimate is, not what is being estimated, and it does not repair a broken randomization.

Why does variance decide how long a test must run?

§ 2 of 7

Because the noise in a metric sets how many users a test needs: the standard error of an average shrinks only with the square root of the number of users. Kohavi et al. (2009)³ give the rule of thumb for 95 per cent confidence and 80 per cent power: the users needed in each variant grow with the variance of the metric and shrink with the square of the change one wants to detect. Their example is a shop where 5 per cent of visitors buy: detecting a 5 per cent change in revenue per user takes “over 409,000 users” in each variant, while the same change in conversion, a less variable metric, takes “less than 122,000 users”.

³ Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140–181.

They list three ways to raise an experiment's sensitivity: more users, a metric with "inherently lower variability", or filtering out users who could not have seen the change. Deng et al.¹ add why this matters in practice: "Small differences in key metrics, on the order of fractions of a percent, may have very significant business implications", and "detecting a 0.5% delta requires 100 times (10 squared) more users" than detecting a 5 per cent one. The [A/B test guide](#) explains how a test is sized; CUPED is a fourth way to the same end.

How does CUPED reduce variance?

§ 3 of 7

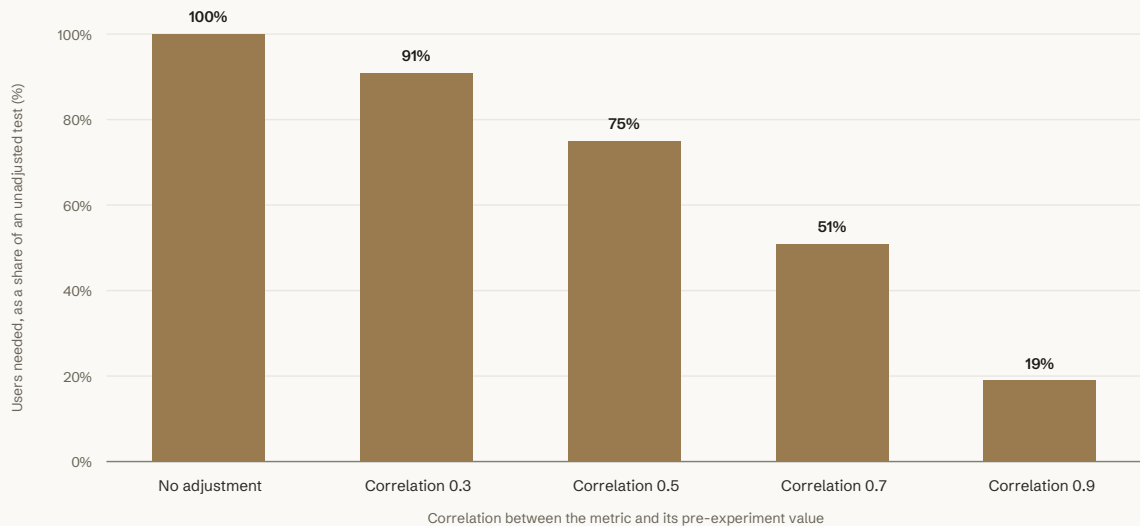
By borrowing a technique from simulation called control variates. Deng et al.¹ show that if a covariate measured before the experiment is subtracted from the outcome, scaled by the right coefficient, the variance of the estimated difference is multiplied by one minus the square of the correlation between the two. They state that "The larger rho, the better the variance reduction", where rho is that correlation, and that the best coefficient "turns out to be the ordinary least square (OLS) solution of regressing (centered) Y on (centered) X", with Y the metric and X the covariate. That makes CUPED close to regression adjustment, with one difference the authors stress: their argument needs no model of the outcome, whereas a linear model "makes strong assumptions that are usually not satisfied in practice".

Because the covariate is measured before the experiment, it does not bias the estimate. "By using only information from before the launch of the experiment to construct the control variate, the randomization between treatment and control ensures" that both groups have the same expected covariate, so its correction cancels in expectation. The choice of covariate follows from the correlation, not from the timing: to make the correlation large, "an obvious approach is to choose X to be the same as Y, which naturally leads to using the same variable during pre-experiment observation window as the control variate". Unbiasedness has one further condition: "the same theta has to be used for both control and treatment", where theta is the coefficient. They also show that stratification, grouping users by a categorical pre-experiment variable, is a special case of the same idea.

Figure 1 turns the variance result into users. It is this guide's arithmetic, not data from the sources: if the variance of the estimate is multiplied by one minus the squared correlation, and the users needed for given power rise in proportion to the variance, a correlation of 0.7 roughly halves the users needed (51 per cent).

Figure 1 Users needed for the same power, by correlation with the pre-experiment metric

With a correlation of 0.7 between a metric and its pre-experiment value, a test needs about half the users. With 0.3 it saves less than a tenth.



Arithmetic from the authors' formula, assuming every user has a pre-experiment value. Missing values lower the gain.

Source Author's computation from the variance result in Deng et al. (2013), pp. 126 and 127, and the sample-size rule in Kohavi et al. (2009), p. 152: users needed equal one minus the squared correlation, as a percentage of an unadjusted test. Illustrative arithmetic, not observed data.

What did CUPED achieve at Bing?

§ 4 of 7

About half the variance, for metrics on which users differ a lot and stay much the same from one week to the next. Deng et al.¹ report that “we can reduce variance by about 50%, effectively achieving the same statistical power with only half of the users, or half the duration”. In the live system, “Three important recent experiments showed variance reductions of 45%, 52% and 49% with one week of experiment and one week of pre-experiment data.” In one experiment that slowed the page down on purpose, the adjusted result was significant “from day 1”, where the unadjusted test barely crossed the threshold after two weeks.

