UNIVERSITY OF BERGEN

Experimental design

Sebastian Jentschke

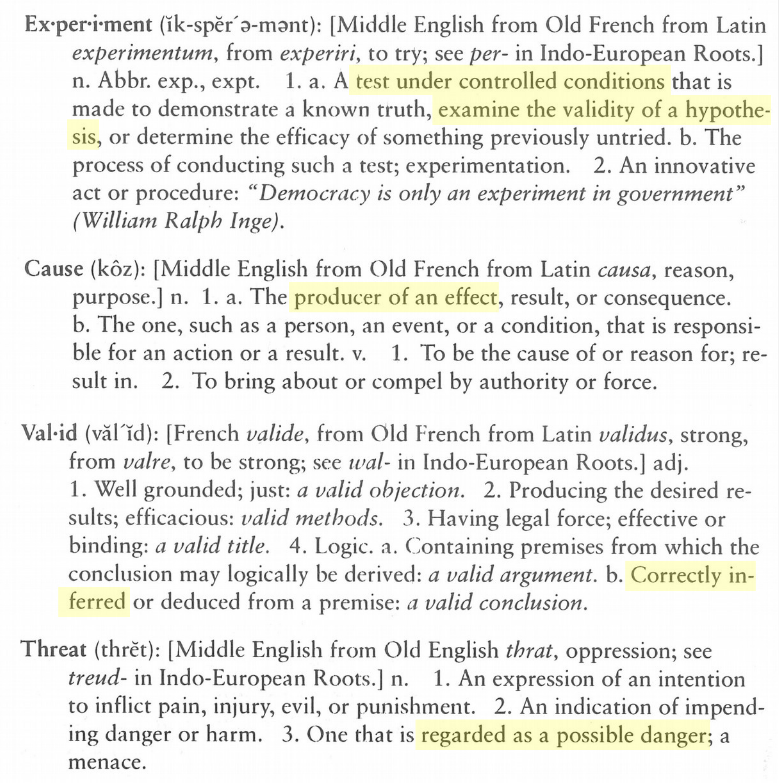
UNIVERSITY OF BERGEN

Agenda

- experiments and causal inference
- validity
- internal validity
- statistical conclusion validity
- construct validity
- external validity
- tradeoffs and priorities

Experiments and causal inference

Some definitions



Ex·per·i·ment (ĭk-spĕr′ə-mənt): [Middle English from Old French from Latin *experimentum*, from *experiri*, to try; see *per-* in Indo-European Roots.] n. Abbr. exp., expt. 1. a. A test under controlled conditions that is made to demonstrate a known truth, examine the validity of a hypothesis, or determine the efficacy of something previously untried. b. The process of conducting such a test; experimentation. 2. An innovative act or procedure: *"Democracy is only an experiment in government" (William Ralph Inge).*

Cause (kôz): [Middle English from Old French from Latin *causa*, reason, purpose.] n. 1. a. The producer of an effect, result, or consequence. b. The one, such as a person, an event, or a condition, that is responsible for an action or a result. v. 1. To be the cause of or reason for; result in. 2. To bring about or compel by authority or force.

Val·id (văl′ĭd): [French *valide*, from Old French from Latin *validus*, strong, from *valre*, to be strong; see *wal-* in Indo-European Roots.] adj. 1. Well grounded; just: *a valid objection.* 2. Producing the desired results; efficacious: *valid methods.* 3. Having legal force; effective or binding: *a valid title.* 4. Logic. a. Containing premises from which the conclusion may logically be derived: *a valid argument.* b. Correctly inferred or deduced from a premise: *a valid conclusion.*

Threat (thrĕt): [Middle English from Old English *thrat*, oppression; see *treud-* in Indo-European Roots.] n. 1. An expression of an intention to inflict pain, injury, evil, or punishment. 2. An indication of impending danger or harm. 3. One that is regarded as a possible danger; a menace.

a hypothesis often concerns a cause-effect-relationship

Scientific revolution

- discovery of America → french revolution: renaissance and enlightenment – Copernicus, Galilei, Newton
- **empiricism**: use observation to correct errors in theory
- **scientific experimentation**:
  taking a deliberate action [manipulation, vary something] followed by systematic observation of what occured afterwards [effect] controlling extraneous influences that might limit or bias observation: random assignment, control groups
- mathematization, institutionalization

