



PRACTITIONER GUIDE
2026

STELLA GROWTH INTELLIGENCE

Nine Questions Your Measurement Vendor Can't Answer.

How to tell if your measurement actually works, before you move a dollar of budget on it.

A STELLA PUBLICATION
STELLAHEYSTELLA.COM

*For CMOs, measurement leads, and finance
partners at enterprise brands.*

What's inside.

•	Executive summary	3
01	The Number You Can't Check	4
02	In-Sample Is the Answer Key	6
03	The Receipt Has Three Layers	8
04	Out-of-Sample MAPE	13
05	Out-of-Sample R-Squared	15
06	The Backtest and the Forward Test	17
07	Collinearity and VIF	20
08	Residual Diagnostics	22
09	Test Design	24
10	Power and Minimum Detectable Effect	26
11	Confidence Bounds	30
12	The Written Path From Model to Budget	32
13	The Tear-Out NEW	34
14	Why Stella Shows All Nine	36
15	What to Do	39

Executive summary.

Enterprise measurement engagements routinely produce results that do not survive technical due diligence. The standard pattern: in-sample model fit presented as proof of accuracy, confidence intervals omitted, MDE undisclosed, residual diagnostics unaddressed, and recommendations issued without explicit uncertainty bounds. Buyers receive reports that look authoritative and cannot be audited.

This guide provides nine validity checks that distinguish measurement work from measurement theater. The checks are organized in three layers: prediction (whether the model works on data it did not train on), identification (whether the model can separate signal from noise across channels), and decision quality (whether the result is strong enough to support budget action).

Each check is presented with the specific question to ask, the strong and weak answers to expect, and the failure mode the check is designed to surface. The guide is written for measurement, finance, and growth leaders evaluating internal or vendor measurement output.

The Number You Can't Check.

Your measurement function exists for one reason. The platforms grade their own homework. Meta reports Meta's contribution. Google reports Google's. Every channel reports efficient in isolation while blended performance moves independently of any of them. At enterprise scale, the gap between platform-reported revenue and actual incremental contribution is no longer a measurement nuisance. It is a nine-figure budget allocation question. So the function exists to settle it. And in most organizations, it does not.

And they handed you a number you still cannot check. Same problem. Nicer dashboard.

Here is the part nobody says out loud. A measurement vendor can show you a clean lift number, a confident budget recommendation, and a model that looks airtight, without ever showing you whether that model survives data it has never seen. You traded one number you could not audit for another.

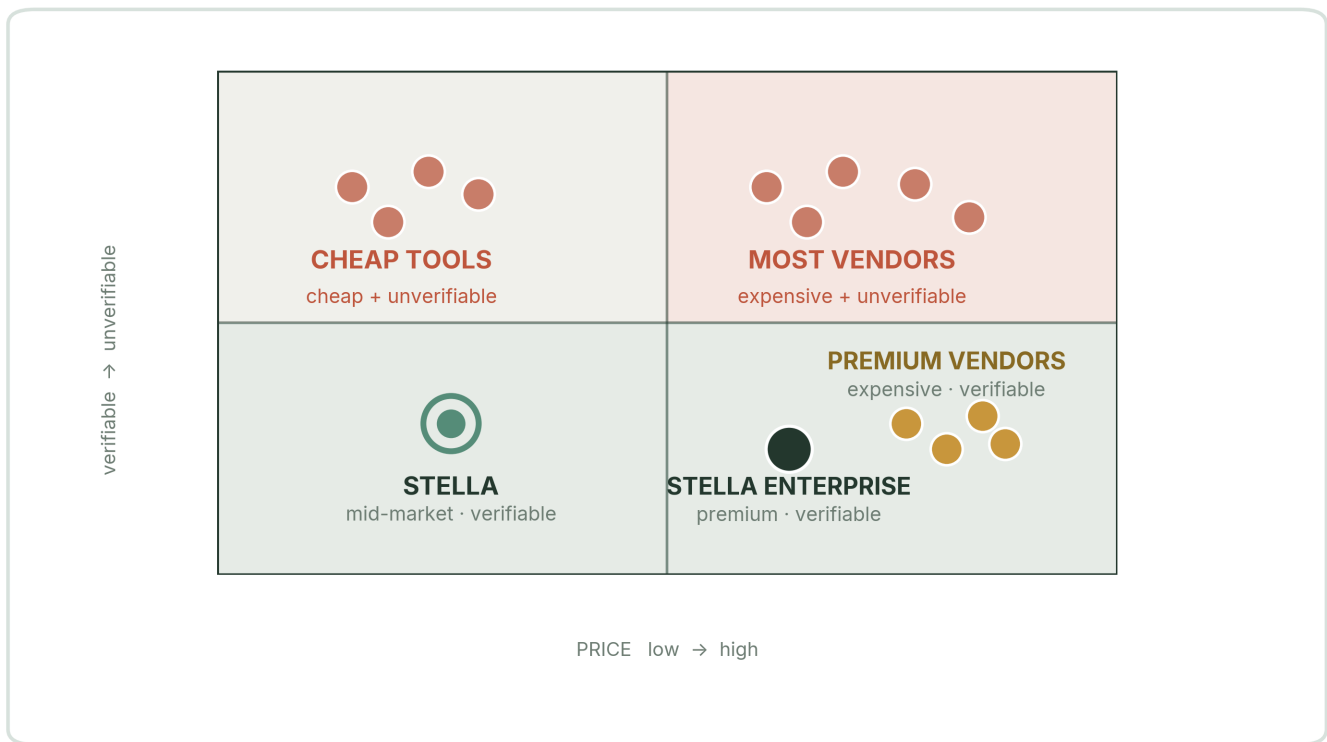
A number you can't check is an opinion with a decimal point.

This playbook is not about which vendor promises the best results. They all promise the best results. It is about the standard of proof they can put on the table when you ask. There are nine checks. Most vendors cannot or will not show all nine, and the ones they skip are the exact ones your budget decisions go wrong on.

One disclosure before we start. Stella runs measurement for a living, so we have a stake in this. We are handing you the questions that are hardest for us to answer too. Every check in here, we show on every model we run, and at the end we tell you exactly how to point all nine back at us.

Now the reframe the rest of this depends on. This is not cheap versus expensive. A weak model at five figures a month is more expensive than a strong one, because the damage is not the fee. The damage is the budget you move on a number that was wrong. It is verifiable versus unverifiable. That is the only comparison that matters, and it is the one your current vendor would rather you did not make.

FIGURE 1 Cheap vs expensive is the wrong axis.



Price is the axis everyone argues on. It is the wrong one. The choice that decides whether your budget moves are sound is verifiable against unverifiable.

In-Sample Is the Answer Key.

Before the nine checks, one idea. Get this and the rest is easy. A model learns by studying a stretch of your history. Spend, revenue, seasonality, promos. It finds the patterns that explain what happened.