The gain depends on the covariate and the metric. In a three-week A/A test on queries per user, the same metric from the week before the test

reduced variance by “more than 45%”, while the day a user first entered the experiment, which is not pre-experiment data but cannot be moved by the treatment, reduced it by about 9 to 10 per cent. Revenue was the exception: “One notable exception is revenue-per-user, where CUPED reduced the variance by less than 5% due to the low correlation of revenue-per-user between the pre-experiment and the experiment periods.” The authors’ recommendation is that “Using a pre-experiment period of 1-2 weeks works well for variance reduction. Too short a period will lead to poor matching, whereas too long a period will reduce correlation with the outcome metric during the experiment period.” The length of the test matters separately: the longer it runs, the more of its users are new and have no pre-experiment record, which lowers the gain.

When does CUPED bias a result?

§ 5 of 7

When the covariate can be moved by the treatment. The rule is stated as a warning: “Never use covariates that could be affected by the treatment, as this could bias the results.” Deng et al.¹ show what happens otherwise. In one Bing experiment that raised queries per user, they adjusted with an in-experiment metric that tracked queries closely, and the estimate came out negative: “a result that is directionally opposite of the known effect”.

Two further limits come from the data. Users without a pre-experiment record, such as new users, have no covariate; the authors add an indicator for whether a user appeared in the pre-period, which splits users into those with a record and those without. The split itself removes some variance, but within the group without a record there is nothing to adjust with, so the lower the coverage (the share of users with a record), the smaller the gain. And metrics computed per page view rather than per user need the delta method for their variance, combined with the adjustment. Kohavi et al. (2013)², closing a review of experimentation at Bing that leaves sensitivity aside, note that “In Bing, we started using pre-experiment data to reduce the variance and improve sensitivity”, and “we believe there is room for significant improvements in this area”. The statistical conclusion validity guide sets out what a more powerful test does and does not establish.

² Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1168–1176.

What goes into a CUPED analysis record?

§ 6 of 7

Six fields, in one record per experiment: the metric, the covariate and why the treatment cannot move it, the pre-experiment window, the share of users who have a pre-experiment value, the unit of analysis, and the checks that are reported with the adjusted result. The fields follow Deng et al.¹'s conditions and recommendations and Kohavi et al.³²'s rules for sizing a test.

Figure 2 A CUPED analysis record

Write down the metric, the covariate, the window, the coverage, the unit and the checks before reading an adjusted result.

Metric	Covariate	Window	Coverage	Unit	Checks
Which metric, and does it vary between users?	Measured before assignment? Could the test move it?	How many weeks before the test start?	What share of users have a prior value?	Per user, account, page or order?	Unadjusted result shown? A/A test passed?

If the treatment could move the covariate, report the unadjusted result only.

One row per experiment. An adjusted result without the unadjusted one beside it cannot be checked.

Source Author's decision worksheet. The covariate and window fields follow Deng et al. (2013), pp. 127 to 130; the metric field follows Deng et al. (2013), p. 130; the coverage and unit fields follow Deng et al. (2013), pp. 127 to 129; the checks field follows Kohavi et al. (2009), pp. 151 and 152, and Deng et al. (2013), p. 128. The fields are reader inputs; no company data is supplied.

An illustration, with no company data: a B2B software firm tests a redesigned report builder with the users of its existing customer accounts, randomizing whole accounts so that colleagues who share reports see the same version. The record names reports created per user as the metric, because heavy and light users differ widely (the case in which Deng et al.¹ find variance reduction works best), and the same metric over the two weeks before the test as the covariate,

which the redesign cannot have touched. It notes that users added to accounts during the test have no pre-period value and carry an indicator instead, analyses per user with the delta method for the variance, as Deng et al. do when the unit analysed is smaller than the unit randomized, and reports the unadjusted and adjusted differences side by side after an A/A test on the same split. Before the firm uses CUPED on revenue per account, it checks how closely each account's revenue before the test tracks its revenue during the test, because CUPED's small gain on revenue at Bing came from a low correlation between the two periods, not from revenue as such. These are the author's examples and have not been tested.

Where has CUPED been tested?

§ 7 of 7

At one search engine, by its inventors. Deng et al. (2013)¹ give the derivation and report results from Bing's experimentation system; Kohavi et al. (2009)³ is a survey and practical guide and Kohavi et al. (2013)² an experience report, both from the same Microsoft experimentation group. The t-test that the paper adjusts also leans on scale: for online experiments, the authors write, the samples of both groups "are at least in thousands, hence the normality assumption on Y is usually unnecessary because of the Central Limit Theorem." On this guide's reading, a B2B test with a few hundred accounts is a different setting, where the adjustment may still reduce variance but the usual test may not hold; the sources do not study it.

 End of essay

References

- 1 Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 123-132. <https://doi.org/10.1145/2433396.2433413>
- 2 Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1168-1176. <https://doi.org/10.1145/2487575.2488217>
- 3 Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140-181. <https://doi.org/10.1007/s10618-008-0114-1>

How to cite this essay

Isoglu, S. (2026, October 18). What is CUPED? Sinan Isoglu.
<https://isoglu.com/insights/journal/what-is-cuped/>
Online version: <https://isoglu.com/insights/journal/what-is-cuped/>

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher
© 2026 Sinan Isoglu · isoglu.com · All rights reserved