Thinking clearly about the value of survey pretesting for practitioners

Ben M. Tappin[1]
London School of Economics and Political Science
b.tappin@lse.ac.uk

Practitioners of various kinds regularly make decisions about whether to deploy "communication" treatments into the world—such as advertisements, reminders, public service announcements, nudges, and advocacy or fundraising appeals—and, if so, which specific version of several different treatment versions to deploy.

Survey pretesting is often used to inform decision-making in these contexts. A survey pretest is where a practitioner runs a randomized experiment in which a sample of people is recruited for an online survey and is exposed to one or more treatments in order to gauge their potential real-world effects. The results then guide decisions about whether to deploy a treatment, which version to deploy, or whether to conduct further treatment development or testing. Despite its common use, the value of survey pretesting for guiding such decisions is often unclear and in many settings contested (*1*–*6*).

In this note, I therefore articulate a simple framework to try and facilitate clearer thinking about the value of survey pretesting for practitioners.

Decisions informed by survey pretesting

To consider the value of survey pretesting for decision-making, I first describe the types of decisions it may be used to inform. There are two key types, which I label (i) "deploy-or-not" decisions and (ii) "what-to-deploy" decisions. These decisions differ in the types of information they require and, therefore, in the features of survey pretesting that are relevant. These decision types are not exhaustive use cases for survey pretesting, but they cover a large proportion of situations in which practitioners might rely on it.

Deploy-or-not decisions

In deploy-or-not decisions, survey pretesting is used to inform **whether or not to deploy** any treatment into the world. Like everyone, practitioners do not have unlimited money and are often budget constrained. Deploying a treatment costs money, and practitioners want to achieve a positive return on investment (ROI). They don't want to waste money. Moreover, practitioners often have multiple actions they could take to achieve their goal, and ideally they want to take the action with the largest ROI. Deploying a communication

[1] I'm grateful to Jonathan Robinson for discussions that prompted me to finally write this note.

treatment into the world could or could not be the largest ROI action available to them. They may use a survey pretest of the treatment to help calculate its expected ROI and thereby inform their decision about whether to deploy or instead take some other action.

For example, a public health practitioner who aims to increase clinic visits could be deciding between deploying a public health campaign or instead upgrading clinic infrastructure. They may use a survey pretest to estimate the effect of the campaign, perhaps estimating that 1 in 100 people exposed to the message would visit the clinic who otherwise wouldn't have. Combining this with the cost of the campaign would yield an ROI value that could be compared against the ROI of upgrading clinic infrastructure.

What-to-deploy decisions

In contrast, in what-to-deploy decisions, survey pretesting is used to inform ***which* treatment to deploy**. The survey is used as a screening tool—a way of identifying the highest impact treatment from among several different versions being considered. The results of the survey pretest may be used to decide which version to deploy in the real world, or which to consider for further testing—perhaps in a slower, more costly but more informative field experiment (i.e., a pretest in a real-world setting, rather than in a survey).

For example, a public health practitioner might have a ring-fenced budget for running a public health campaign and therefore want to maximise its impact. They may use a survey pretest to estimate the effects of different types of messaging and select the top performing message for deployment in the campaign.

Features of survey pretesting that determine its value for decision-making

The value of survey pretesting for informing these decisions is determined in large part by two features of the survey: its **calibration** and its **discrimination**. Specifically, the survey's value for informing deploy-or-not decisions is determined by its calibration, whereas its value for what-to-deploy decisions is determined by its discrimination. Calibration and discrimination are conceptually distinct features of a survey pretest and have distinct implications for decision-making—yet people often get them confused. I think this confusion drives some of the disagreement over the decision value of surveys.

Calibration determines value for deploy-or-not decisions

Calibration refers to the extent to which treatment effects in a survey correspond (on average) to the effects that would be observed in the real world. A well-calibrated survey produces effects that are close in magnitude to real-world effects, whereas a poorly calibrated survey systematically overstates or understates those effects. Calibration concerns *levels* rather than rankings: it is about whether effects are broadly on the right scale, not whether the effects of different treatments are correctly ordered.

Calibration is central to deploy-or-not decisions because these decisions require practitioners to assess whether deploying a treatment is worthwhile relative to its cost. When survey effects are well calibrated, practitioners can combine them with information about deployment costs to approximate the expected ROI of an intervention. When calibration is poor, survey effects are systematically distorted, and if practitioners are unsure of the size of this bias they can end up overinvesting in ineffective treatments or forgoing interventions that would in fact be worthwhile. Figure 1 illustrates.

