RESEARCH · QUANTITATIVE AND MIXED METHODS

Internal and External Validity in Research Design

What internal and external validity mean in research design, the main threats to each, how designs protect against them, the trade-off between the two, and how to judge whether findings apply to a new setting.

ModQu research article · 1,356 words · 14 references (APA 7th edition)

A study can go wrong in two broad ways. Its conclusion about cause and effect may be mistaken, because something other than the supposed cause produced the result. Or its conclusion may be correct for the people and setting studied but not hold anywhere else. The first problem concerns **internal validity**; the second concerns **external validity**. Campbell (1957) introduced the distinction, and it remains central to how research designs are planned and judged. This article explains both concepts, the main threats to each and how researchers balance them.

Internal validity

Internal validity is the extent to which a study supports the conclusion that a relationship between two variables is causal. Shadish et al. (2002) set out three conditions for a causal claim: the cause must come before the effect, the cause and effect must be related, and other plausible explanations must be ruled out. The third condition is the hardest. Most threats to internal validity are alternative explanations that could produce the same pattern of results.

Campbell and Stanley (1963) listed the classic threats, which Cook and Campbell (1979) and later Shadish et al. (2002) refined. The table below summarises the most common.

Threats to internal validity

Alternative explanations a design must rule out

Threat	Example	Remedy
Selection	Volunteers for training were already more motivated	Random assignment; matching
History	A competitor's price rise during the study	Control group measured at the same time
Maturation	Staff improve with experience anyway	Control group; several baseline measures
Testing	A pre-test teaches people the answers	Groups without a pre-test
Instrumentation	Survey wording changes between waves	Keep measures constant
Regression to the mean	Worst-performing stores improve on their own	Compare with similar untreated stores
Attrition	Dissatisfied users drop out before follow-up	Track and analyse who leaves

Based on Campbell & Stanley (1963); Shadish et al. (2002) ModQu.com

Figure 1. Common threats to internal validity, with examples and remedies. Based on Campbell and Stanley (1963) and Shadish et al. (2002); table by ModQu.

Selection is the most serious threat in many studies. If people who receive a treatment differ from those who do not, any difference in outcomes may reflect the initial differences rather than the treatment. Customers who choose a premium service may already be more loyal; employees who volunteer for training may already be more motivated.

History refers to events other than the treatment that occur during the study. **Maturation** refers to natural changes over time, such as growing experience. **Testing** effects arise when taking a measure affects later scores on it. **Instrumentation** threats arise when the measure itself changes, for example when survey questions are reworded between waves. **Regression to the mean** occurs when groups are selected for extreme scores, which tend to move toward the average on retesting regardless of any treatment. **Attrition** is the loss of participants, which can bias results if those who leave differ from those who stay.

Some threats come from the research situation itself. Orne (1962) described **demand characteristics**: cues in an experiment that suggest to participants what is expected, leading them to behave accordingly. A result produced by participants trying to help the researcher is not evidence of the effect being studied.

Protecting internal validity

Research designs differ in how well they rule out these threats. **Random assignment** is the strongest protection against selection, because it makes groups equivalent on average in every respect, measured or not. A **control group** that does not receive the treatment but experiences the same history, maturation and testing allows these effects to be separated from the treatment effect.

When randomisation is impossible, **quasi-experimental designs** use other strategies: comparing groups that are as similar as possible, measuring outcomes several times before and after a change, or exploiting natural cut-offs that determine who receives a treatment. Cook and Campbell (1979) described many such designs and the threats each leaves open. The general principle is to identify the plausible alternative explanations and design the study so that each can be ruled out or measured.

External validity

External validity is the extent to which a causal relationship holds across variations in persons, settings, treatments and outcomes. Cronbach (1982) emphasised that the question of interest is usually not whether a result generalises everywhere, but whether it applies to a particular target: this population, this organisation, this version of the programme.

