FROM THE RESEARCH BENCH

Evidence over anecdote — what a number has to survive.

What a doctorate in cross-border growth teaches about proof — and how to hold the numbers in a commercial call to the same bar.

Sinan Isoglu, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher

PUBLISHED
1 August 2026

READING TIME
12 min

WORDS
2,646

https://isoglu.com/journal/evidence-over-anecdote/

CONTENTS

MANAGEMENT SUMMARY

The number in the deck and the number in the paper are both survivors of a selection process — and only one of the two pages shows what its number survived. This essay assembles the measured record: the one setting with an honest denominator — every trial two US nudge units ran, averaging 1.4 percentage points against a published sample averaging 8.7 — and what checking did to vetted findings across psychology, economics, medicine and cancer biology, by amounts that refuse a single constant. It hands over three questions that grade any number entering a commercial call — out of how many, decided when, compared to what — and states the limit plainly: they grade coarsely, they produce no corrected number, and where none of the three can be answered, the honest basis is judgement, declared as judgement.

Keywords: Evidence quality · Selection bias · Replication · Commercial decision-making · Case studies

A vendor's deck and a journal paper land on your desk in the same week, ahead of the same decision. The deck carries a customer that looks like you and a double-digit lift. The paper carries an average effect and a page of limitations. If you are honest, the deck moves you more. The rule doctoral training installed in me says the instinct is aimed at the wrong axis: a number is graded by the selection process it survived, not by the document it arrived in. The real difference between the two pages is that one of them shows you what its number survived, and the other shows you nothing.

"Evidence over anecdote" is usually read as papers over decks. That is not the claim here. The deck is evidence too, of something, for some question. The claim is that numbers reach decisions through filters, that the filter decides what a number can prove, and that the deck genre shows none of the filtering. So the working question is not which document to respect. It is what each number had to survive to reach the page.

What selection alone does to a number, measured once

The cleanest measurement of that filtering comes from an unglamorous corner: government nudge units. Stefano DellaVigna and Elizabeth Linos assembled every randomised trial run by two of the largest units in the United States — 126 trials covering 23 million people, nothing left in a drawer — and held them against a sample of published nudge trials drawn from two academic meta-analyses. In the published sample, the average intervention improved take-up by 8.7 percentage points, a 33.4% lift over control. Across everything the units actually ran: 1.4 points, an 8.0% lift — in their words, "still sizable and highly statistically significant."

The gap is not fraud, and it is mostly not even the interventions. Their model attributes about 70% of the difference to selective publication exacerbated by low statistical power: individually honest studies, filtered by which results were striking enough to write up. The units' number is honest for one reason only. They kept the denominator, so every attempt counted, and nothing could be quietly missing.

Two cautions travel with the pair, and they matter more than the pair. These are US government nudge units moving take-up rates, not campaigns moving revenue; the sizes do not transfer. And 8.7-to-1.4 is a fact about a published sample against a full census in one field, not a discount rate for your pilot. No such rate exists, as the next section shows.

Even vetted numbers shrink when someone checks — by no single amount

A published finding has already survived hostile review, and on average it still shrinks when somebody re-runs it. That is the uncomfortable half of the replication record. The half that matters for a commercial reader is what the shrinkage refuses to be: one number.

Table 1 **What checking did to published numbers, setting by setting**

Setting	What was done	What happened to the published numbers	What that quantity is
Psychology — 100 studies, three journals	Re-run with high-powered designs and original materials where available	"Replication effects were half the magnitude of original effects." 97% of originals were significant; of the replications, 36% were significant, 47% of original effects sat inside the replication's 95% confidence interval, 39% were rated as having replicated, and 68% stayed significant with original and replication evidence combined	Effect-size ratio plus four success measures — the authors report four precisely so that no single verdict exists
Laboratory economics — 18 studies, two top journals	Re-run at 90%-plus power under pre-defined plans	Replicated effect sizes averaged 66% of the original; 61% showed a significant effect in the original direction; four further replicability indicators ran 67–78%	Effect-size ratio and pass rates — a floor-condition decay, with almost nothing about the setting changed
The most-cited clinical research — 49 studies examined	Held against later, larger or better-controlled studies	Of the 45 the paper counts: 7 contradicted (16%), 7 initially stronger than what followed (16%), 20 replicated (44%), 11 never seriously re-tested (24%)	Verdict counts on famous findings, selected for fame, not a base rate. Most were never overturned
Preclinical cancer biology	Independent re-run of published experiments	Median replication effect 85% smaller than the original; 92% of replication effects came out smaller than their originals	The extreme of the range, in the setting furthest from a commercial reader
Nudges — a published sample against practice	A full census of two US units' 126 trials, held against a sample from two meta-analyses	8.7 percentage points in the published sample; 1.4 across everything run; about 70% of the gap traced to selective publication with low power	A selected sample against an honest denominator — not a discount rate

