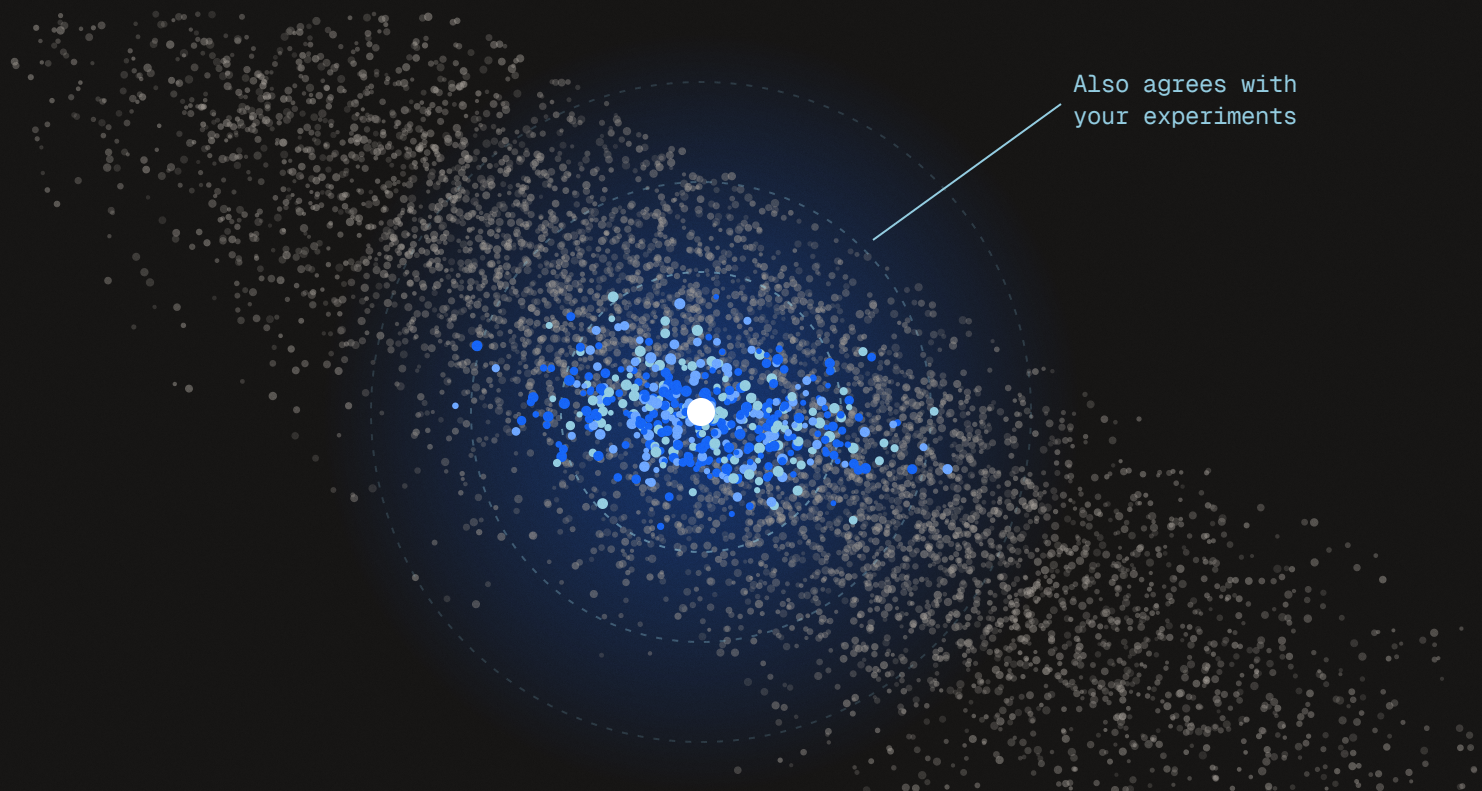


Each dot: a model that matches your sales



Report

Your MMM uses experiments. But does it actually learn from them?

As experiment calibration becomes standard, what matters is whether those experiments actually anchor what the model learns.



Ittai Shacham, PhD
Staff Applied Scientist, Haus

Adding experiments to marketing mix models (MMM) is quickly becoming table stakes.

That is a good thing. The work marketers put into incrementality testing should carry real weight in the model informing their budget decisions.

But accepting experiments as inputs does not guarantee the model's conclusions will agree with them. A model can “incorporate” your test results, fit historical sales beautifully, and still get the channel response curves wrong.