Notably, uncertainty about calibration does not affect all survey results equally. Consider a practitioner who suspects that their survey overstates real-world effects—that is, calibration is poor; the survey effect is upwardly biased. A null survey result can remain informative even without knowing the precise degree of bias, because deflating a zero effect still yields zero, and the decision is therefore: *don't deploy*.[2] A positive survey effect, by contrast, is much harder to interpret. The practitioner suspects the real-world effect is smaller, but how much smaller matters enormously: the same survey result could imply a worthwhile ROI or a negligible one depending on the degree of bias.

Discrimination determines value for what-to-deploy decisions

In contrast, discrimination refers to the extent to which a survey pretest can distinguish higher-impact treatments from lower-impact ones. A survey exhibits high discrimination when treatments that perform better in the survey also tend to perform better in the real world; and low discrimination when survey performance is weakly related, or unrelated, to real-world performance. Discrimination concerns *rank ordering* rather than levels: it is about identifying relatively better treatments, not estimating their absolute effects.

Discrimination is central to what-to-deploy decisions because these require practitioners to choose among multiple treatments. When survey pretests discriminate well, they can be used as screening tools to prioritize treatments that are more likely to perform well in the real world. When discrimination is poor, survey pretesting offers limited information about which treatment is superior and may be no better than random choice (Figure 1).

[2] This logic depends on the direction of the suspected bias; if the practitioner instead had reason to believe the survey *under*stated real-world effects, then a null result would not support the same inference, since the real-world effect could be positive. I sidestep that here to ease interpretation and because upward bias in surveys is much more likely (as discussed below; see also Table 1).