Causal relationships

Definitions and some philosophy:

- causal relationships are recognized intuitively by most people in their daily lives
- Locke: „A **cause** is which makes any other thing, either simple idea, substance or mode, begin to be; and an **effect** is that, which had its beginning from some other thing“ (1975, p. 325)
- Stuart Mill: A causal relationship exists if (a) the **cause preceded the effect**, (b) the **cause was related to the effect**, (c) we can find **no plausible alternative explanantion** for the effect other than the cause.
  Experiments: (a) manipulate the presumed cause, (b) assess whether variation in the cause is related to variation in the effect, (c) use various methods to reduce the plausibility of other explanations for the effect.
  Non-experimental methods (e.g., correlation analyses) have weaknesses with (a) unclear which variable came first, and (c) can‘t rule out alternative explanations (third moderating variable) and can‘t provide evidence for causation
- Popper: regarding (c) – falsificionist logic: confirmation is often difficult (because we might not observe all instances) → one disconfirming instance is sufficient to falsify the hypothesis / conclusion → prove – provide evidence

Causal relationships

- cause: constellation - many factors are usually required and we rarely know all of them and how they relate (e.g., psychotherapy)
- inus condition (an insufficient but non-redundant part of an unnecessary but sufficient condition)
  insufficient: a match can not start a fire → adding a non-redundant part: fire-promoting factors (oxygen, dry leaves) → unnecessary: there might be other sets of conditions → sufficient condition to start a fire
- causes must be manipulable to be used in experiments – non-manipulable causes can still be studied and provide evidence (observe an effect and search for its cause)

Causal relationships

- effect: Hume – counterfactual model
  experiment: we observe what did happen after people got treatment, but we don‘t know what would have happened (counterfactual) if they had not received treatment
  effect: difference between what did happen and what would have happened – but: can not observe the counterfactual and need a reasonable approximation (e.g., treatment and control group)
- two central tasks of experimental design: (a) creating a high-quality (but necessarily imperfect) source of counterfactual inference; (b) understand how this source differs from the treatment condition

Causal relationships

- experiments can provide a **causal description** – describing the consequences attributable to deliberately varying a treatment – but are **less suited** to provide a **causal explanantion** – clarifying mechanisms through which and the conditions under which a causal relationship holds
- analogy to molar (as a whole) and molecular (decomposed into parts) causation: causal description = describe bivariate relationship between molar treatment and molar outcome; causal explanation = breaking molar causes into molecular parts to determine what causes the change (drug vs. placebo: decomposing medication effects and verbal interaction / social support)
- no clear dichotomy between causal description and explanation
- causal explanation is not always required for practical solutions

Components of experiments

- control of treatment → manipulating (one or more) independent variable to observe the effect on (one or more) dependant variable; caveat: observations / measurements are not theory-neutral (what is measured and how is influenced by, e.g., the researchers theoretical assumptions, available measures, etc.)
- experiment: randomized assignment to the experimental units → create two groups that are probabilistic similar to each other → outcome differences are likely due to the treatment not to already exisiting group differences
- quasi-experiments: share most features of an experiment (e.g., control group, pretest) but lack random assignment – cause is manipulable and occurs before the effect, but less compelling support for counterfactual inference (control group may differ even though many alternative explanations are controlled for; solution: assess pre- and post-test scores and assess whether they vary in commonality with the hypothesized cause)
- natural experiments: naturally-occuring difference between treatment and comparison
- non-experimental designs: correlational / passive observational design → identify presumed cause and effect without structural features of experiments (randomization, control group)

Experiments and generalizability

- experiments: strength – illuminate causal inference vs. weakness – how far generalizes that causal relationship
- **highly localized and particularistic**: restricted range of settings, theory-laden measures, convenient samples, conducted at a particular point in time **vs. derived theories**: abstract constructs with broad conceptual applicability, population
  construct validity: how well does the research operation represent the underlying theoretical / abstract construct?
- Cronbach (1982): decomposing experiments into units / persons, treatments, observations / outcomes and settings (UTOS)
  external validity: does the causal relationship hold over variations in persons, treatments, observations and settings?
- random selection as solution? persons (but requires clearly delineated population and opportunity to sample from these – but self-selection); treatments (conflicts with „optimal" treatment), outcomes (multi-method), settings (prototypical vs. heterogeneous instances)