So when someone tells me their MMM is calibrated, I want to know: Do its final estimates of channel impact actually line up with the experiments?

That is the distinction between a model that is experiment-calibrated and one that is truly experiment-anchored.

01

Predictive vs. causal fit

A model can match total sales and still misread channel impact

During a promotion, a marketer might increase spend across search, social, and video while underlying demand rises at the same time. The model's job is to separate each channel's contribution from the demand the promotion would have generated anyway.

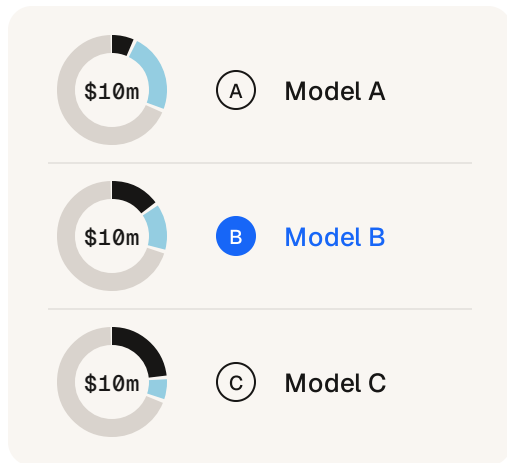
The problem is that several explanations can fit the same history. \$1 million from search and \$3 million from social could be read the same as \$3 million from search and \$1 million from social. A model can predict the sum perfectly and still misread which channel contributed what.

Experiments help distinguish those possible explanations by measuring the effect of deliberately changing spend. They provide evidence about channel impact that the sales total alone cannot supply.

That is why agreement with experimental evidence is a first-order requirement for our models. Predictive fit asks whether the model can reproduce sales. Causal fit asks whether its estimates of channel impact agree with experimental evidence. You need both.

Different mixes predict the same sales total

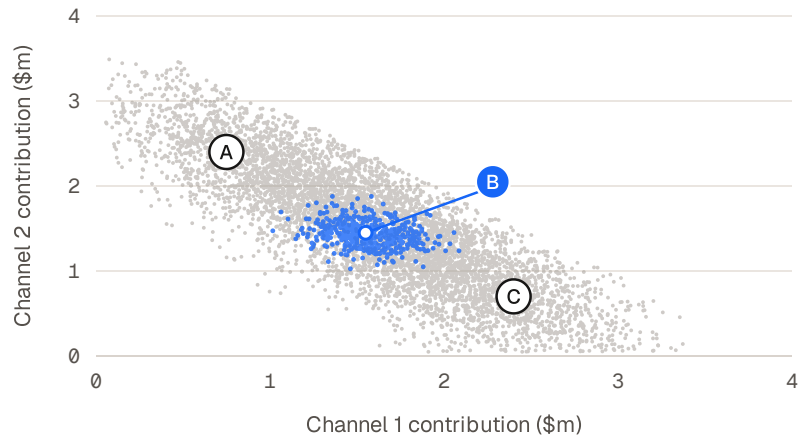
Channel 1
 Channel 2
 Other



Other = other media + organic

Experiments narrow the plausible explanations

Matches sales total
 Also agrees with experiments



A, B and C correspond to the three models at left.

Each dot represents a possible model of the same sales total, shown for two channels in a schematic illustration.

02

Posterior matching

Experiments should anchor the channel response curves

Saying that an MMM “uses experiments” doesn’t tell you *how much* they shape the model’s final result.

A common approach is to use experimental evidence to inform a prior, then update it with the historical data. But an experiment-informed prior does not guarantee an experiment-consistent posterior: If the sales history exerts enough influence, the fitted estimates can drift away from the evidence. The experiment may influence the model without adequately anchoring the result.

At Haus, our posterior matching algorithm checks whether the fitted response curves agree with your experiments. When they conflict, it updates the model and refits so that agreement with the experiments is an explicit objective of fitting.

Crucially, experiments are considered at the spend level they were tested. That's because a test at \$20,000 in spend and another around \$60,000 provide information that can help reveal the response curve: how incremental sales change with spend and where diminishing returns begin. Some approaches blend results into one average ROI, which loses that useful context.

