INSIGHTS · INSIGHT QUALITY AND PITFALLS

Validating Insights Before Acting on Them

A practical guide to checking whether an insight is true before an organisation acts on it, covering data checks, reproduction, robustness tests, triangulation, out-of-sample validation, replication and market tests, and how to match the depth of validation to the stakes of the decision.

ModQu research article · 2,432 words · 23 references (APA 7th edition)

Insights are claims. "Customers who use the app spend more because of it." "Price is the main reason people leave." "Younger buyers want sustainable packaging." Each sounds like knowledge, but each could be wrong: the result of an error in the data, a quirk of one sample, an analysis that tried many options until one worked, or a correlation mistaken for a cause. Organisations that act on unvalidated insights waste money and can do real harm. This article sets out a practical sequence of checks that help establish whether an insight is true, and explains how to decide how much validation a decision needs.

Why validation matters

Science has learned this lesson painfully. The Open Science Collaboration (2015) attempted to replicate 100 published psychology studies and found that, although 97% of the original studies had reported statistically significant results, only 36% of the replications did, and the average effect size in the replications was about half that of the originals. Ioannidis (2005) argued that, when studies are small, effects are modest, many relationships are tested and analyses are flexible, a large share of published findings are likely to be false.

Insight work shares many of these conditions. Data sets are often small or unrepresentative, many variables are explored, analyses are flexible and there is pressure to find something interesting. Simmons et al. (2011) showed how undisclosed choices in data collection and analysis, each defensible on its own, can together make it easy to find a "significant" result where none exists. Gelman and Loken (2014) described how this can happen even without deliberate searching, because the analysis that seems natural depends on the data that happened to be observed.

Business experience points the same way. Kohavi and Thomke (2017) reported that, in large companies that run many online experiments, most ideas for improvements fail to produce the expected benefit when tested, and many have no effect or a negative one. If ideas that seemed promising to experienced people usually fail when tested, insights that have not been tested deserve caution.

A sequence of checks

Validation is best thought of as a sequence of checks, each more demanding than the last.