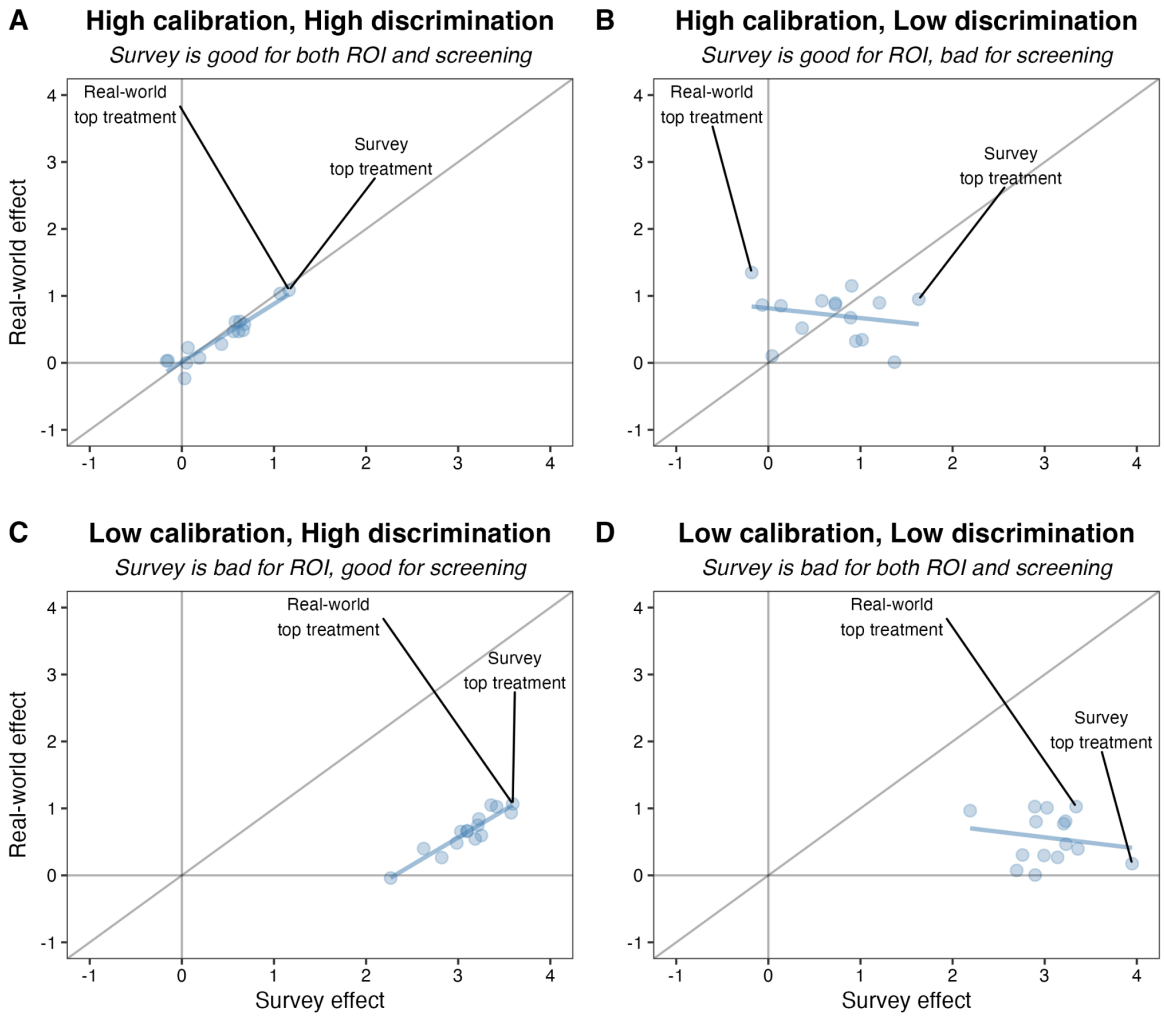

Figure 1. Simplified conceptual illustration of calibration and discrimination in survey effects relative to real-world effects, and the implications for the expected value of survey pretesting. Each panel shows multiple treatments (blue points). The grey diagonal line indicates perfect calibration. Annotated is the top-performing treatment in the survey and in the real world. ROI = return on investment.

Factors that determine a survey's calibration and discrimination

A survey's calibration and discrimination determine its value for decision-making—but what determines its calibration and discrimination? There are many factors, and they depend on the specific context of the survey. Table 1 describes some common factors.

Table 1. Factors which can affect a survey's calibration and discrimination (not an exhaustive list).

Factor	Description	How it can affect calibration	How it can affect discrimination
Sample composition	Whether the survey sample matches the real-world target population in features that affect treatment receptivity (e.g., demographics, attitudes, etc.).	A mismatched sample can inflate or deflate overall effect size.	Treatment rankings may change if subgroups respond differently to different treatments.
Attention	Surveys typically guarantee high attention to the treatment, whereas real-world attention is likely to be partial or fleeting.	Guaranteed high attention inflates effects.	Rankings may shift if treatments differ in their ability to capture/hold attention in the real world.
Effect decay	Surveys usually measure outcomes immediately, but real-world outcomes may be relevant over days or weeks.	Immediate measurement inflates effects.	Rankings may shift if the effects of different treatments decay at different rates.
Demand effects	Respondents may adjust their answers based on perceived survey objectives or social desirability.	This can inflate effects.	Treatments where the "desired" response is more transparent may be disproportionately inflated.
Outcome type	Surveys typically measure self-reported intentions or attitudes rather than real-world behaviors.	The intention-behavior gap can distort effect sizes (typically inflating them).	Rankings may shift if the treatments that move intentions most aren't the ones that change behavior most.
Forced exposure	Surveys guarantee exposure to the treatment, but in the real world people may never see it. (This is different from attention; exposure is a precursor).	Guaranteed exposure can inflate effects.	Rankings may shift if some treatments are more likely to receive exposure than others in the real world (e.g., through re-sharing behavior).
Exposure dosage	Surveys usually present a treatment once, whereas real-world campaigns often involve repeated exposure.	Single-exposure effects may not reflect cumulative impact.	Rankings could shift if the effects of different treatments wear out or accumulate at different rates.
Prior survey context	Survey content that comes before treatment exposure (prior questions, instructions) could prime or frame how respondents react to the treatment.	This can inflate or deflate effects.	Priming may interact differently with different treatments.
Competitive environment	Surveys typically present treatments in isolation, whereas in the real world people may simultaneously be exposed to counter treatments.	Isolation from competition can inflate survey effects.	Some treatments may perform better than others under real-world competition.
Statistical precision	Sample size and experimental design determine how precisely treatment effects are estimated.	Individual estimates may be far from the true value due to noise.	Low precision degrades the ability to reliably rank the true effects of treatments.
Spillover effects	In the real world, treatments can generate indirect effects through social contagion or behavioral ripples that surveys typically cannot capture.	Surveys may underestimate total real-world impact by measuring only direct individual-level effects.	Rankings may shift if treatments differ in their capacity to generate spillover.