In-sample fit asks: how well does the model explain the data it was allowed to study? That is the answer key. The model already saw it. Of course it looks good.

Any MMM can hit 95% in-sample R-squared. That doesn't mean it works. It means it memorized the past.

Out-of-sample fit asks a different question. Hide a stretch of data during training. Then check whether the model can predict that stretch it never saw. That is the real test. It is the one that tells you whether the model will hold up next quarter, where your budget actually goes.

Here is how to make it concrete. Picture a student who gets the exam with the answers already filled in, scores 98, and tells you they are ready for the real thing. You would not trust that. You would want to see them sit a test they have not seen. A model is the same. In-sample is the filled-in answer key. Out-of-sample is the real exam.

And to be fair to the model, some in-sample fit is necessary. A model that cannot explain the past it studied is broken before it starts. The problem is not that vendors report in-sample fit. It is that they stop there, lead with it, and hope you do not ask for the out-of-sample number behind it. The nine checks that follow are mostly about one thing: making them show you the hard one.

FIGURE 2 The model that aced the test it studied for.



The model looks perfect right up until the moment it has to predict something it did not study. The lines sit on top of each other through training, then the prediction drifts.

The Receipt Has Three Layers.

Not every check carries the same weight. There are three layers, in the order they should worry you when one is missing.

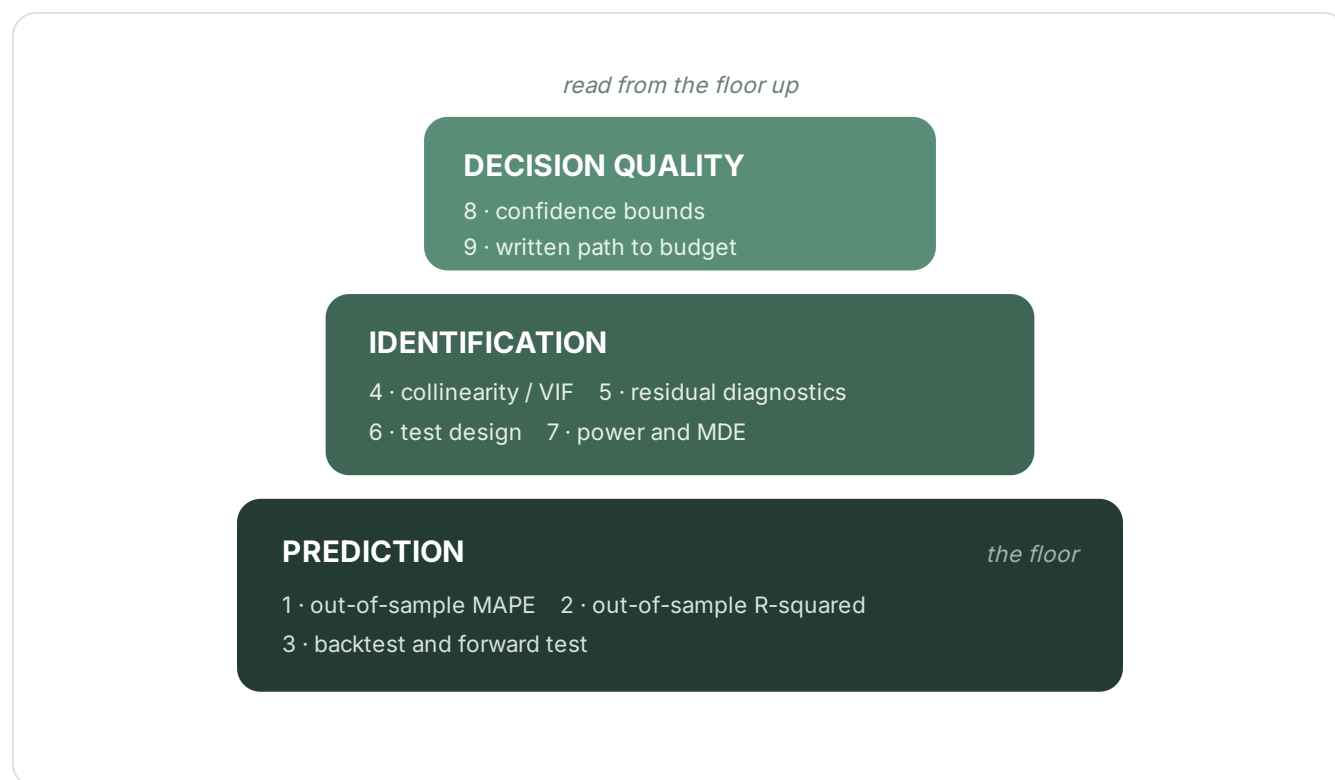
Prediction. Can the model predict data it did not train on? If it has only been graded on the period it learned from, that is a memory test, not evidence. This is the floor. Fail here and nothing above it matters, because the number is fiction.

Identification. Can the model tell signal from noise? Can it separate which channel actually did the work, or are the channels too tangled to read? A model can predict total revenue well and still credit the wrong channel. Fail here and the total looks right while every reallocation you make on it is wrong. This is the layer that quietly wastes budget.

Decision quality. Is the result strong enough to move money, or just interesting enough to put in a deck? Fail here and you paid for analysis that never became a decision.

Nine checks. Three layers. Here is how they map.

FIGURE 3 Three layers of the receipt. Read from the floor up.



A vendor can ace the top and fail the bottom. If the model can't predict data it never saw, every layer above it is built on fiction.

At enterprise scale, two more layers sit underneath these three.

Data integrity. Before any model can predict anything, the input signal has to be trustworthy. Multi-brand portfolios, multi-region reporting, multi-currency revenue, retailer sell-through versus DTC sell-out, marketplace data, channel attribution conflicts between agency and platform reporting: enterprise measurement starts with an integration problem that a \$750K DTC brand does not have. The model is downstream of this. If the input data has lineage gaps or definition conflicts, no amount of model validation upstream fixes the result.

Governance. At enterprise scale, measurement output has to survive legal review, finance audit, board-level inspection, and in some cases regulatory disclosure. A defensible deliverable in the SMB context is one a smart CMO trusts. A defensible deliverable in the enterprise context is one that holds up when a Big Four auditor reviews the methodology and a CFO presents the conclusions to the board. These are different bars.

The nine checks in the rest of this guide apply at both scales. The data integrity and governance layers apply primarily at the enterprise scale. Smaller brands can usually compress these into a single workstream. Multinationals cannot.

Two quick definitions so the labels on each check make sense. MMM models all your channels at once from historical data. Incrementality tests one channel with a controlled experiment. iROAS is the lift number both produce. iROAS is a revenue multiple, not a profit one, and that distinction decides budget moves later, so hold onto it.