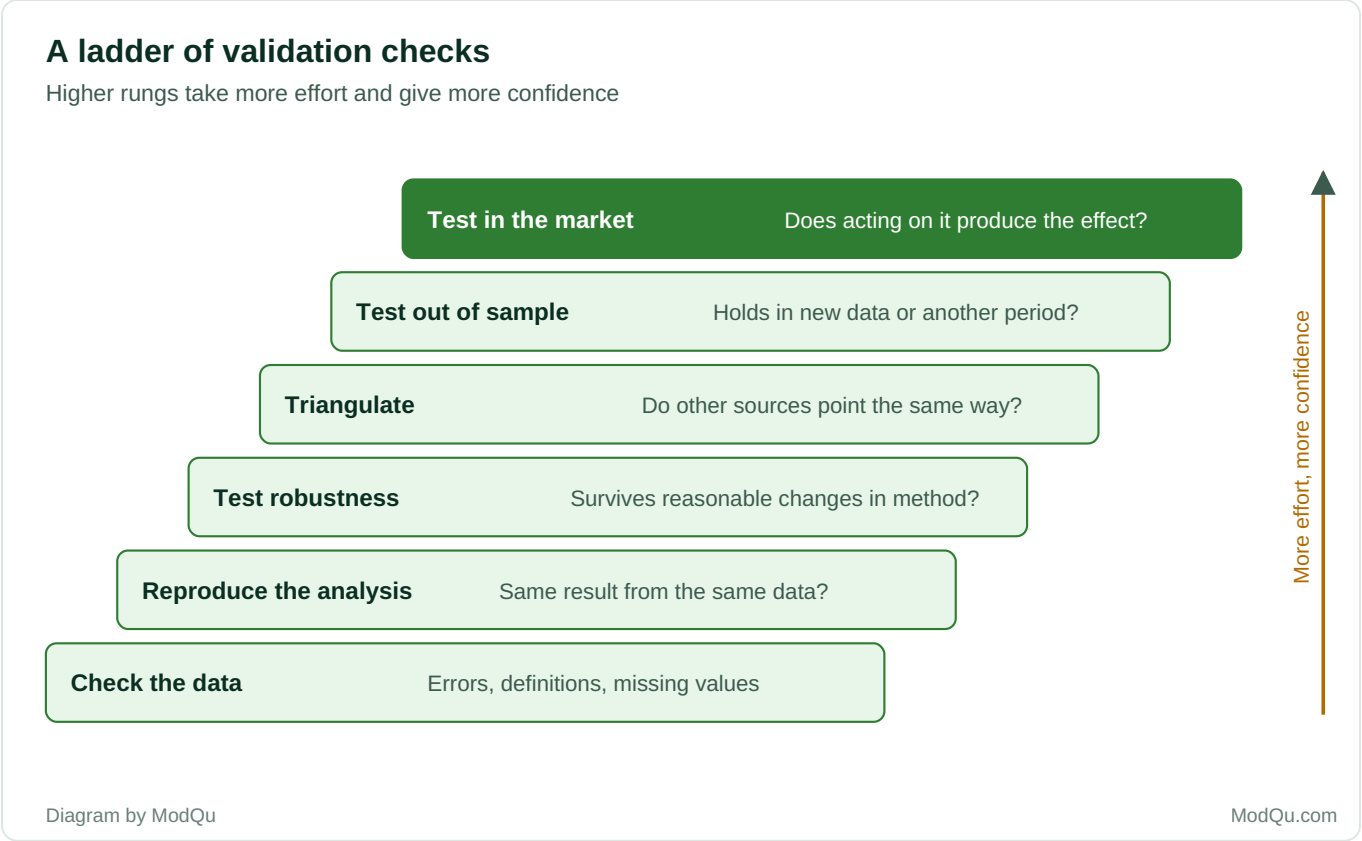


Figure 1. A ladder of validation checks. Higher rungs take more effort but give more confidence. The stakes of the decision determine how far to climb. Diagram by ModQu.

1. Check the data. Many surprising insights turn out to be errors. Check definitions, units, time periods, duplicates, missing values and how the data were collected. A sudden change in a measure may reflect a change in a tracking system rather than in customer behaviour.

2. Reproduce the analysis. Ask a colleague to repeat the analysis from the same data without seeing the original code. If two analysts cannot reach the same result, at least one has made an error or an undocumented choice.

3. Test robustness. Change reasonable analytic choices, such as the time period, the definition of the outcome, the treatment of outliers or the control variables, and see whether the result holds. An insight that appears only under one particular combination of choices is fragile.

4. Triangulate. Look for confirmation in independent sources. Campbell and Fiske (1959) argued that a finding is more credible when different methods of measuring the same thing agree. If survey data suggest that customers value speed of delivery, do behavioural data show that faster options are chosen, and do complaints mention delays?

5. Test out of sample. A pattern found in one data set should be checked in another: a later period, a different region or a holdout sample that was not used in the original analysis. Hastie et al. (2009) explained that models fitted to one sample almost always perform worse on new data, because they capture some of the noise as well as the signal; only performance on new data shows how well they will work in practice.

6. Test in the market. For insights about cause and effect, the most direct validation is an experiment: act on the insight for a randomly selected group and compare the result with a control group. Kohavi et al. (2020) described how organisations can run trustworthy online experiments, and the pitfalls, such as sample ratio mismatches and novelty effects, that can make even experimental results misleading. Thomke (2020) described how a culture of experimentation helps organisations test ideas cheaply before committing to them.

Common threats to validity

Shadish et al. (2002) set out a framework of threats to the validity of causal conclusions, and several are especially common in insight work.

Selection. Groups being compared differ in ways other than the factor of interest. Customers who use a feature may differ from those who do not.

Confounding. A third factor causes both the supposed cause and the effect. Stores with new layouts may also be in growing areas.

