10

Complementary Techniques

If all you have is a hammer, everything looks like a nail
– *Abraham Maslow*

Why you care: *When running experiments, you also need to generate ideas to test, create, and validate metrics, and establish evidence to support broader conclusions. For these needs, there are techniques such as user experience research, focus groups, surveys, human evaluation, and observational studies that are useful to complement and augment a healthy A/B testing culture.*

The Space of Complementary Techniques

To have successful A/B experiments, we not only need the care and rigor in analysis and in creating the experimentation platform and tools, but we also need:

- Ideas for experiments, that is, an *ideas funnel* (Kohavi et al. 2013).
- Validated metrics to measure the effects we care about.
- Evidence supporting or refuting hypotheses, when running a controlled experiment is either not possible or insufficient.
- Optionally, metrics that are complementary to the metrics computed from controlled experiments.

For an idea funnel, you want to use every method at your disposal to generate ideas, including methods like observing users in a user experience study. For ideas that are easy to implement, we recommend testing them directly by running a controlled experiment; however, for ideas that are expensive to implement, you can use one of these complementary techniques for early evaluation and idea pruning to reduce implementation cost.

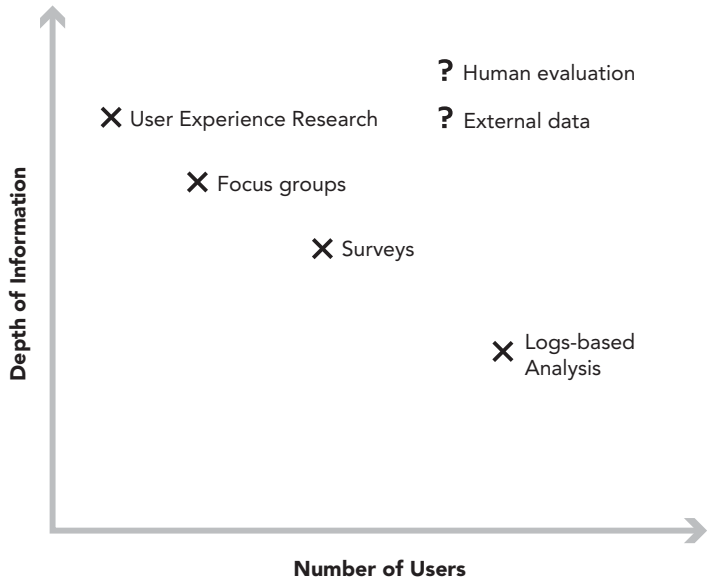

Figure 10.1 Number of users versus depth of information per user

As another example for using complementary techniques, what if you want a reliable proxy metric for user satisfaction, a concept that is quite difficult to measure. You can run a survey and gather self-reported user satisfaction data, and then analyze instrumented logs data to see what large-scale observational metrics correlate with the survey results. You can extend this further by running controlled experiments to validate the proposed proxy metrics.

The methods we discuss in this chapter vary along two axes: scale (i.e., number of users) vs. depth of information per user, as summarized in Figure 10.1, and as we discuss each in turn, we will see the tradeoff in terms of the generalizability that comes from the scale relative to the details we can get from lower-scale methods.

Logs-based Analysis

One pre-requisite for running trustworthy A/B experiments is having proper instrumentation of user views, actions, and interactions to compute metrics for evaluating controlled experiments. The same is true for logs-based analyses, also called *retrospective* analyses. These help with:

- **Building intuition**: You can answer questions, such as the following, to define metrics and build intuition:
  - What is the distribution of sessions-per-user or click-through rate?
  - What is the difference by key segments, such as by country or platform (see Chapter 3)?
  - How do these distributions shift over time?
  - How are users growing over time?