In practice, monetary cost also matters

Calibration and discrimination determine whether survey pretesting is valuable **in principle** for deploy-or-not and what-to-deploy decisions. However, the value of survey pretesting in practice also depends on its monetary cost relative to the decision it informs.

The intuition can be felt by considering extreme cases. Imagine two practitioners: one has a total budget of $3 million, while the other has $4000. Each of them runs a survey pretest with multiple treatments to inform their decision about which treatment of several to deploy into the world (i.e., a **what-to-deploy** decision). The survey costs $3000. Suppose the survey has poor discrimination: only 5% of the time would it result in the practitioner identifying a higher-performing treatment than simply choosing at random.

For the wealthy practitioner, exchanging 1/1000th of their budget to identify a treatment that is 5% stronger in expectation looks like a good trade. By contrast, for the less wealthy practitioner, the survey consumes 3/4 of their total budget! Following the survey pretest, their deployment capability has been severely reduced; they may no longer be able to deploy *any* treatment into the world. For them, the survey is harmful—it wouldn't matter even if it had excellent discrimination, it is simply too costly relative to their budget.

Many practitioners facing what-to-deploy decisions in the real world will fall somewhere between these extremes. A pretest costs some fraction of their budget, and the question is whether the imperfect signal it provides is sufficient to be worth it. Even weakly discriminating surveys can be worthwhile if they are sufficiently cheap vs. the budget.

When it comes to **deploy-or-not** decisions, the logic is slightly different. As noted earlier, the informativeness of a survey result for this type of decision depends on whether the result is null or positive. For null results, monetary cost matters because precision is what makes a null informative—and precision requires larger samples, which cost more. A practitioner who spends very little on a pretest may get a null result too imprecise to act upon; one who spends more can obtain a tight null that supports a clear "don't deploy" decision even without knowing the exact degree of upward bias.

For positive results, cost matters differently. Imagine calibration is poor: the survey overestimates real-world effects by 20-fold. If the $3 million practitioner doesn't anticipate this, and decides to deploy based on an ROI calculation from the biased survey effect, then this could be a very costly mistake. It is a more costly mistake than for the $4000 practitioner, who had considerably less money to spend (read: mis-spend) in the first place. On the other hand, if calibration is poor and practitioners can adjust for this, or if calibration is *good*, then spending a fraction of the budget on a survey pretest (e.g., the $3 million practitioner) can be very wise: it can confirm a non-zero real-world effect and feed into an ROI calculation. Like with discrimination, however, if the pretest costs a *large* fraction of the budget (e.g., for the $4000 practitioner), it may not be worth it either way.

To summarise, cost moderates the importance of calibration and discrimination as criteria for evaluating survey pretesting in practice. For what-to-deploy decisions, lower costs make weaker signals acceptable, whereas higher costs demand stronger signals for value. For deploy-or-not decisions, cost governs how informative null results are (through precision) and how damaging positive results can be (through the scale of potential mis-spending). Notably, the advent of synthetic survey participants simulated by large language models may further shrink the cost of pretesting in the future, albeit potentially at some reduction in calibration or discrimination or both (*7*, *8*).

Implications for practitioners

The framework in this note distinguishes two decision types, two corresponding features of survey pretesting that determine its value, and the moderating role of monetary cost. But what should practitioners actually do with this information? Unfortunately, the specifics depend heavily on context—there is no one-size-fits-all prescription. However, the framework can be put to practical use in three ways, at increasing levels of specificity.

1. Questions to ask

At the most general level, the framework provides a set of questions that practitioners should ask themselves before commissioning or interpreting a survey pretest.

First: what decision am I trying to make—and therefore what feature of the survey matters most? If I am trying to determine whether to deploy a treatment at all, then I need the survey to be well calibrated (or I need to know the rough magnitude of bias), because I will be using the effect estimate to assess ROI. If I am trying to identify the best treatment from among several candidates, then I need the survey to discriminate well, because I will be using it to rank treatments rather than to estimate their absolute effects.