How to read this, and which checks apply to you. First, the thresholds in here are Stella's working bands, not laws of nature. They come from running this hundreds of times. We say so each time one appears. The exact cutoff matters less than whether your vendor can tell you where their number landed and why. A vendor who cannot name the number at all is the real red flag. Second, not every check applies to every method. Five of the nine are MMM checks. Two are incrementality checks. Two apply to both. If your vendor only runs one method, you only need that column plus the two that apply to everyone.

FIGURE 3B Which checks apply to your vendor.

#	CHECK	APPLIES TO
1	Out-of-sample MAPE	MMM
2	Out-of-sample R-squared	MMM
3	Backtest and forward test	MMM
4	Collinearity / VIF	MMM
5	Residual diagnostics	MMM
6	Test design	INCREMENTALITY
7	Power and MDE	INCREMENTALITY
8	Confidence bounds	BOTH
9	Written path to budget	BOTH

MMM vendor? Use 1 to 5, plus 8 and 9. Incrementality only? Use 6 and 7, plus 8 and 9. Doing both, like we do? All nine.

PART ONE

Prediction.

*Can the model predict what it hasn't
seen?*

Out-of-Sample MAPE.

What it is. MAPE is the average size of the model's miss, written as a percentage. Predict \$100K, actual was \$88K, that is a 12% miss. MAPE averages those misses across the window the model never trained on.

Why it matters. It answers the most practical question you can ask a model. When this thing makes a prediction, how far off is it, usually? In a percentage anyone in the room can read.

The number to ask for is the out-of-sample MAPE, on a held-back window. Low in-sample MAPE means almost nothing, because the model already studied that data. Industry working bands, drawn from public methodology references and from our own benchmark across 225 tests: 10 to 15% is strong, over 20% is a problem.

You might be thinking those cutoffs are arbitrary. Fair. They are working bands from running this hundreds of times, not physics. The exact number matters less than the question. A vendor who can tell you where their out-of-sample MAPE landed, and why, is doing the work. A vendor who cannot tell you at all is the actual red flag. Treat every band in this playbook the same way: a starting point, and a reason to make them defend their own.

One honest caveat, because we would rather say it than have a sharp reader catch it. MAPE has quirks. It punishes over and under-forecasting unevenly, and it gets unstable when the actual numbers run near zero. The right error metric depends on your KPI. But as a first screen, out-of-sample MAPE tells you fast whether the model is predicting or guessing.

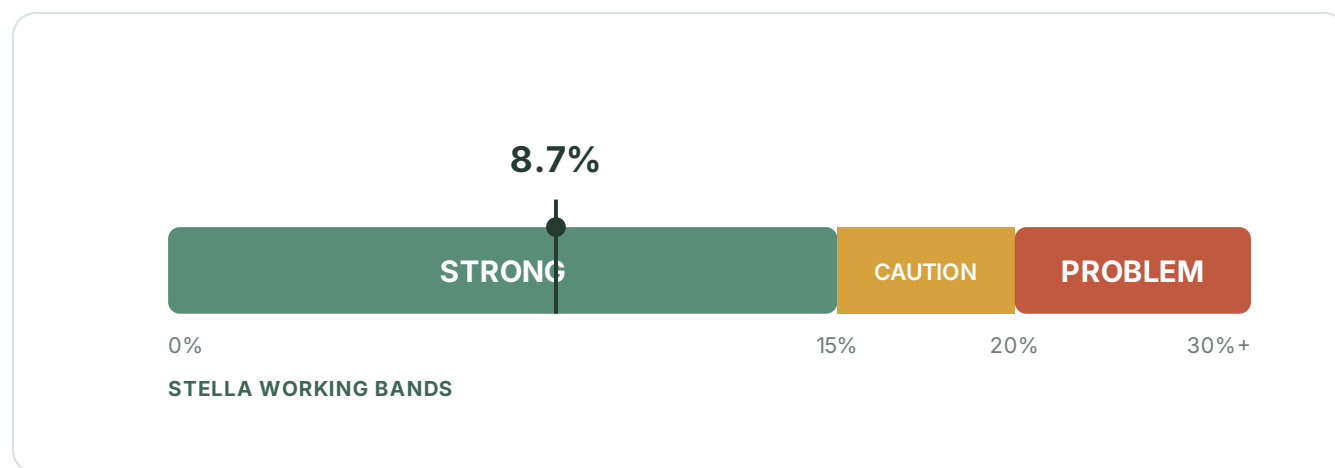
At enterprise scale, MAPE has to be computed by segment, not aggregated. A model with 8% blended MAPE can have 22% MAPE on the EMEA region and 4% on North America. The blended number hides the EMEA problem until budget reallocation triggers it. Serious enterprise reporting publishes MAPE by region, by brand, by channel, not as a single rolled-up number.

THE ASK What was your out-of-sample MAPE, and which window did you hold back?

Strong answer *"MAPE was 8.7% on a four-week holdout, and we beat a naive seasonal baseline."*

Weak answer *"The model fit was strong."*

FIGURE 4 What a MAPE number is telling you.



Ask where your model's out-of-sample MAPE lands, not its in-sample fit.

Out-of-Sample R-Squared.

What it is. R-squared tells you how much of the rise and fall in your sales the model can actually account for, versus just guessing the average every time. Out-of-sample means it was scored on data it never saw.

Why it matters. This is the cleanest single read on whether the model learned something real or memorized the past.

Here is the move most vendors make. They show you a gorgeous in-sample R-squared. 0.95, 0.97, 0.98. It looks like brilliance. It is usually overfitting. The model memorized the wiggles in the training data, including the random ones, and it will break the moment it meets new data.

The tell is not the number itself. It is the gap. A model with 0.97 in-sample and 0.62 out-of-sample did not get smarter, it got memorized. A model with 0.91 in-sample and 0.89 out-of-sample is being honest with you.

A suspiciously perfect fit isn't a strong model. It's a model that memorized the noise and will break on contact with reality.

So a high out-of-sample number is good. But treat an unusually high one as a question, not a trophy, and check it against holdout behavior. We do this on our own models.

THE ASK What was the out-of-sample R-squared, and how did you check for overfitting?

Strong answer "Out-of-sample R-squared was 0.89. We treat unusually high values as possible overfit and check them against holdout behavior."

Weak answer "The R-squared was 0.98."

FIGURE 5 The gap is the tell.



Don't ask for the R-squared. Ask for both, and look at the distance between them.

The Backtest and the Forward Test.

This is the pair that separates someone who built a model from someone who operates one. A backtest and a forward test both grade the model on data it never trained on. They just do it at different points in time.

The forward test. Rebuild the model with the last stretch of real history withheld. Run it over that window using the spend you actually ran. Then compare the prediction to the revenue you actually earned. That gap is the model's honesty on a stretch it never studied. Predicted \$2M and you did \$2.1M, that is a strong model. Predicted \$2M and you did \$3.4M, the model is broken, and every budget call you made on it was a bet you did not know you were placing.