  Building this intuition helps you understand your product and system baseline, what the variance is, what is happening organically independent of experimentation, what size change might be practically significant, and more.
- **Characterizing potential metrics**: Building intuition is the precursor for characterizing potential metrics. Characterization helps you understand the variance and distributions, how new metrics correlate with existing metrics. Log-based analyses establish understanding of how a potential metric might perform on past experiments. For example, is it useful for making decisions? Does it provide new/better information than existing metrics?
- **Generating ideas for A/B experiments based on exploring the underlying data:** You can examine the conversion rate at each step of the purchase funnel to identify large drop offs (McClure 2007). Analyzing sessionized data can uncover that a particular action sequence took longer than expected. This discovery path leads to ideas of how to make your product better, whether you're introducing new features or UI design changes.
- You can explore whether ideas generated using these complementary techniques happen at-scale and are worth investing time implementing and evaluating using an A/B experiment. For example, before investing in making an e-mail attachment easier to use, get an upper-bound sizing of the impact by analyzing the number of attachments sent.
- **Natural experiments**: These occur occasionally, either due to exogenous circumstances (e.g., an external company changing a default) or bugs (e.g., a bug that logs all users out). In those cases, run an observational analysis (see Chapter 11) to measure the effect.
- **Observational causal studies** (see Chapter 11): You can run these studies when experiments are not possible, for example, you can use quasi-experimental designs. When you use quasi-experimental designs in combination with experiments, they can lead to an improved inference of a more general result.

Logs-based analyses can serve many purposes complementary to A/B experiments. One limitation is that these analyses can only infer what will

happen in the future based on what happened in the past. For example, you may decide not to further invest in the e-mail attachment feature because current usage is small; however, the current low usage might have been caused by the fact that it is difficult to use, which logs-based analysis may not reveal. Combining logs-based analysis with user and market research, as we discuss later in this chapter, gives a more comprehensive picture.

Human Evaluation

Human evaluation is where a company pays a human judge, also called a *rater*, to complete some task. The results are then used in subsequent analysis. This is a common evaluation method in search and recommendation systems. Simple ratings can be questions such as, "Do you prefer side A or side B," or "Is this image pornographic?" and can get progressively more complicated, such as, "Please label this image," or "How relevant is this result for this query." The more complicated rating tasks may have detailed instructions to ensure more calibrated ratings. Typically, multiple raters are assigned the same task, as raters may disagree; you can use various voting or other disagreement resolution mechanisms to obtain high-quality aggregate labeling. For example, the quality of data from pay-systems, such as Mechanical Turk (Mechanical Turk 2019), varies depending on the incentives and payment amount, increasing the importance of quality-control and disagreement resolution (Buhrmester, Kwang and Gosling 2011).

One limitation of human evaluation is that raters are generally not your end users. Raters are doing tasks assigned to them – often in bulk, whereas your product is something your end users come by organically to their lives. In addition, raters can miss the local context of real users. For example, the search query "5/3" to many raters is an arithmetic query and will expect the result 1.667, yet users living near the "Fifth Third Bank," whose logo is "5/3," are looking for the bank information. This is an example of how hard it is to evaluate personalized recommendation algorithms. However, this limitation can also be an advantage, as raters can be trained to detect spam or other harmful experiences that users may not be able to perceive or detect. It is best to think that your human evaluation provides calibrated labeled data to complement data gathered from real users.

You can use metrics based on human evaluation as additional metrics for evaluating A/B experiments (Huffman 2008). Again, let's use search ranking changes. You can ask raters to rate results from either Control or Treatment for

a given query and aggregate the ratings to see which variant is preferred; or use a side-by-side experiment, where Control and Treatment search results are shown side by side, and raters asked which side is "better." For example, Bing and Google's scaled-out human evaluation programs are fast enough for use alongside the online controlled experiment results to determine whether to launch the change.

Human evaluation results are also useful for debugging: you can examine the results in detail to understand where changes perform well and poorly. In our search query example, results rated a poor match for a query can be examined to help determine why the algorithm returned the result. You can also pair human evaluation with log-based analysis to understand what observed user actions correlate with highly relevant results for a query.

User Experience Research (UER)

While user experience research (UER) uses a variety of methods, we focus here on a subset of field and lab studies that typically go deep with a few users, often by observing them doing tasks of interest and answering questions in either a lab setting or in situ (Alvarez 2017). This type of research is in-depth and intensive typically with at most tens of users, and is useful for generating ideas, spotting problems, and gaining insights from direct observation and timely questions. For example, if your website is trying to sell something, you can observe users trying to complete a purchase, and develop ideas for metrics based on observing where they struggle: Do we observe the purchase taking a long time? Are users struggling and going down a rabbit hole, such as looking for coupon codes?