Validity

Definition

- (approximate) truth of an inference (i.e., a property of the inference not design, methods, etc.)
→ judgement about the extent to which the empirical evidence supports this inference
→ always an approximation: no method guarantees the validity of an inference
- philosophical theories of truth:
(1) correspondence: a claim is true if it corresponds to the world → gathering data to assess how well knowledge claims match the world
(2) coherence: a claim is true if it belongs to a coherent set of claims → must cohere with exisiting knowledge, scepticism if new contradicts established knowledge
(3) pragmatism: a claim is true if it is useful to believe it → assigns meaning or permits predictability; convince others to use it
- correspondence: empirical evidence → abstract inference
- various degrees and types / aspects of validity: use of a method may affect more than one type of validity simultaneously (e.g., internal vs. external validity) → we may not anticipate all consequences

Validity typology

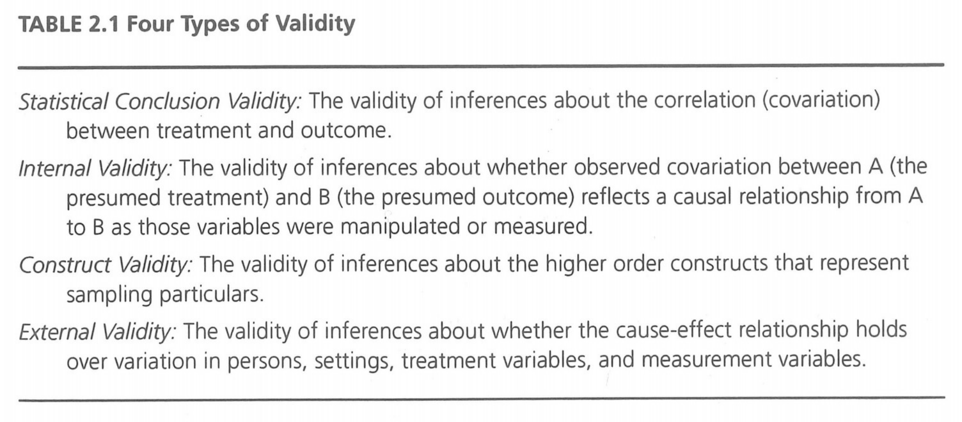


TABLE 2.1 Four Types of Validity

Statistical Conclusion Validity: The validity of inferences about the correlation (covariation) between treatment and outcome.

Internal Validity: The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.

Construct Validity: The validity of inferences about the higher order constructs that represent sampling particulars.

External Validity: The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

1) How reliable and large is the covariation between presumed cause and effect?
2) Is the covariation causal or would the same covariation have been obtained without or with another treatment?
3) How well reflect the persons, treatments, observations and settings the underlying general constructs?
4) How generalizable is the locally embedded causal relationship over varied persons, treatments, observations and settings?

Threats to validity

- threats can be identified conceptually or empirically – empirically-based threats change over time and the likelihood of occurence varies across contexts
- list of validity threats have a heuristic function: help anticipating likely criticism of the inferences
  → minimize amount and plausibility of occurence
  (1) design controls (e.g., randomization)
  (2) statistical controls
- explore role and influence of threats:
  (1) How would the thread apply?
  (2) Is the threat plausible to occur (not just possible)?
  (3) Does it operate in the same direction as the observed effect (confound)?

Internal validity

Definition

- does the observed covariation (between cause and effect) reflect a causal relationship?
  (1) cause must precede effect, (2) cause and effect must covary, (3) there is no plausible alternative explanation for the relationship
- internal validity as local molar causal validity: local (limited to particular treatments, outcomes, settings and persons), molar (treatment as a complex package, e.g. psychotherapy)