The backtest. Deploy what the optimizer recommends. Then check, monthly, whether revenue is tracking what the model predicted. If yes, the loop is closing. If no, you catch a drifting model before the quarter closes, not after.

Scoring a model on the same history it was fit on is neither of these. That is the answer key from Chapter Two, wearing a more impressive name.

*Running one of these makes you look data-driven.
Operating both changes the business.*

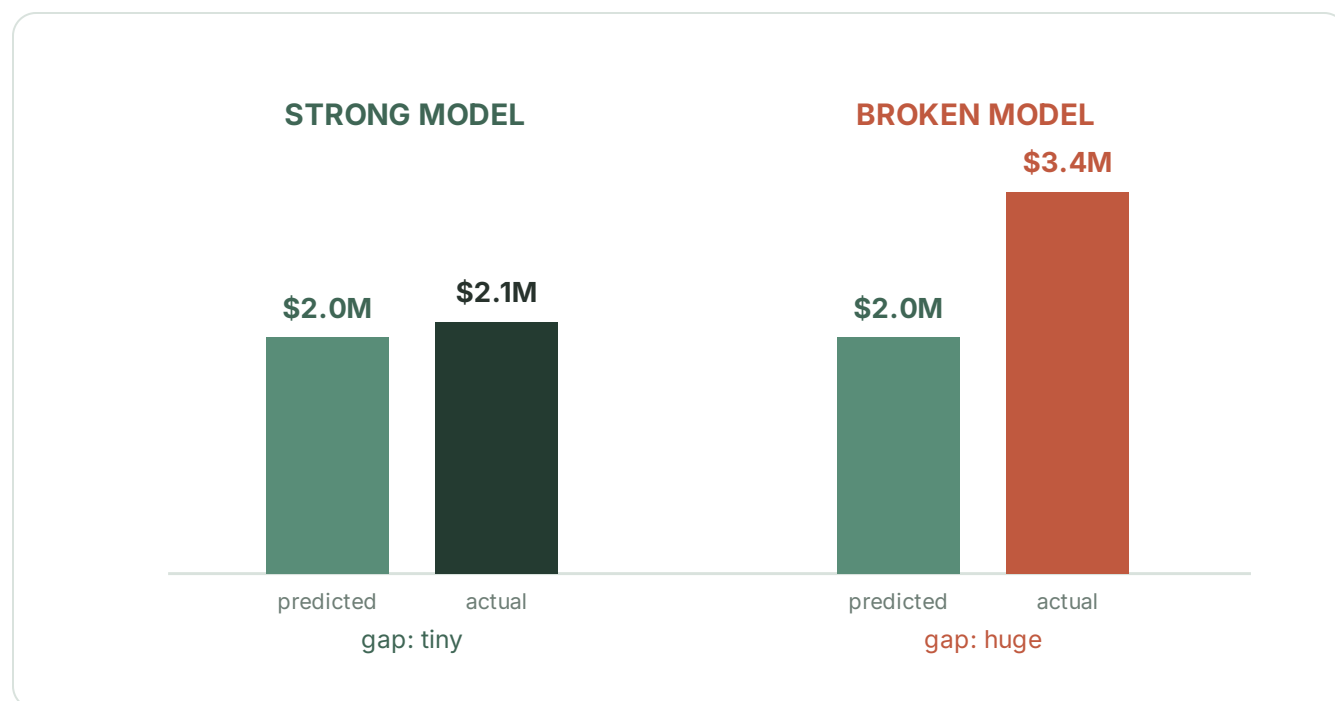
A vendor who has done neither has been running a model, not operating one. They can describe the past. They have never once checked whether the model can call the future.

THE ASK Was the model forward-tested on a held-out window, and is a backtest running against live revenue, rather than scored on the history it was trained on?

Strong answer *"We withheld the last 12 weeks, predicted them, and the gap was small. The live recommendation is being checked against actuals monthly."*

Weak answer *"We validated on the full history."*

FIGURE 6 What honesty looks like. The forward test gap.



The forward test turns a confident model into a measured one. The gap between predicted and actual on a held-out window is the part the deck never shows you.

PART TWO

Identification.

Can it tell what actually caused what?

Collinearity and VIF.

What it is. Your channels do not move on their own. Meta, Google, email, promos, and seasonality tend to ramp together during a launch or a sale. VIF, the variance inflation factor, measures how tangled they are.

Why it matters. When everything moves at once, the model can still predict total revenue fine. But the per-channel numbers, the ones you actually reallocate budget on, get shaky. The model cannot cleanly tell which channel did the work, because they all moved together.

Stella's working threshold: above 5, the model cannot separate those channels cleanly, and any precise per-channel split it gives you is precision it has not earned. A high VIF does not mean a channel is worthless. It means there is not enough independent movement in the data to estimate that channel's effect with confidence. The honest move is to say so, not to publish a clean-looking number for two channels that rose and fell together all year.

Enterprise multi-brand portfolios compound the collinearity problem. Cross-brand promotions, shared seasonality, and coordinated platform launches push VIF higher than single-brand models typically see. A multinational running synchronized Q4 campaigns across six brands routinely has channel-level VIF above 8 on at least half the channels, and any reported per-brand-per-channel ROI inside that range is approximation dressed as estimate.

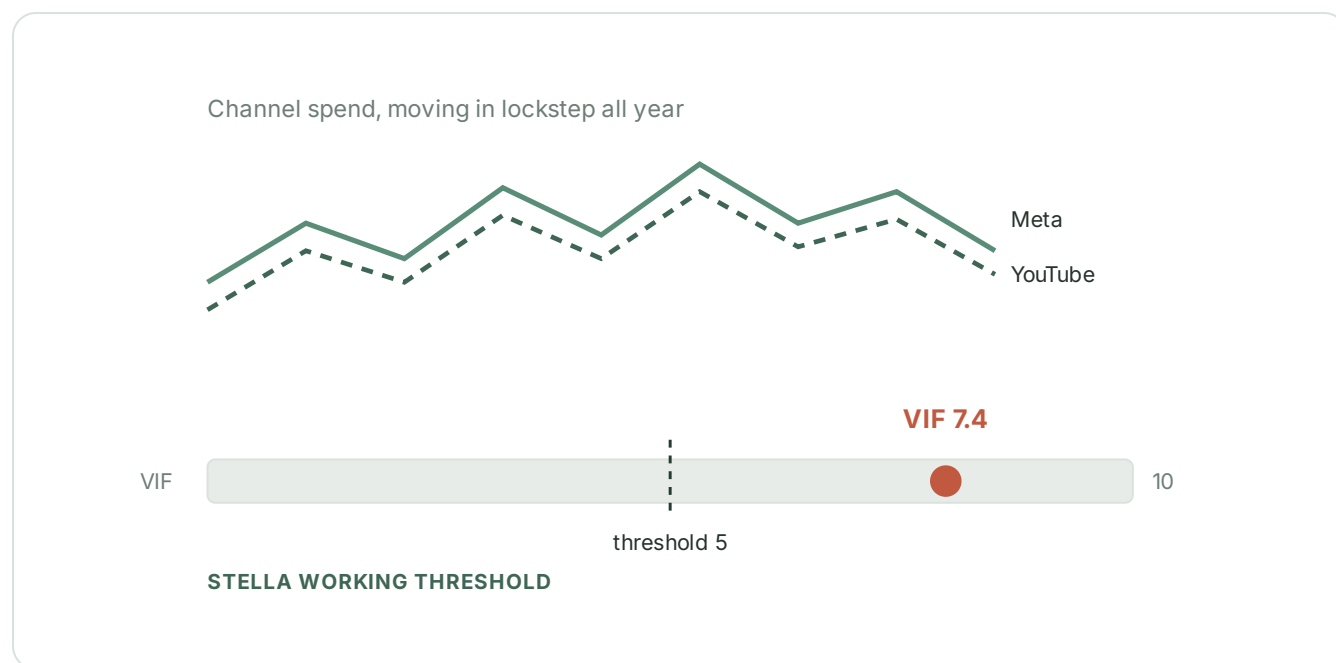
Separate, precise ROI for two channels that obviously move together is precision the model hasn't earned.