Second: which of the factors described in Table 1 are most likely to be problematic in my specific context, and how strongly and in which direction might they affect the relevant feature of the survey (calibration or discrimination)?

Third: what fraction of my budget does this pretest consume, and what is my alternative if I skip it—deploying without any pretesting, relying on intuition, or something else?

These questions do not yield automatic answers, but they impose useful discipline on thinking that can prevent common errors, such as concluding that a survey pretest is uninformative for what-to-deploy decisions merely because it has poor calibration.

2. Rules of thumb for clear cases

Although many real situations will involve difficult unknowns and trade-offs, the framework identifies cases where the direction of the answer is relatively clear.

When a survey pretest costs a trivial fraction of the total budget and the practitioner faces a **what-to-deploy** decision, it is difficult to construct a scenario in which pretesting is harmful: even weak discrimination provides some information advantage over choosing at random, and the cost of obtaining that advantage is negligible. At the other extreme, when a pretest consumes a large fraction of the budget, it may not be worthwhile regardless of its discrimination—the practitioner's remaining deployment capability may be too diminished for the informational gain to be beneficial on net.

For **deploy-or-not** decisions, the general expectation should be that survey pretests will overstate real-world effects. Most surveys guarantee exposure and high attention to the treatment, measure outcomes immediately, and rely on self-reported intentions or attitudes—all of which tend to inflate effect sizes relative to what would be observed in the field. Practitioners making deploy-or-not decisions should therefore usually apply a discount to survey effect estimates, especially when using them as inputs to an ROI calculation. The size of that discount will depend on the specifics of their context: how many of the factors in Table 1 are present and how strongly they point toward inflation.

3. Quantitative analysis for the less clear cases

For practitioners who fall between these extremes—where the pretest is neither trivially cheap nor prohibitively expensive, and where the direction and/or magnitude of bias in calibration is highly uncertain—more precise guidance requires quantitative analysis tailored to the specific context. Such an analysis involves formalising the decision problem and specifying assumptions about the key unknown parameters. For instance, for **what-to-deploy** decisions, the expected correlation between survey effects and real-world effects and the distribution of true treatment effects across the candidates being tested (which jointly govern discrimination); for **deploy-or-not** decisions, the expected degree of systematic over- or under-estimation of effects (which governs calibration). Together with colleagues I have previously conducted such an analysis for what-to-deploy decisions in the context of public health communication (*9*), which may be helpful as a template for similar decisions in other domains. The framework presented in this note can help practitioners identify which inputs to that analysis matter most for their particular decision.

References

1. H. Dai, S. Saccardo, M. A. Han, L. Roh, N. Raja, S. Vangala, H. Modi, S. Pandya, M. Sloyan, D. M. Croymans, Behavioural nudges increase COVID-19 vaccinations. *Nature* **597**, 404–409 (2021).

2. M. G. Findley, B. Laney, D. L. Nielson, J. C. Sharman, External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation. *J. Polit.* **79**, 856–872 (2017).

3. S. Saccardo, H. Dai, M. A. Han, S. Vangala, J. Hoo, J. Fujimoto, Field testing the transferability of behavioural science knowledge on promoting vaccinations. *Nat. Hum. Behav.* **8**, 878–890 (2024).

4. N. Carnes, G. L. Henderson, Not Getting the Message on Climate? Attention as a Key Barrier to Mass-Marketing Experimentally-Validated Messages. *Br. J. Polit. Sci.* **55**, e106 (2025).

5. J. Barabas, J. Jerit, Are Survey Experiments Externally Valid? *Am. Polit. Sci. Rev.* **104**, 226–242 (2010).

6. T. Incerti, Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design. *Am. Polit. Sci. Rev.* **114**, 761–774 (2020).

7. L. Hewitt, A. Ashokkumar, I. Ghezae, R. Willer, Predicting Results of Social Science Experiments Using Large Language Models. (2024).

8. J. R. Anthis, R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, J. Evans, E. Brynjolfsson, M. Bernstein, LLM Social Simulations Are a Promising Research Method. arXiv arXiv:2504.02234 [Preprint] (2025). https://doi.org/10.48550/arXiv.2504.02234.

9. B. M. Tappin, L. B. Hewitt, Using survey experiment pretesting to support future pandemic response. *PNAS Nexus* **3**, pgae469 (2024).