Threats

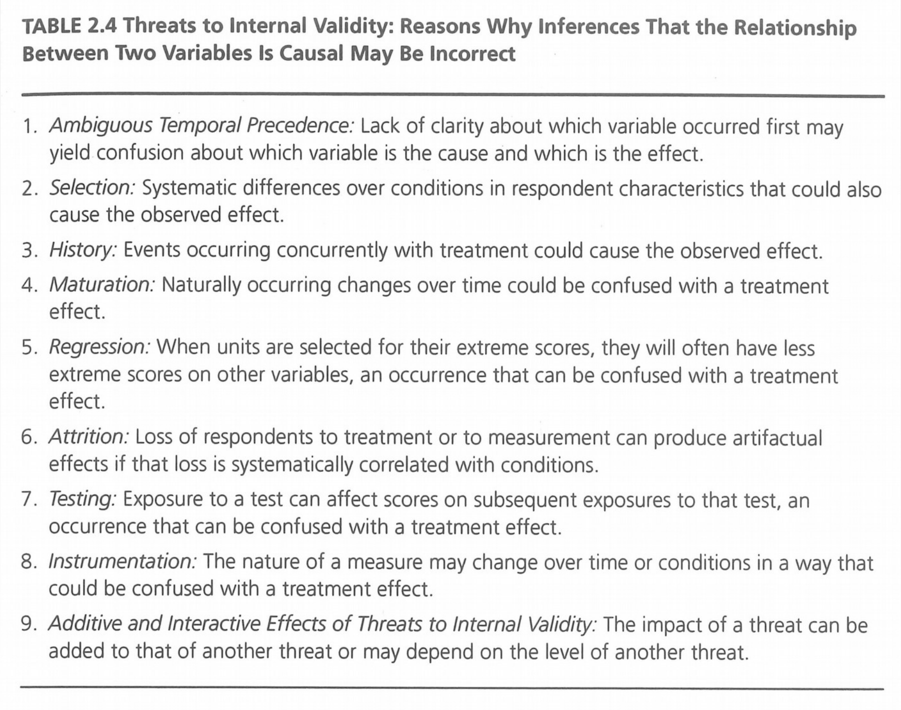


TABLE 2.4 Threats to Internal Validity: Reasons Why Inferences That the Relationship Between Two Variables Is Causal May Be Incorrect

1. *Ambiguous Temporal Precedence:* Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2. *Selection:* Systematic differences over conditions in respondent characteristics that could also cause the observed effect.
3. *History:* Events occurring concurrently with treatment could cause the observed effect.
4. *Maturation:* Naturally occurring changes over time could be confused with a treatment effect.
5. *Regression:* When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.
6. *Attrition:* Loss of respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with conditions.
7. *Testing:* Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect.
8. *Instrumentation:* The nature of a measure may change over time or conditions in a way that could be confused with a treatment effect.
9. *Additive and Interactive Effects of Threats to Internal Validity:* The impact of a threat can be added to that of another threat or may depend on the level of another threat.

- generally randomization works well (except for differential attrition by treatment group or due to that different testing procedures are required by the treatment groups)
- indentifying and quantifying possible threats (and statistically controlling for them)

Statistical conclusion validity

Definition

- statistical inferences about the covariation component (between presumed cause and effect) of causal inferences:
  (1) whether (significance)
  (2) how strongly (effect size)
- null hypothesis significance testing: is the group difference (between treatment and control) large enough to assume it did not occur by chance? → does not mean that cause and effect do not covary if non-significant (to close to call)
- type-I-error (α): falsely inferring the existence of an effect (H0 is true)
  type-II-error (β): falsely inferring the absence of an existing effect (H1 is true)
- effect size: difference between the conditions relative to the standard deviation
- there is a relation between statistical power / effect size (1 – β), sample size and α