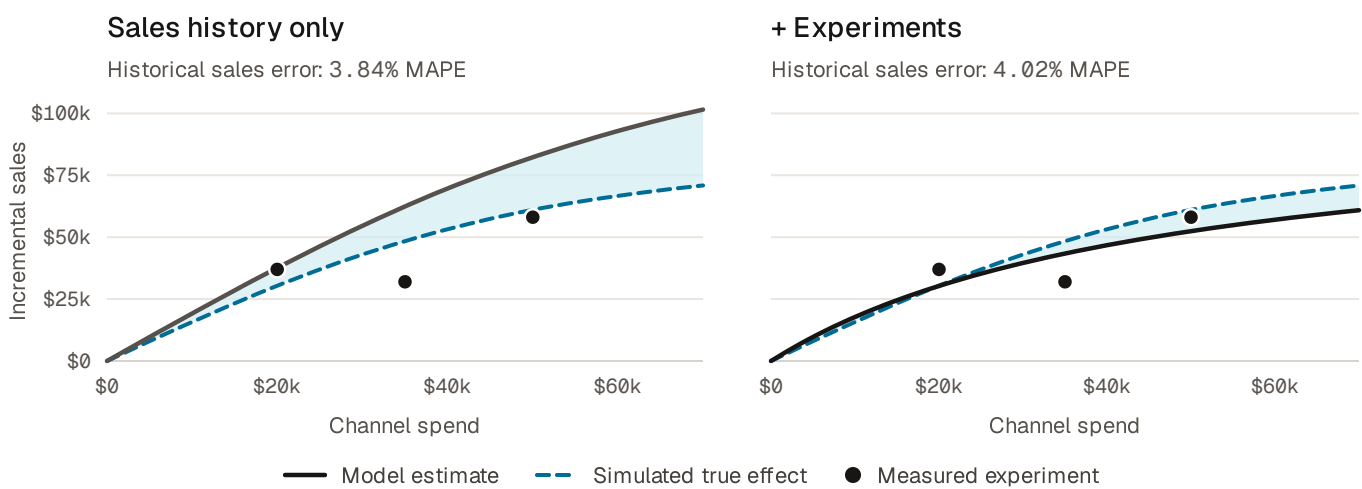
Uncertainty is also factored in. A model shouldn't force its response curve through every experiment's point estimate, and more precise experiments should carry more weight.

The model's final explanation of your business should hold up against your experiments.

Your brand's experiments are our strongest causal anchors, and priors from Haus' Incrementality Index, our experiment library, support channels you haven't tested. Observational data still has an important role in explaining the rest of the business. It just never overrules our strongest causal evidence.

To make this mechanism concrete, we use a simulated business where we know each channel's true impact. That gives us a way to check what the model learns against what advertising actually caused.

Adding experimental constraints changes the learned response curves even though historical sales fit barely moves. The constrained model gets closer to the true channel effect around the tested spend levels, an improvement predictive fit alone would miss.



Illustrative response curves for one channel, with dots showing experimental estimates and shading showing the gap from the simulated true effect.

But even a model that remains anchored to the experiments can interpret them too simplistically.

03

Time-varying effectiveness

An experiment's timing is critical context

Suppose you test the same channel in a quiet month and again during a major promotion. The second test finds a higher incremental return. Which result should you include in the model?

Choosing whichever test seems most relevant to your next planning period feels sensible, but it discards potentially useful evidence. If you wait for a test that perfectly matches the spend level, season, and market conditions you're forecasting, you'll rarely have one when you need it. Useful experiments start to look "stale" simply because the context has changed.

Controlling for organic seasonality does not fully solve this problem. More customers may buy without advertising during a promotion, masking the fact that the advertising itself was also more effective. You might end up under-crediting media during productive periods and over-crediting it during weaker ones.

Haus' time-varying return curves estimate how media effectiveness changes through time. Before posterior matching fits the response curve, it uses those estimates to put each experiment on a common timing basis. This helps separate the effect of spending more from the effect of advertising at a more productive time.