THE ASK Can you show me the VIF for our channels, and what did you change because of it?

Strong answer *"Meta and YouTube were highly correlated, so we didn't report separate ROI as if they were cleanly identified."*

Weak answer *"The model accounts for that."*

FIGURE 7 When the model is guessing which channel did the work.



When two channels move together, the model can't split the credit. VIF tells you when to stop trusting the split.

Residual Diagnostics.

What it is. A residual is one miss. What the model predicted minus what actually happened, for every single data point. MAPE and R-squared summarize the misses into one number. Residuals show you the shape of them.

Why it matters. A model can post strong fit metrics and still be missing something real. The residuals are where it shows. In a healthy model, the misses look like random scatter. No direction, no clustering. When the misses cluster, the model is missing structure.

For a marketing model, that usually means something specific. It underpredicts during promos. It overpredicts during holidays. It never adjusted for a price change or an inventory crunch. The fit metrics can look fine while the model is quietly wrong in exactly the windows where you make your biggest decisions.

“The R-squared was high” is not an answer about residuals. It’s a way of not answering.

A defensible vendor can tell you a problem they caught in the residuals and fixed. That is what checking actually sounds like.

THE ASK Did you check the residuals for patterns, and what did you find?

Strong answer *“Residuals clustered around promos, so we added a promo term and re-fit.”*

Weak answer *“The R-squared was high.”*

FIGURE 8 Random misses vs misses with a shape.



Every model misses. The question is whether the misses have a shape.

Test Design.

What it is. For an incrementality test, the iROAS you get out is shaped by how the test was built, not just by what happened during it. So the design has to be visible.

Why it matters. You can run a “holdout” and still get garbage. The number will look authoritative either way. What separates a clean read from a contaminated one happens before the test starts, in how the markets were chosen and how the control was built.

The most common failure is picking a test market because it “feels” right, or hitting randomize on a list of regions. That is not a design. That is vibes with a confidence interval attached.

A proper test answers three questions before a single ad turns off. Does the test region historically move with the rest of the business? Can a lift even be detected here, or is it too noisy? If you ran a placebo test with no real change, would test and control still track cleanly?

And the grading matters as much as the setup. A real read compares test region before versus after, and test versus a weighted synthetic control, together. If a tool reports “test minus control,” that is subtraction, not incrementality.

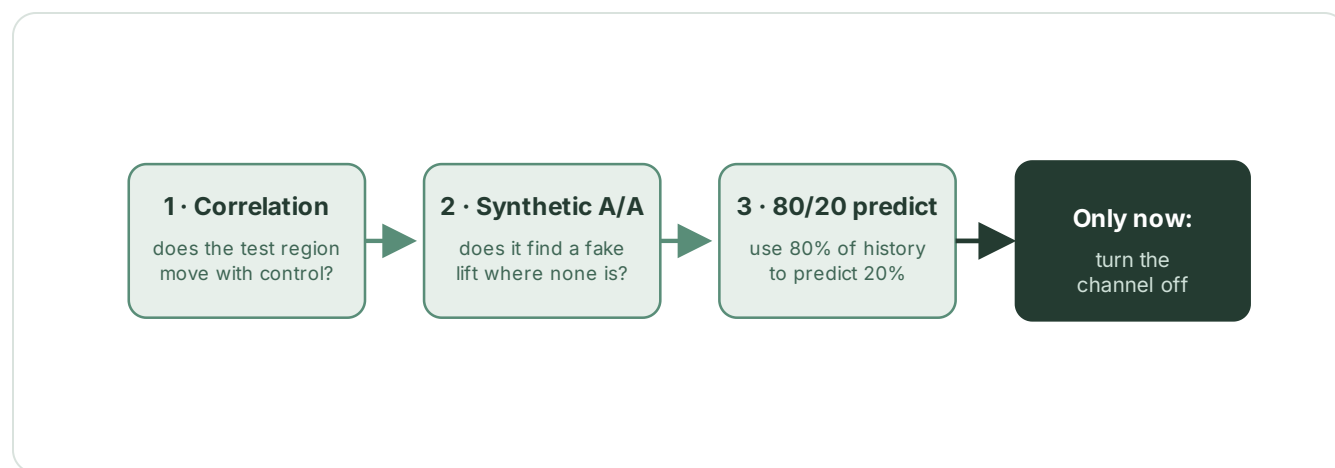
If your tool reports test minus control, it isn't an incrementality tool. It's a subtraction calculator.

THE ASK What was the design, how was the control built, what was excluded, and what could have biased the result?

Strong answer *“Here are the treatment and control markets, the pre-period correlation, the synthetic control, the exclusions, and the contamination risks.”*

Weak answer *“Trust the result.”*

FIGURE 9 Three checks before you turn anything off.



A holdout is only as good as the design that came before it.

Power and Minimum Detectable Effect.

What it is. Two pre-test design questions. Power is the chance the test catches a real effect if one exists. MDE, the minimum detectable effect, is the smallest lift the test could reliably see, given your geographies, spend change, and how long it runs.

Two flavors matter. Ex-ante MDE is the planning number, computed before the test from historical baseline variance: the smallest lift this design could detect if everything goes as expected. Ex-post MDE is the realized number, computed after the test from the actual residual variance during the window: the smallest lift the test actually could have detected, given the noise that showed up. A defensible deliverable reports both. Ex-ante alone tells you whether the design was sound on paper. Ex-post tells you whether the test you actually ran was capable of supporting the result you are reporting.

Why it matters. A flat result does not automatically mean the channel did not work. It can mean the test was too small or too short to see the lift. Those are two completely different conclusions, and they lead to opposite budget decisions.