Threats

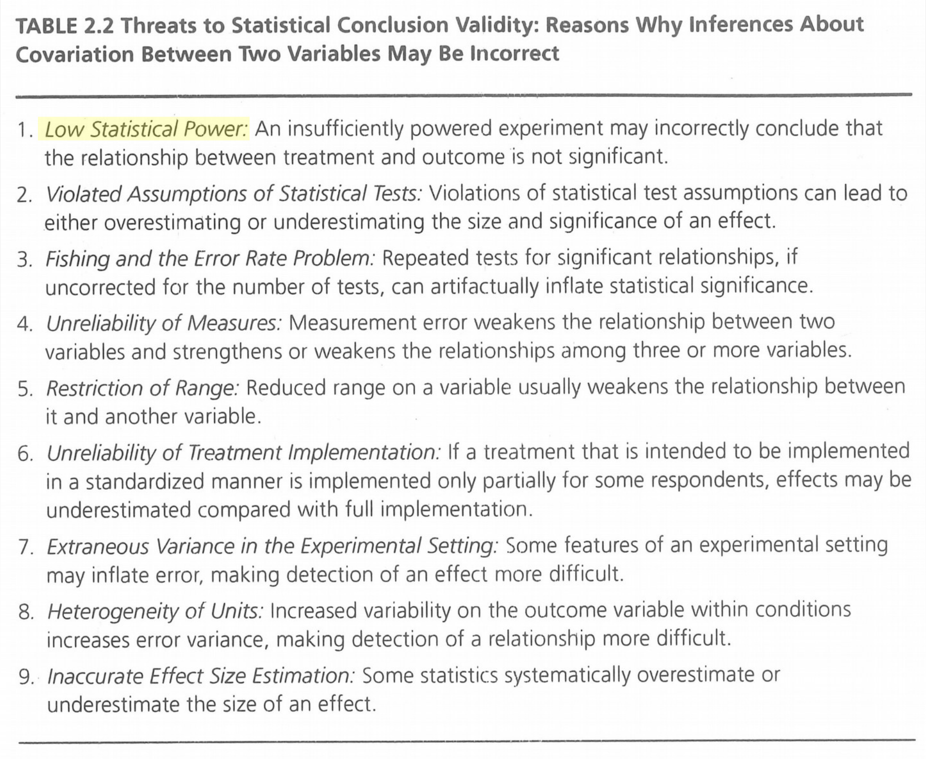


TABLE 2.2 Threats to Statistical Conclusion Validity: Reasons Why Inferences About Covariation Between Two Variables May Be Incorrect

1. *Low Statistical Power:* An insufficiently powered experiment may incorrectly conclude that the relationship between treatment and outcome is not significant.
2. *Violated Assumptions of Statistical Tests:* Violations of statistical test assumptions can lead to either overestimating or underestimating the size and significance of an effect.
3. *Fishing and the Error Rate Problem:* Repeated tests for significant relationships, if uncorrected for the number of tests, can artifactually inflate statistical significance.
4. *Unreliability of Measures:* Measurement error weakens the relationship between two variables and strengthens or weakens the relationships among three or more variables.
5. *Restriction of Range:* Reduced range on a variable usually weakens the relationship between it and another variable.
6. *Unreliability of Treatment Implementation:* If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation.
7. *Extraneous Variance in the Experimental Setting:* Some features of an experimental setting may inflate error, making detection of an effect more difficult.
8. *Heterogeneity of Units:* Increased variability on the outcome variable within conditions increases error variance, making detection of a relationship more difficult.
9. *Inaccurate Effect Size Estimation:* Some statistics systematically overestimate or underestimate the size of an effect.