Regression to the mean. Units selected for extreme values tend to move toward the average on the next measurement, regardless of any intervention.

History. Something else changed at the same time as the intervention, such as a competitor's price change or a news story.

Leakage. In predictive work, information that would not be available at the time of prediction finds its way into the model. Kaufman et al. (2012) described how such leakage makes models appear far more accurate in testing than they are in use.

Multiple comparisons. When many relationships are tested, some will appear significant by chance. Wasserstein and Lazar (2016), in a statement for the American Statistical Association, warned that p-values alone do not measure the size or importance of an effect, and that conclusions should not rest on whether a single p-value passes a threshold.

The winner's curse

Insights that are selected because they look impressive tend to be less impressive on further testing. When many ideas are tested and the most successful is chosen, its measured effect includes a share of luck, which will not repeat. The same applies to analyses that explore many variables and report the strongest relationship.

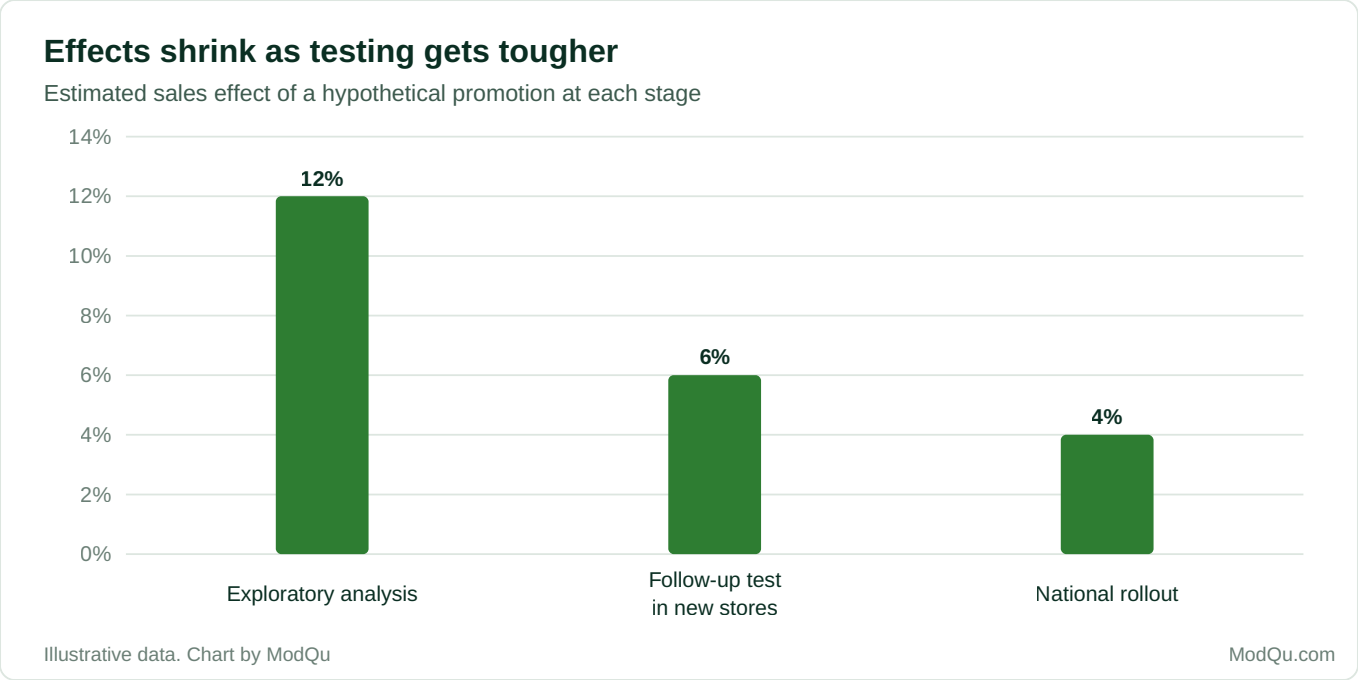


Figure 2. The estimated effect of an insight often shrinks as it is tested more rigorously. Plan for smaller effects than the first analysis suggests. Illustrative data, chart by ModQu.