If a vendor reports “inconclusive” or “no lift” without telling you the MDE, they cannot tell you whether the signal was absent or whether the test was simply blind to it. And these are pre-test numbers. They have to be calculated before the test runs, not explained after it disappoints.

For a geo test, the thing that drives power is the number of geographies and how much they vary, plus the size of the spend change and the duration. Not raw customer counts.

At enterprise scale, MDE calculations have to account for geographic and brand-level effect heterogeneity. The same channel can have a 15% incremental lift in one market and 3% in another, and an aggregate MDE that averages them tells you nothing about either. Properly powered enterprise tests are designed at the cell level, market by brand by channel, which is computationally and operationally heavier than the single-test designs sufficient for smaller brands.

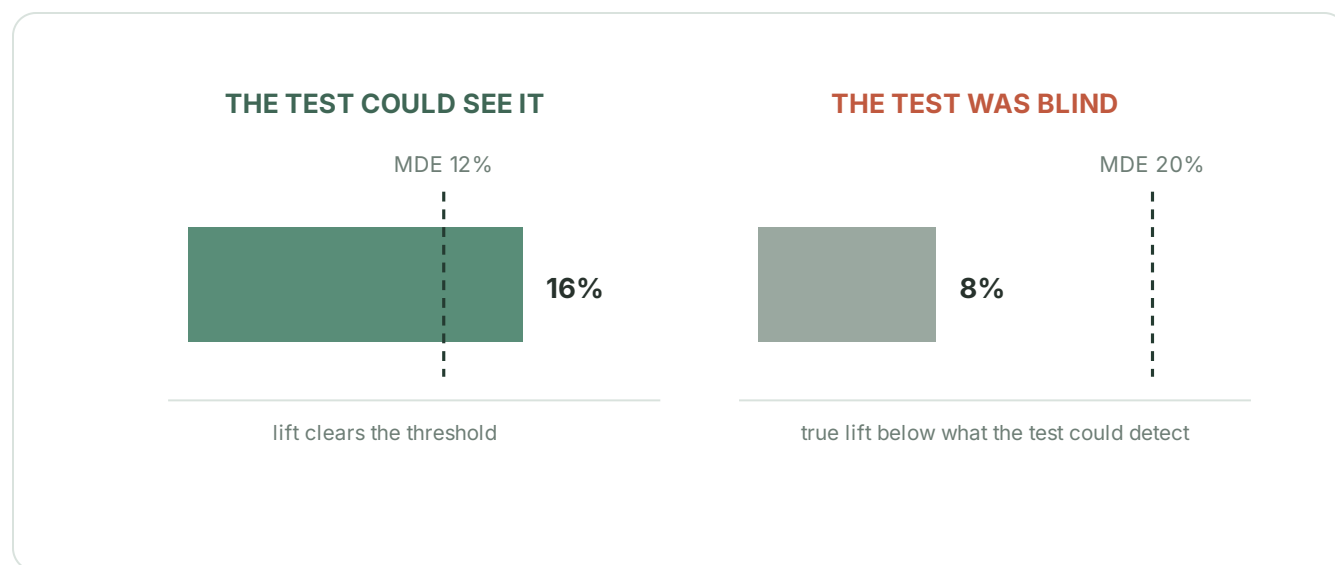
Stella’s read on significance. Below 90% confidence, a result is not actionable. 90% is our operational threshold. Academic convention is 95%; most marketing vendor reporting uses lower. We settled at 90% because it balances false-positive risk against the cost of paralysis on budget decisions that have to be made monthly.

THE ASK What was the MDE, and was the test powered for this specific decision?

Strong answer *"Ex-ante MDE was 8%, ex-post realized at 7.2% from the in-test variance, observed lift was 16%. The lift cleared both thresholds and the credible interval was tight enough to support a 15% budget move."*

Weak answer *"It reached significance."*

FIGURE 10 A flat result can mean two very different things.



Without the MDE, "no lift" and "couldn't see the lift" look identical. They are not.

PART THREE

Decision Quality.

Is it strong enough to move money?

Confidence Bounds.

What it is. Every iROAS or lift number is the model's single best guess. The bounds are the range of plausible answers around that guess.

Why it matters. The point estimate alone hides how sure the model is. A 2.4x iROAS with a tight band, say 1.9x to 2.9x, supports a confident move. The same 2.4x with a band of 0.8x to 4.0x includes outcomes you would never act on. Same headline number. Completely different bet.

A point estimate without an interval is half an answer dressed as a whole one.

One precision that trips people up, and that you should hold a vendor to. iROAS is a revenue number, not a profit number. A 1.0x iROAS means you earned a dollar back for every dollar spent, which at any margin below 100% is a loss. So breakeven is not 1.0x. It is 1 divided by your contribution margin. A 2.4x iROAS at a 40% margin is comfortably profitable. The same 2.4x at a 15% margin is barely above water. When you size a move off an interval, size it against your real breakeven multiple, not against 1.0x. A vendor who tells you a range "clears breakeven" without knowing your margin is using the wrong line, and so is one who calls a range profitable when its lower bound sits below your breakeven multiple.

Because Stella's MMM is Bayesian, these come through as credible intervals from the model's posterior, and the incrementality tests report confidence intervals. The language differs by method. The decision rule does not. Size the move off the interval and the downside, not the single number.