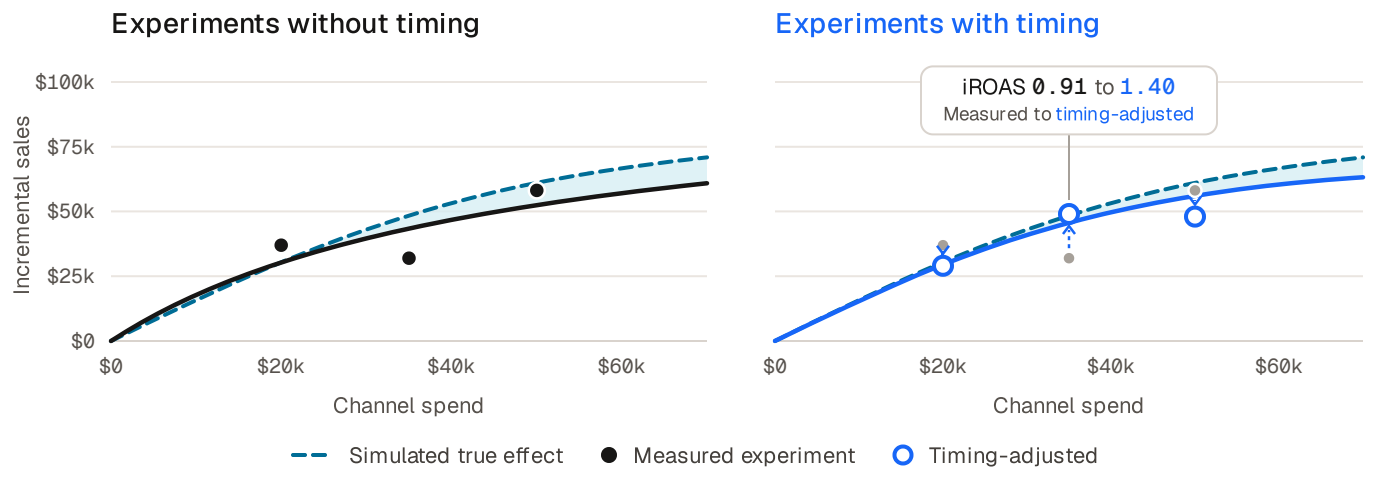
Account for when each test ran so experiments from different periods can inform the same response curve.

Take a \$35,000 test from the simulation. It measured \$32,000 in incremental sales, or 0.91 iROAS. The model estimates that media effectiveness during the test was 65% of its level in the reference period. Dividing by that factor puts the test on a common timing basis: about \$49,000 in incremental sales, or 1.40 iROAS.

The experiment measured 0.91; the 1.40 figure is *not* a revised result. It represents what the test implies on the common basis used to compare experiments from different periods.

A quiet-period test can then inform peak-season planning without assuming effectiveness is constant. A peak-season test can inform ordinary periods without treating favorable conditions as the year-round norm.

Note: Changes in tactic, targeting, or product can still limit how comparable older experiments are.



Tests shift vertically at the same spend, with iROAS measured as incremental sales divided by spend. Shading shows the gap from the simulated true effect.

04 Validation

Test the model against experiments it hasn't seen

A model can agree with experiments it was explicitly fitted to and still fail to generalize. We therefore leave each experiment out in turn, fit the model using the remaining evidence, and compare its predicted lift for the test period with the experiment's measured lift. This tests whether the model learned relationships that extend beyond its calibration points.

Withholding a few weeks of sales is also a useful predictive check. But predicting another period's total does not establish whether the channel contributions are right. Withholding an experiment tests whether the model can anticipate the effect of deliberately changing spend on a particular channel.

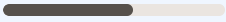
We ran that evaluation in the same simulated business, with three channels, nine experiments, and 520 days of sales history.

The evaluation surfaces a known tradeoff of anchoring a model to experiments: a slightly looser fit to historical sales in exchange for a more credible estimate of channel impact. In this simulation, adding experiments increases historical sales error from 3.84% to 4.02% MAPE, while reducing withheld-experiment error from 2.63 to 1.46. For a model that directs budget, we would accept that modest change in historical fit to better recover channel incrementality.

Accounting for experiment timing improves both measures in this run. Withheld-experiment error falls to 0.91, a further 38% reduction, while historical sales error falls to 2.93% MAPE. Predictive fit still matters, but a slightly better fit to the total should not outweigh a more credible explanation of what each channel caused.

A small tradeoff in sales fit, until timing is added

Lower error is better; the highlighted model performs best in this simulation

Model	Historical sales error		Withheld-experiment error	
Sales history only	3.84%		2.63	
+ Experiments	4.02%		1.46	
+ Experiments + timing Best	2.93%		0.91	

Adding experiments increases historical MAPE by 0.18 percentage points while improving prediction of withheld experimental results. Accounting for experiment timing then improves both measures in this simulation.

Study details

In this simulation, MAPE summarizes percentage error in the model's fit to historical sales used for training, while causal error compares predicted lift with results from experiments withheld during fitting. The response curves are explanatory illustrations, and this study does not directly compare posterior matching with conventional experiment-informed priors.