The practical lesson is to expect shrinkage. An effect estimated in an exploratory analysis or a small pilot should be discounted when forecasting the benefit of acting on it, and confirmed in a test designed specifically to measure it.

Validating measures

Many insights rest on measures, such as a satisfaction index, an engagement score or a segmentation, whose meaning is taken for granted. Messick (1995) argued that validity is not a property of a measure in itself but of the interpretations and uses made of its scores, and that establishing it requires several kinds of evidence: whether the content covers what it should, whether responses reflect the intended processes, whether the structure of the measure fits the theory, whether it relates to other measures as expected, and what consequences follow from using it. For insight work, a practical version of this asks whether a measure means what decision makers think it means, whether it moves when the thing it measures changes, and whether it predicts the outcomes it is supposed to predict. A customer health score that does not predict renewals, for example, may be measuring activity rather than health.

Validating predictive models

Predictive models need validation of their own. Collins et al. (2015) set out reporting standards for prediction models in medicine that are useful well beyond it: describe the data and the outcome clearly, report how the model was developed, and, above all, report its performance on data not used to build it, including how well its predicted probabilities match observed outcomes. A model that ranks customers well but systematically overstates their risk can lead to poor decisions about how many to target. Models should also be monitored after deployment, because their performance tends to decline as customer behaviour and business conditions change.

Validating qualitative insight

Qualitative insights need validation too, though the methods differ. Lincoln and Guba (1985) proposed criteria for judging the trustworthiness of qualitative research, including credibility, which can be strengthened by checking interpretations with participants and by triangulating sources; transferability, which depends on describing the context richly enough for readers to judge whether findings apply elsewhere; and dependability, which depends on a clear record of how the research was done. A qualitative insight that has been checked with participants, found in several sources and examined by a second researcher is far more credible than one drawn from a single analyst's reading.

Validating before big decisions

Validation also applies to the reasoning that turns insights into decisions. Kahneman et al. (2011) proposed a checklist of questions that decision makers can ask about recommendations brought to them, including whether the team has fallen in love with its proposal, whether credible alternatives were considered, and whether the base case is overly optimistic. Lovallo and Kahneman (2003) argued that managers often base forecasts on the specifics of a project rather than on the outcomes of similar projects, and recommended taking an "outside view" by looking at the record of comparable cases. Flyvbjerg (2006) described reference class forecasting, a formal method for taking the outside view, and showed how it reduces the optimism that leads large projects to overrun.

Klein (2007) proposed the premortem, in which a team imagines that a decision has failed and lists the reasons why. Heuer (1999) recommended the analysis of competing hypotheses, which asks analysts to evaluate evidence against all plausible explanations rather than only the favoured one. Both techniques are forms of validation applied to reasoning rather than data.

Building a habit of validation

Validation works best as an organisational habit rather than an occasional exercise. Pfeffer and Sutton (2006) argued that many management practices are adopted on the basis of habit, imitation or belief rather than evidence, and that organisations benefit from routinely asking what evidence supports a proposal. Tetlock and Gardner (2015) found that the most accurate forecasters in a large forecasting tournament shared habits such as breaking problems into parts, seeking out information that might prove them wrong, updating their views in small steps as evidence arrived, and keeping score of their own accuracy. Insight teams can adopt similar habits: recording the predictions implied by their insights, checking later whether they came true, and learning from the cases where they did not.

It also helps to separate the people who generate an insight from those who check it. A colleague who did not do the original analysis, and who is asked explicitly to find reasons it might be wrong, will spot problems that the original analyst cannot see.

Matching validation to the stakes

Not every insight needs every check. Validation takes time and money, and some decisions are cheap to reverse. A useful approach is to match the depth of validation to the stakes.