Open Science Collaboration (2015), Science 349(6251), aac4716, abstract; Camerer et al. (2016), Science 351(6280), 1433–1436, abstract; Ioannidis (2005), JAMA 294(2), 218–228; Errington et al. (2021), eLife 10:e71601; DellaVigna & Linos (2022), Econometrica 90(1), 81–116, abstract. Each row reports a different quantity; no shared scale exists, and that is the finding.

Read the table's direction, not its cells. One detail deserves its own sentence. In the clinical set — Ioannidis's JAMA survey — the six highly cited non-randomised studies fared far worse than the trials: five were later contradicted or had reported effects stronger than what followed, against nine of thirty-nine randomised trials. The weaker the design, the harder the fall. Measured on six and thirty-nine, so treat it as a signpost rather than a law.

Note what the strongest of these checks enjoyed. The economics replications ran at ninety-percent--plus power under pre-defined plans; the psychology project worked from the original materials where it could; in both, the failures were published. The clinical row differs in method as well as in fame — not a re-run but a comparison against later, larger studies, which is closer to how a commercial number ages. Where a published number met real checking machinery and still shrank, the shrink was measured under conditions built to catch it. The numbers in a commercial deck have usually faced nothing of the kind. That is my reasoning, not a measured result, and the next section says why I think it holds.

Three doors stand between a commercial number and the truth

The first door is selection, the mechanism measured above. A deck's case study is chosen from the population of attempts, and chosen for how it reads. In research, the size of that filter could be measured because two units kept their denominator. Commercial reporting rarely keeps one: in the pipelines I have worked in and around, nobody was ever asked to, and nothing in the deck genre compels the attempts that failed onto the page. So the filter's direction is known. Its size, for the number in front of you, is not.

The second door is confounding, and its sign is unknown. Even your own dashboard's numbers, with no vendor involved, routinely disagree with what an experiment finds. Brett Gordon and colleagues took fifteen US advertising experiments at Facebook, with the platform's own user-level data, and compared the experimental answer with what industry-standard observational methods report: "in half of our studies, the estimated percentage increase in purchase outcomes is off by a factor of three across all methods" — and off in both directions, which for a decision-maker is worse than a consistent bias. The data predate 2019's privacy changes, which have since made observational measurement harder, not easier.

At eBay, Blake, Nosko and Tadelis found experimental returns on paid search a fraction of the non-experimental estimates, and ads on the company's own brand keywords showed "no measurable short-term benefits". One advertiser with a dominant brand: carry the setting, not a law of paid search.

The third door is the metric found after the fact. Run enough attempts and watch enough metrics, and something will always have moved; a success declared after the results are in is a different kind of claim from a success defined before them. Research meets this door at industrial scale — Microsoft's Bing, with "over 200 concurrent experiments" running on a given day, treats "a high occurrence of false positives" as an operational fact to be managed. And it is the door John Ioannidis's PLoS Medicine essay models: flexibility "in designs, definitions, outcomes, and analytical modes" erodes the reliability of what gets claimed. His headline conclusion is genuinely disputed — Goodman and Greenland's objections are in the references — but the direction of the flexibility problem is not the disputed part.

Why build machinery against your own numbers at all? Because judgement kept failing at scale. The Bing team reports that "many ideas impact key metrics by 1% and are not well estimated a-priori", and that the system caught negative features "despite key stakeholders' early excitement". Ron Kohavi and colleagues compressed the lesson into a phrase that deserves its fame: returns come when teams listen to their customers, "not to the Highest Paid Person's Opinion (HiPPO)".

Notice what the example concedes. The correction machinery exists in commercial life, and Bing ran it industrially. The asymmetry is not research versus business; it is the artefact. A published finding is bound to the machinery that checked it: named methods, a review, a record the next team can attack, as the replication projects in the research cluster's sibling essay attacked theirs. A deck can be assembled with or without any of this behind it, and it looks the same either way. That likeness is what the next section prices.