05

Budget decisions

Better channel estimates uncover more incremental sales

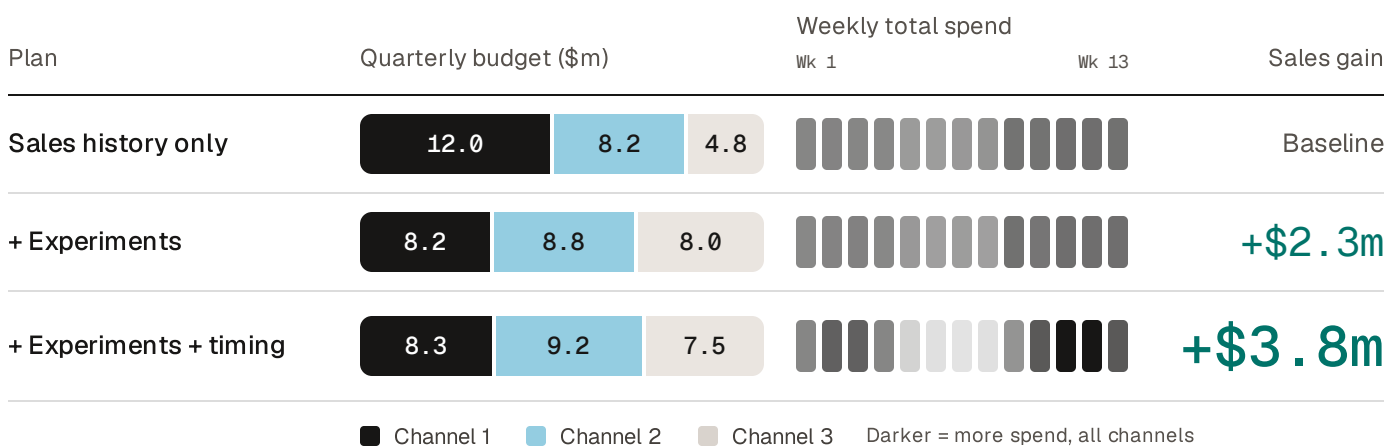
Response curves ultimately feed into budget decisions. Better causal estimates don't guarantee a better allocation. But if a curve is wrong near the spend levels you're weighing, it misstates what the next dollar will return and the eventual budget inherits that error.

While better response curves inform how much each channel should receive, time-varying effectiveness also informs *when* that budget is likely to work hardest.

In a separate constructed illustration, three channels share a \$25 million quarterly budget under the same constraints. Compared with the plan based on sales history alone, the experiment-informed plan generates about \$2.3 million more incremental sales in the quarter, a 3% gain. Accounting for experiment timing raises that gain to about \$3.8 million, or 5%.

More incremental sales from the same \$25m quarterly budget

Additional incremental sales versus the sales-history-only plan



This constructed planning illustration is separate from the model-fit study; each plan has the same \$25m quarterly budget and constraints. Gains are measured against the sales-history-only plan, with values rounded.

Make your experiments count

Marketers accept real costs to build incrementality programs. They hold back spend, work around operational constraints, and trade short-term results for the cumulative gains of learning. Don't let your MMM ignore that work.

Before trusting an MMM to direct the next dollar, don't stop at asking whether it's calibrated. Ask whether it's anchored: Do its channel effects stay consistent with your experiments, and do they hold up when an experiment is withheld?

Key terms

Predictive fit

How closely the model reproduces observed sales. Strong predictive fit shows the model can explain the total, not that it has credited each channel correctly.

Response curve

The estimated relationship between a channel's spend and the incremental sales it drives, including where diminishing returns begin. Budget decisions are made along this curve.

Posterior matching

Haus' iterative fitting process. It checks the fitted response curves against your experiments at the spend levels you tested and refits when they conflict, so experiments shape the final estimates as well as the starting assumptions.

MAPE

Mean absolute percentage error: the average percentage gap between modeled and observed sales, ignoring direction. It measures predictive fit.

Causal fit

How well the model's estimated channel effects agree with experimental evidence, given that evidence's uncertainty.

iROAS

Incremental return on ad spend: incremental sales generated per dollar of advertising spend. It counts only sales that wouldn't have happened without the ads.

Time-varying return curves

Estimates of how a channel's effectiveness changes through time. They let experiments from different seasons, promotions, or demand conditions be compared on a common timing basis.

Withheld-experiment error

How far the model's predicted lift lands from an experiment's measured lift when that experiment is left out of fitting, averaged across experiments. It measures causal fit on evidence the model hasn't seen.



About the author

Ittai Shacham, PhD

Staff Applied Scientist, Haus

Ittai earned his PhD in econometrics at Tilburg University, then worked as a senior researcher at Meta, where he built internal causal modeling tools. At Haus, he works on the science behind Causal MMM.