External validity can fail in several ways:

- **People.** Participants may differ from the target population. Henrich et al. (2010) showed that most published psychology research used participants from Western, educated, industrialised, rich and democratic societies, who are unusual on many psychological measures, so findings may not generalise to most of the world's people.
- **Settings.** An effect found in a laboratory or a single company may not appear in other environments. Levitt and List (2007) argued that laboratory experiments on generosity and fairness overstate how much these preferences matter in real markets, because participants know they are observed and the stakes are small.
- **Treatments.** A programme delivered by enthusiastic researchers may work differently when scaled up by routine staff.
- **Outcomes.** A treatment may improve a short-term measure but not the outcome that matters in the long run.

Rothwell (2005) showed that randomised trials in medicine often exclude many of the patients who will later receive the treatment, such as older people or those with several conditions, and that reports seldom give enough detail for clinicians to judge whether the results apply to their own patients.

The trade-off

Internal and external validity often pull in opposite directions. Tight control, homogeneous participants and standardised procedures strengthen internal validity but make the study less like the settings where findings will be used. Field studies with diverse participants are closer to real conditions but harder to control.

Mook (1983) argued that external validity is not always the goal. Some studies aim to test whether something can happen, or whether a theory's prediction holds under specified conditions. For such studies, artificial conditions are not a weakness. The question to ask is what the researcher is trying to conclude, and whether the design supports that conclusion.

Judging whether findings apply elsewhere

Deaton and Cartwright (2018) cautioned that a well-conducted randomised trial shows that a treatment worked on average in the study population, but not that it will work in a new setting. Cartwright and Hardie (2012) argued that to predict whether a policy will work in a new place, decision-makers need to know why it worked where it was tested, and whether the necessary supporting factors are present in the new setting. A school programme that worked because of small class sizes will not work where classes are large.

Several approaches help. Findley et al. (2021) reviewed strategies for assessing external validity, including testing the same intervention across several sites, comparing the characteristics of study samples and target populations, and specifying the mechanisms through which effects operate. Pearl and Bareinboim (2014) developed formal methods, called **transportability**, for deciding when and how results from one population can be transferred to another, given knowledge of how the two differ.

Two further kinds of validity

Shadish et al. (2002) placed internal and external validity within a larger scheme of four. **Statistical conclusion validity** concerns whether the statistical analysis correctly shows that cause and effect are related, and is threatened by small samples, unreliable measures and running many tests. **Construct validity** concerns whether the treatments, measures and samples in a study represent the concepts named in its conclusions. A study can have strong internal validity and still mislabel what it found: an experiment may show that a training session improved sales, but if the improvement came from the attention participants received rather than the content of the training, the conclusion about "training" is mistaken. All four kinds of validity need attention for a conclusion to be trusted.

An applied example

A hypothetical example: a software company tests a new onboarding email with randomly selected new users and finds that it increases activation. Random assignment gives strong internal validity. But the test ran during a quiet month, among users who signed up through the website rather than through partners, and measured activation after one week. Before rolling the email out to all users, the team asks whether partner sign-ups behave similarly, whether the effect lasts beyond a week, and whether results hold in busy periods. Running the test again with partner users and a longer follow-up addresses the main external validity questions.

Conclusion

Internal validity asks whether a study's causal conclusion is sound; external validity asks whether it holds beyond the study. Designs that strengthen one can weaken the other. Good research identifies the threats that matter most for its purpose, chooses a design that addresses them, and states clearly the people, settings and conditions to which its conclusions apply.

References

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Findley, M. G., Kikuta, K., & Denly, M. (2021). External validity. *Annual Review of Political Science, 24*, 365–393. https://doi.org/10.1146/annurev-polisci-041719-102556

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*(2), 153–174. https://doi.org/10.1257/jep.21.2.153

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379–387. https://doi.org/10.1037/0003-066X.38.4.379

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*(11), 776–783. https://doi.org/10.1037/h0043424

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579–595. https://doi.org/10.1214/14-STS486

Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet*, *365*(9453), 82–93. https://doi.org/10.1016/S0140-6736(04)17670-8

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