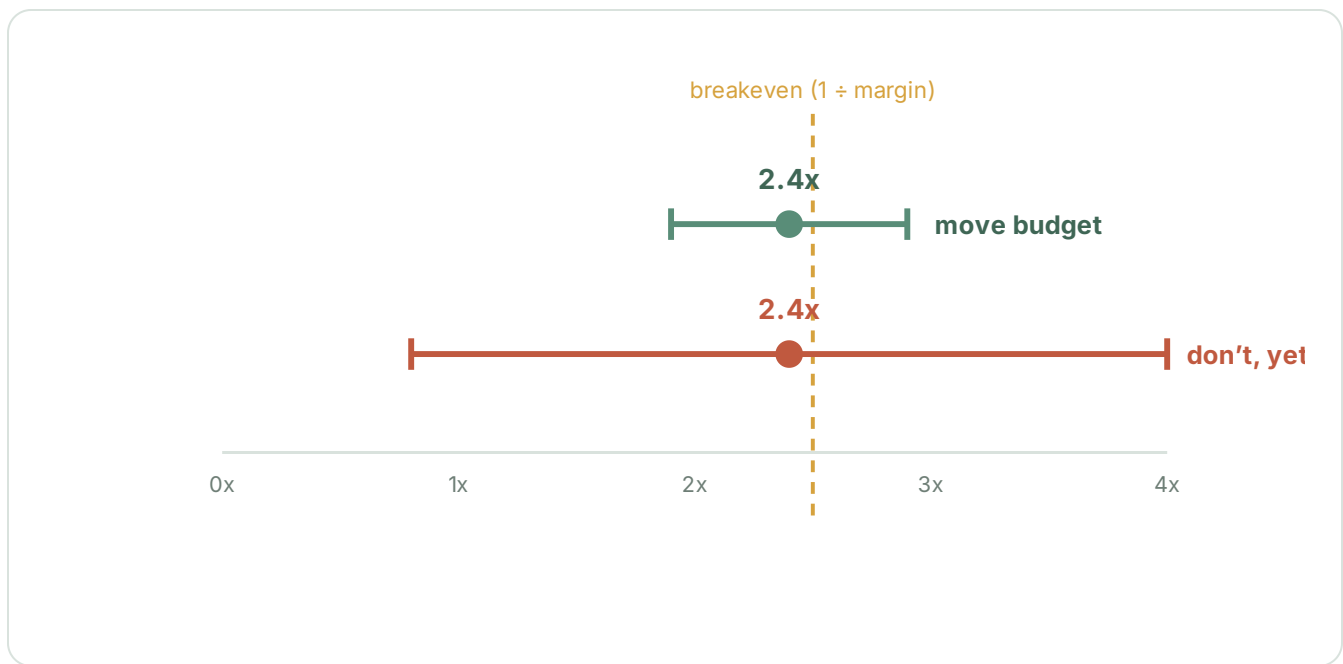
There is a related check worth naming: prediction coverage. Does the model's predicted range actually contain the real outcomes when you look back? A model whose "90% interval" misses the real number half the time is not giving you a real interval. Coverage is the calibration check sitting underneath the bounds, not a tenth check.

THE ASK What are the upper and lower bounds on each number, and where's my breakeven multiple inside that range?

Strong answer "2.4x iROAS, with a 1.9x to 2.9x interval, and your breakeven at a 40% margin is 2.5x, so the whole interval clears it. Tight enough to move budget."

Weak answer "It was significant."

FIGURE 11 Same number, different decision.



The point estimate is the part everyone quotes. The interval, measured against your real breakeven multiple, is the part that decides whether you should act on it. The wide interval crosses breakeven, so the same headline number isn't yet a move.

The Written Path From Model to Budget.

What it is. A lift number is not a decision. The output only matters when it says what to move, in which direction, by how much, and with how much confidence.

Why it matters. This is where most engagements quietly fail. You get a number and a slide, and nothing changes, because nobody connected the analysis to an action.

“Meta is 2.9x incremental ROAS” is a number. It is not a decision. “Meta is estimated at 2.9x, enough confidence to raise spend 15% next month but not enough to double it, while branded search is over-credited by platform reporting and should hold or fall first.” That is a decision. Someone can act on it and defend it in a QBR.

A finished recommendation carries the uncertainty all the way through. It accounts for diminishing returns on the channel you are scaling, for whether the channel-level number was even cleanly identified (back to VIF), for whether the number clears your breakeven multiple and not just 1.0x, and for whether the evidence is experimental, modeled, or both. Experimental and modeled evidence do not carry equal weight, and a serious recommendation says which it is leaning on.

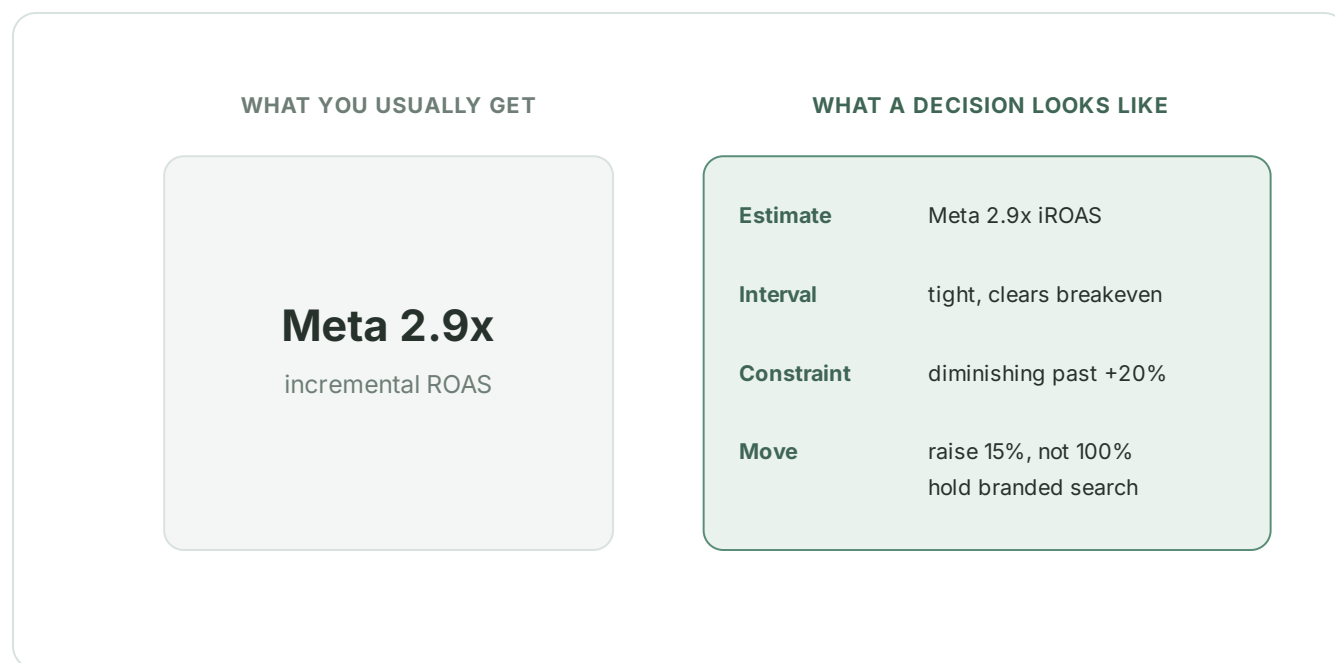
If the work ends at the lift number with no recommendation and no uncertainty, nothing will change. You bought a history textbook.

THE ASK What budget change do you recommend, and what uncertainty sits around that move?

Strong answer “Increase Meta 15%, hold branded search, and re-test YouTube before scaling.”

Weak answer “Meta is 2.9x iROAS.”

FIGURE 12 A number is not a decision.



The receipt isn't complete until the number turns into a move someone can defend.

The Tear-Out.

FIGURE 13 The tear-out card.

TAKE THIS INTO YOUR NEXT REVIEW

The nine questions.

Ask them verbatim. Watch which get a straight answer and which get a subject change.