The bar that travels: three questions

Doctoral training does not equip you to run a randomised trial on a market entry, and the experimenters themselves flag the limits of their instrument — Kohavi's survey names them "both technical and organizational". What the training actually changes is cheaper. It installs three questions, one per door, askable in any meeting without a line of statistics:

1. Out of how many?
2. Decided when?
3. Compared to what?

Out of how many prices the selection door. The census shows what a real answer looks like: every attempt counted, nothing missing. The commercial version is smaller. How many pilots, accounts or tests does this number stand on, and how many ran that I am not seeing? Often the honest answer is that nobody kept one. A refusal here tells you the denominator will not reach the decision — not that it does not exist, and not that the number is false. The difference matters.

Decided when prices the flexibility door. Was the success metric fixed before the result existed, or found afterwards in whatever moved? The norm is imported from research practice, where the pre-registered design exists because its absence proved expensive; its practitioner echo is Kohavi's "clear evaluation criteria" fixed in advance. What AI actually changes in revenue operations ran this question against one literature; here it becomes standing equipment.

Compared to what prices the confounding door. What would this metric have shown if the initiative had done nothing — a holdout, a stagger, a geography, or an honest "we cannot know"? That is the question the Facebook study answers with a factor of three, on platform-grade data. A number with no comparison is not wrong; it is unpriced.

Three questions grade coarsely. They produce no corrected number; no shrinkage constant exists to hand over, and Table 1 is the refusal. And they grade a number only as evidence about its own setting: whether Bing's or eBay's or a nudge unit's world reaches yours is a judgement no question automates. What the questions buy is economic. A pipeline can decline to answer, or go vague, and vagueness against a direct question is itself an answer. Silent omission is cheap and deniable. A false answer to a direct question is neither. That, not statistical sophistication, is what the bar buys.

Where the questions stop

The biggest commercial calls are the ones the instrument serves least. A market entry, an acquisition, a pricing reset: one attempt, no honest denominator, often no counterfactual. There, the three questions return “unanswerable”, and that output is the useful one. Jacob Cohen spent a career watching researchers wring certainty out of the significance test — published research’s own ritual — and wrote the sentence that generalises: the test “does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!”

When the questions come back empty, the honest basis for the call is judgement: experience, structure, appetite for the downside — declared as judgement, not dressed as a borrowed number. And the deck keeps its real job. A case study selected for relevance can be honest evidence for existence — someone like you has made this work — provided its own number survived the other two doors. What it cannot carry is an estimate of magnitude, of what the average attempt would deliver. Knowing which question a number can answer is most of what “evidence over anecdote” ever meant.

Figure 1 **Three questions before a number enters the decision**

THE NUMBER, AS IT ARRIVED Verbatim, with its source and date.	OUT OF HOW MANY? Attempts behind it — seen and unseen.	DECIDED WHEN? Metric fixed before or after the result.	COMPARED TO WHAT? What it would show if nothing worked.

A blank cell is a finding. Three blanks: the number answers an existence question at best — decide on declared judgement.

Author’s own worksheet.

Run the sheet on the next number that asks for budget. Not to catch anyone: most pipelines were never asked to keep denominators, and finding that out is diagnosis, not indictment. The point is smaller and harder. The bar was never where a number was printed. It is what the number survived, and whether anyone can say.

REFERENCES

Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. Econometrica, 83(1), 155–174. https://doi.org/10.3982/ECTA12423

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. Science, 351(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. Econometrica, 90(1), 81–116. https://doi.org/10.3982/ECTA18709

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. eLife, 10, e71601. https://doi.org/10.7554/eLife.71601

Goodman, S., & Greenland, S. (2007). Why most published research findings are false: Problems in the analysis. PLoS Medicine, 4(4), e168. https://doi.org/10.1371/journal.pmed.0040168

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science, 38(2), 193–225. https://doi.org/10.1287/mksc.2018.1135

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. JAMA, 294(2), 218–228. https://doi.org/10.1001/jama.294.2.218

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

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

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

END OF ESSAY

HOW TO CITE THIS ESSAY

Isoglu, S. (2026, August 1). Evidence over anecdote — what a number has to survive. Sinan Isoglu. https://isoglu.com/journal/evidence-over-anecdote/

KEEP READING

From the research bench

What peer review deleted, the web kept.

https://isoglu.com/journal/what-peer-review-deleted/

From the research bench

A vendor page supplied the source for the 5x retention rule

https://isoglu.com/journal/a-vendor-page-invented-the-source/

Revenue operations & AI

What AI actually changes in revenue operations

https://isoglu.com/journal/what-ai-changes-in-revenue-operations/

Sinan Isoglu, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher

Online version: https://isoglu.com/journal/evidence-over-anecdote/