These type of field and lab studies can include:

- Special equipment to gather data, such as eye-tracking that you cannot gather from your instrumentation
- Diary studies, where users self-document their behavior longitudinally, are useful for gathering data analogous to online instrumentation but augmented with data you cannot gather via instrumentation, such as user intent or offline activities.

These techniques can be useful for generating metric ideas based on correlating "true" user intent with what we observe via instrumentation. You must validate these ideas using methods that scale to more users, such as observational analyses and controlled experiments.

Focus Groups

Focus groups are guided group discussions with recruited users or potential users. You can guide discussion to any range of topics, ranging from open-ended questions about user attitudes, "What is commonly done or discussed amongst their peers," to more specific questions, maybe using screenshots or a demo walk-through to elicit feedback.

Focus groups are more scalable than a UER study and can handle a similar level of ambiguous, open-ended questions that can guide product development and hypotheses. However, given the group nature and discussion format, less ground can be covered than in a UER study, and can fall prey to group-think and convergence on fewer opinions. What customers say in a focus group setting or a survey may not match their true preferences. A well-known example of this phenomenon occurred when Philips Electronics ran a focus group to gain insight into teenagers' preferences for boom box features. The focus group attendees expressed a strong preference for yellow boom boxes during the focus group, characterizing black boom boxes as "conservative." Yet when the attendees exited the room and were given the chance to take home a boom box as a reward for their participation, most chose black (Cross and Dixit 2005).

Focus groups can be useful for getting feedback on ill-formed hypotheses in the early stages of designing changes that become future experiments, or for trying to understand underlying emotional reactions, often for branding or marketing changes. Again, the goal is to gather information that cannot be measured via instrumentation and to get feedback on not-yet-fully-formed changes to help further the design process.

Surveys

To run a survey, you recruit a population to answer a series of questions (Marsden and Wright 2010). The number of questions can vary, as can the type of questions. You can have multiple-choice answers, or open-ended questions where users give a free-form response. These can be done in-person, over the phone, or online directly on your app or site or via other methods of reaching and targeting users (such as Google Surveys (Google 2018)). You can also run surveys from within products, potentially pairing them with controlled experiments. For example, the Windows operating system prompts users with one or two short questions about the operating system and about other Microsoft products; Google has a method to ask a quick question tied to a user's in-product experience and satisfaction (Mueller and Sedley 2014).

While surveys may seem simple, they are actually quite challenging to design and analyze (Marsden and Wright 2010, Groves et al. 2009):

- Questions must be carefully worded, as they may be misinterpreted or unintentionally prime the respondents to give certain answers, or uncalibrated answers. The order of questions may change how respondents answer. And if you want to get data over time, you need to be careful about changes to the survey, as the changes may invalidate comparisons over time.
- Answers are self-reported: Users may not give full or truthful answers, even in anonymous surveys.
- The population can easily be biased and may not be representative of the true user population. This is exacerbated by "response bias," that is, which users respond may be biased (e.g., only people who are unhappy respond). Because of this bias, relative survey results (e.g., time period over time period) may be more useful than absolute results.

These pitfalls suggest that surveys are almost never directly comparable to any results observed from instrumentation. You can use surveys to reach larger numbers of users than UERs or focus groups, but they are primarily useful for getting answers to questions you cannot observe from your instrumented data, such as what happens when a user is offline or a user's opinion or trust and satisfaction levels. Questions might include what other information a user used when making a purchase decision, including offline actions such as talking to a friend, or asking about a user's satisfaction level three months post-purchase.

Surveys are also useful for observing trends over time on less-directly-measurable issues, such as trust or reputation, and are sometimes used to correlate with trends on highly aggregate business metrics, such as overall usage or growth. This correlation can then drive investment in a broad area such as how to improve user trust, but not necessarily generate specific ideas. You can use targeted UER studies for idea generation once you define the broad area.

Depending on the consent of survey participants, you may be able to pair survey results with observational analysis to see which survey responses correlate with observed user behavior, but the bias of the survey respondents will impact the believability and generalizability of the results.