- 1 What was your out-of-sample MAPE, and which window did you hold back?
MMM
- 2 What was the out-of-sample R-squared, and how did you check for overfitting?
MMM
- 3 Was the model forward-tested on a held-out window, and is a backtest running against live revenue?
MMM
- 4 Can you show me the VIF for our channels, and what changed because of it?
MMM
- 5 Did you check the residuals for patterns, and what did you find?
MMM
- 6 What was the test design, how was the control built, and what could have biased it?
INCREMENTALITY
- 7 What was the ex-ante MDE, the ex-post MDE, and was the test powered for this decision?
INCREMENTALITY
- 8 What are the bounds on each number, and where is my breakeven multiple inside that range?
BOTH
- 9 What budget change do you recommend, and what uncertainty sits around that move?
BOTH

Six red flags to keep next to you on the call.

- **In-sample fit presented as proof.** A memory test, not a prediction.
- **A suspiciously perfect R-squared.** Usually memorized noise, not brilliance.
- **One number, no interval.** Half an answer dressed as a whole one.
- **A range called profitable against 1.0x** instead of your real breakeven multiple. Wrong line.
- **Clean per-channel ROI for channels that obviously move together.** Precision they haven't earned.
- **A deck instead of a decision.** If it ends at the lift number, nothing changes.

If they can't answer half of these, you don't have measurement. You have a dashboard with opinions.

Directional, not a verdict. A full set of answers does not prove the conclusion is right. It proves the work is auditable. Across Stella's 225-test benchmark, the relationship between pre-test fit quality and eventual significance was visibly stronger than the relationship between either budget or duration and significance. A strong first screen, not proof of causal correctness.

Why Stella Shows All Nine.

The measurement vendor landscape at the enterprise scale falls into three categories. Large agencies offer measurement as one service among many, typically delivered by junior staff with senior names attached, with methodology that does not survive technical due diligence. Specialist consultancies offer methodological depth with limited operational integration: strong analysis that ends in a deck rather than a budget decision. Self-serve platforms offer scale and speed at the cost of customization and rigor.

Stella sits in a fourth category: a measurement consultancy operating at the technical depth of a specialist firm, with the operational integration of a platform, and the personnel model of a senior advisory engagement. Every model is built and reviewed by PhD data scientists. Every methodology choice is auditable. Every output ties to a budget decision the client's finance function can defend.

This is what the rest of this guide has been describing, the standard of work, applied end to end, on every engagement.

Every other consultancy asks you to trust them. We show you the receipts so you can fact-check us.

That is the whole posture. We are a measurement consultancy, so we own the problem end to end. We design the holdouts, build the model, and turn the output into budget decisions. But we do not ask you to take any of it on faith. Every number we put in front of you comes with the nine checks attached, so you can grade our homework the same way this playbook just taught you to grade everyone else's.

Some consultants told us to hide these metrics. "Marketers will be overwhelmed. They won't understand it." Too bad. You need to know whether the number you are about to base a \$500K decision on is accurate or garbage. If we are going to show you a number, we are going to show you the confidence around it.

There is also a reason to believe the checks are not just hygiene. Across our 225-test benchmark, the quality of a model's pre-test fit predicted whether a test reached significance better than the test's budget or its duration did. In plain terms: whether the setup was sound mattered more than how much was spent or how long it ran. That is the empirical case for checking the setup before you trust the result, and it is why these nine are the nine.

FIGURE 14 What a complete receipt looks like. (Illustrative values.)

STELLA · THE RECEIPT		TYPICAL VENDOR	
Out-of-sample R-squared	0.89	In-sample R-squared	0.98
Out-of-sample MAPE	8.7%	<i>nothing else shown</i>	
Forward test gap	small		
Backtest	live, monthly		
VIF	under 5		
Residuals	checked		
Test design	3 pre-checks		
Significance	92%		
Bounds + breakeven	shown		
Written recommendation	attached		

Nine checks versus one. That's the difference between auditable and trust-me.

So here is what is on every model we run. Validity metrics front and center, out-of-sample R-squared, MAPE, VIF, statistical significance, and prediction coverage, on every output, not in an appendix you have to ask for. A Bayesian MMM, calibrated with your own holdout results as priors, so the portfolio model is anchored to causal proof, not just historical correlation. The backtest and forward test built in, so you verify predictions against actual revenue before the quarter ends. Incrementality graded properly, with three location checks before any test and two comparisons at the end, not subtraction. And the people reading your results built the models. A dedicated PhD data scientist on a managed account, not a junior analyst handed the file after onboarding.

A consultancy that hides its accuracy isn't a data science partner. It's a vendor with a nicer deck.

We are not asking you to trust us instead of the math. We are asking you to make every partner, us included, show the math. So point the nine back at this very playbook. Every threshold in here is labeled as our working band. Check us on it.

What to Do.

You have three options. Most people pick the first one without noticing.

Keep trusting the number you can't check. Keep moving budget on in-sample fit and platform ROAS. Keep explaining to your CFO why growth is uneven. Most brands will do exactly this. That is your edge if you do not.

Audit your current vendor yourself. Take the nine questions into your next review. You will learn more from which ones they dodge than from any answer they give you. This costs nothing and you can do it tomorrow.

Hand it to a consultancy that shows the receipts. If you want a partner to own the measurement end to end, that is what managed is for. A dedicated PhD data scientist, custom models built around your KPIs, holdouts and MMM built and run by our team, turned into budget decisions and presented to your board. Done for you, with every one of the nine checks on the table, every time.

We meet you wherever you are on that.

Pricing.

Stella engagements at enterprise scale begin at six figures annually, structured as a 12-month minimum commitment with quarterly checkpoints. Scope is defined per engagement: number of brands in the portfolio, number of markets, number of measurement workstreams, level of platform integration. We do not publish pricing because enterprise engagements are not commodity engagements. The work is scoped to what the business actually needs.

Pricing for mid-market brands (\$25M to \$250M annual revenue) is available on a published-tier basis at stellaheystella.com.

Who Stella is not for.

Brands with under \$25M in annual revenue, or under \$200K per month in paid media spend, will likely find the engagement cost exceeds the marginal value of the analysis. Direct-response businesses with single-channel exposure (for example, 90% of spend on Meta) do not have the channel diversity to benefit from full-stack causal measurement. Organizations without a dedicated measurement or analytics function will struggle to operationalize Stella's recommendations, regardless of methodological quality.

We turn down approximately 40% of enterprise inbound for one of these reasons. The fit question matters more than the contract.

60 minutes. Both founders. Your data on screen.

We connect Stella to your ad accounts, or work from a CSV, and run incrementality, MMM, and always-on attribution on your real mix. You see what the platform does, what it shows, and what it would tell you to do differently next month. Every number with its receipt attached.

Engagements begin with a 60-minute scoping discussion with both founders, focused on your current measurement architecture and the specific decisions that depend on it. We will tell you within that hour whether we believe Stella is the right partner for your situation, and which alternatives we would recommend if not.

Book your call → stellaheystella.com

STELLA GROWTH INTELLIGENCE

Know what's causing growth. Know what to do next.

© 2026 Stella. All rights reserved.