For small, reversible decisions, such as changing the wording of an email, checking the data and running a quick test may be enough. For moderate decisions, such as changing a product feature or a service process, reproduction, robustness checks and triangulation are worthwhile, followed by a test where possible. For large or irreversible decisions, such as a new pricing strategy, a major investment or a change that affects many customers, all the checks should be applied, including an experiment or a staged rollout with a comparison group.

Nosek et al. (2018) argued for preregistration, stating in advance what will be tested and how. In business, a lighter version is often enough: before a validation test, write down the expected result, the measure and the threshold for acting. This prevents the goalposts from moving after the results are known.

An example

Consider a hypothetical online retailer whose analysts find that customers who watch product videos are twice as likely to buy. The marketing team proposes producing videos for every product.

The insight team applies the checks. The data are sound, and a second analyst reproduces the result. But robustness checks show that the effect shrinks sharply once the analysis accounts for how far customers had already progressed toward buying: people who watch videos are often those already close to a decision. Triangulation with survey data shows that customers find videos useful mainly for complex products. An experiment, in which videos are shown to a random half of visitors for a sample of products, finds a modest increase in purchases for complex products and none for simple ones.

The retailer produces videos for complex products only, at a fraction of the original cost. The insight was partly true, but acting on the first analysis would have wasted most of the budget.

A validation checklist

Before an insight is used to support a significant decision, a short checklist can help. Has the underlying data been checked for errors and changes in definition? Has a second analyst reproduced the main result? Does the result hold under reasonable alternative analyses? Do independent sources point the same way? Has the pattern been confirmed in data not used to find it? If the insight is about cause and effect, has it been tested with a comparison group, or is there a plan to do so as part of the rollout? What is the plausible range of the effect, and has the forecast been discounted for likely shrinkage? And what result, from a test or from the rollout itself, would show that the insight was wrong?

Not every answer needs to be yes. But writing the answers down makes clear how much confidence the insight deserves, and allows decision makers to judge whether that confidence is enough for the decision in front of them. An insight presented with an honest account of how far it has been validated is more useful than one presented with unwarranted certainty, because it allows the organisation to manage the risk of being wrong.

Key takeaways

Insights are claims that may be wrong, and experience in both science and business shows that many plausible findings do not hold up. A sequence of checks, from checking the data and reproducing the analysis to testing robustness, triangulating, testing out of sample and testing in the market, builds confidence step by step. Common threats include selection, confounding, regression to the mean, history, leakage and multiple comparisons. Effects found in exploratory analyses tend to shrink when tested. Qualitative insights and the reasoning behind decisions also need validation. The depth of validation should match the stakes: the larger and less reversible the decision, the further up the ladder an insight should climb before the organisation acts on it.

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine, 162*(1), 55–63. https://doi.org/10.7326/M14-0697

Flyvbjerg, B. (2006). From Nobel prize to project management: Getting risks right. *Project Management Journal, 37*(3), 5–15. https://doi.org/10.1177/875697280603700302

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

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

Kahneman, D., Lovallo, D., & Sibony, O. (2011). Before you make that big decision... *Harvard Business Review, 89*(6), 50–60.

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), Article 15. https://doi.org/10.1145/2382577.2382579

Klein, G. (2007). Performing a project premortem. *Harvard Business Review, 85*(9), 18–19.

Kohavi, R., & Thomke, S. (2017). The surprising power of online experiments. *Harvard Business Review, 95*(5), 74–82.

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press. https://doi.org/10.1017/9781108653985

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.

Lovallo, D., & Kahneman, D. (2003). Delusions of success: How optimism undermines executives' decisions. *Harvard Business Review, 81*(7), 56–63.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

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

Pfeffer, J., & Sutton, R. I. (2006). Evidence-based management. *Harvard Business Review, 84*(1), 62–74.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown.

Thomke, S. H. (2020). *Experimentation works: The surprising power of business experiments*. Harvard Business Review Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108