External Data

External data is data relevant to you and what you are looking at that a party external to your company has collected data and analyzed. There are several sources of external data:

- Companies providing per-site granular data (such as, the number of users to a website or detailed information about user online habits) based on data gathered from recruiting a large panel of users who agreed to have all online behavior tracked. One question has been around the representativeness of these users – while they are sampled from clear demographic buckets, there may be other differences in the users who agree to be tracked at this level of detail.
- Companies providing per-user granular data, such as user segments, that can be potentially joined with logs-based data.
- Companies running surveys and questionnaires either to publish themselves or who you can hire to run custom surveys. These companies use a variety of methods to answer questions you might be interested in, such as how many devices users have or their perspective on how trustworthy a brand is.
- Published academic papers. Researchers often publish studies of something of interest. There are a lot of papers out there, for example, papers comparing eye tracking – what the user looked at in a lab, with how they clicked on a search engine (Joachims et al. 2005) give you a good sense of how representative your click data is.
- Companies and websites providing lessons learned, often crowd-sourcing results to validate the lessons. This can be UI design patterns (Linowski 2018b)

External data can help validate simple business metrics if your site or industry appears in one of these lists. For example, if you want to look at total visitors to your site, you can compare your number computed from an internal observational analysis with the numbers provided by comScore or Hitwise, or you could compare the fraction of shopping traffic in each "vertical" category to what you see on your site. Rarely will these numbers exactly match. A better way to do validation is to look at a time series of both internal and external data to see whether the time series aligns in terms of the trend or seasonal variability. You can also provide supporting evidence for your business metrics, either directly measurable quantities or to get ideas for which measurable metrics make good proxies for other harder-to-measure quantities.

Publicly available academic papers, such as those pertaining to User Experience, often establish a general equivalence between different types of metrics. One example compares user-reported satisfaction with a search task to the measured task duration (Russell and Grimes 2007), that gives a good general correlation for satisfaction with duration, though with caveats. This study helped validate a metric, *duration*, that can be computed at scale and correlates with a metric that cannot be computed at scale, user-reported satisfaction.

External data can also add to the hierarchy of evidence. For example, companies could use the published work from Microsoft, Google, and others to establish that latency and performance is important without necessarily needing to run their own online controlled experiments (see Chapter 5). Companies would need to run their own experiments to understand specific tradeoffs for their product, but the general direction and investment could be based on external data for a smaller company without those resources.

External data can also provide competitive studies about how your company compares with your competitors, which can help provide benchmarking on your internal business metrics and give you a sense of what is attainable.

One caveat: because you do not control the sampling or know the exact methods used to do the analysis, the absolute numbers may not always be useful, but trends, correlations, and metric generation and validation are all good use cases.

Putting It All Together

There are many ways to gather data about users, so the question is how to choose which one(s) to use. In large part, this depends on your goal. Do you want to figure out how to measure a particular user experience? Do you want to validate metrics? If you have no idea about what metrics to gather in the first place, more detailed, qualitative, brainstorming type of interactions, such as UER studies or focus groups work well. If you have no way of getting the data, because the interactions aren't on your site, a survey may work well. For validating metrics, external data and observational analyses work well since the data is usually collected over a large enough population that there are fewer sampling biases or other measurement issues.

All these techniques have different tradeoffs. You should consider how many people you are able to collect data from. This affects the generalizability of the results; in other words, whether you can establish external validity. The number of users is often a tradeoff against what type of detail you can get. For example, logs usually have user actions at-scale but not "why" a user acts a particular way that you might get in a UER field study. Where you are in a product cycle may also be a consideration. Early on, when you have too many ideas to test, more qualitative methods such as focus groups and user experience research may make more sense. And as you move towards having quantitative data, then observational studies and experiments make more sense.

Finally, remember that using multiple methods to triangulate towards a more accurate measurement – establishing a hierarchy of evidence – can lead to

more robust results (Grimes, Tang and Russell 2007). Since no method can fully replicate the results from another method, use multiple methods to establish bounds for the answer. For example, to see whether users are happy with your personalized product recommendations, you must define signs of "happiness." To do that, you might observe users in an UER study, see whether they use the personalized recommendations, and ask them questions about whether they found the recommendations useful. Based on that feedback, you can look at the observational data for those users and see what behavioral signals you might see, such as a longer time reading the screen or certain click orders. You can then run a large observational analysis to validate the metric ideas generated from the small-scale UER study, see the interplay with the overall business metrics, and then potentially bolster that with an on-screen survey to reach a larger set of users with simple questions about whether they liked the recommendations. Accompanying this with learning experiments that change the recommendations, will allow you to better understand how user happiness metrics relate to the overall business metrics and improve your OEC.