Threats

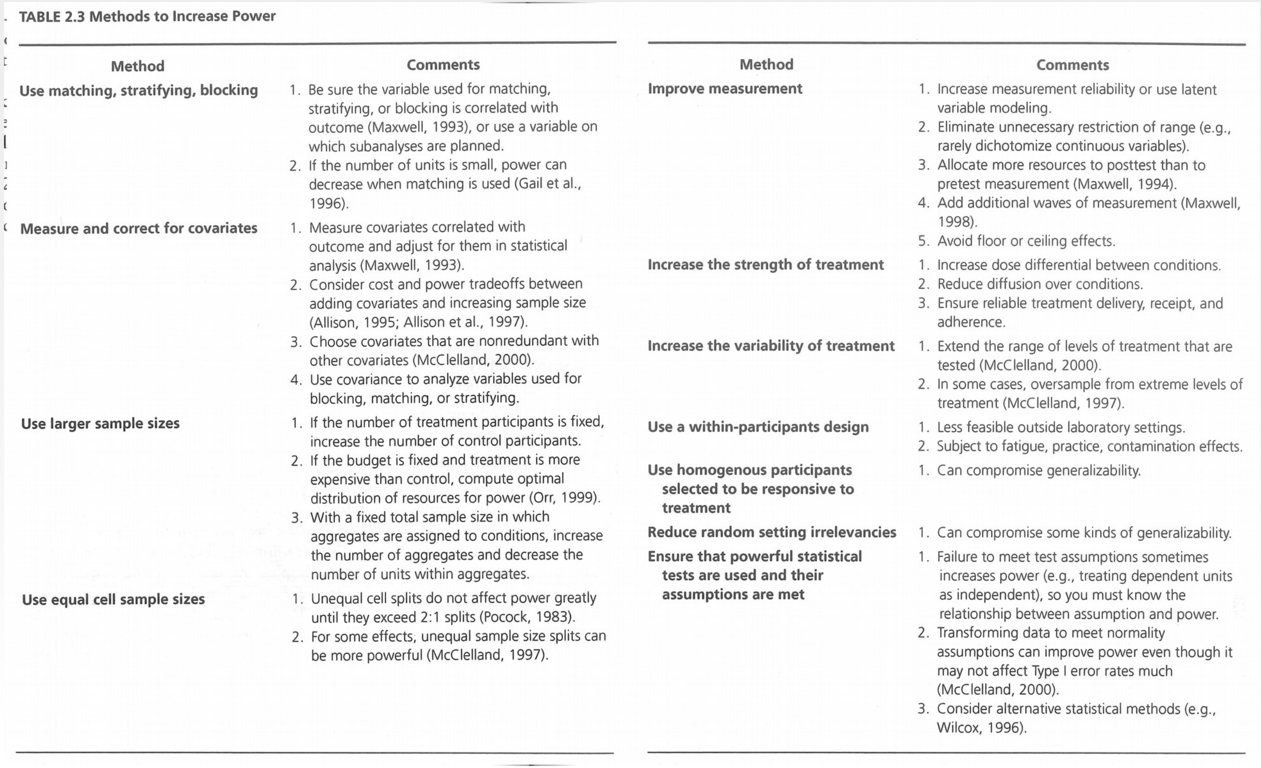


TABLE 2.3 Methods to Increase Power

Method	Comments
Use matching, stratifying, blocking	1. Be sure the variable used for matching, stratifying, or blocking is correlated with outcome (Maxwell, 1993), or use a variable on which subanalyses are planned. 2. If the number of units is small, power can decrease when matching is used (Gail et al., 1996).
Measure and correct for covariates	1. Measure covariates correlated with outcome and adjust for them in statistical analysis (Maxwell, 1993). 2. Consider cost and power tradeoffs between adding covariates and increasing sample size (Allison, 1995; Allison et al., 1997). 3. Choose covariates that are nonredundant with other covariates (McClelland, 2000). 4. Use covariance to analyze variables used for blocking, matching, or stratifying.
Use larger sample sizes	1. If the number of treatment participants is fixed, increase the number of control participants. 2. If the budget is fixed and treatment is more expensive than control, compute optimal distribution of resources for power (Orr, 1999). 3. With a fixed total sample size in which aggregates are assigned to conditions, increase the number of aggregates and decrease the number of units within aggregates.
Use equal cell sample sizes	1. Unequal cell splits do not affect power greatly until they exceed 2:1 splits (Pocock, 1983). 2. For some effects, unequal sample size splits can be more powerful (McClelland, 1997).
Improve measurement	1. Increase measurement reliability or use latent variable modeling. 2. Eliminate unnecessary restriction of range (e.g., rarely dichotomize continuous variables). 3. Allocate more resources to posttest than to pretest measurement (Maxwell, 1994). 4. Add additional waves of measurement (Maxwell, 1998). 5. Avoid floor or ceiling effects.
Increase the strength of treatment	1. Increase dose differential between conditions. 2. Reduce diffusion over conditions. 3. Ensure reliable treatment delivery, receipt, and adherence.
Increase the variability of treatment	1. Extend the range of levels of treatment that are tested (McClelland, 2000). 2. In some cases, oversample from extreme levels of treatment (McClelland, 1997).
Use a within-participants design	1. Less feasible outside laboratory settings. 2. Subject to fatigue, practice, contamination effects.
Use homogenous participants selected to be responsive to treatment	1. Can compromise generalizability.
Reduce random setting irrelevancies	1. Can compromise some kinds of generalizability.
Ensure that powerful statistical tests are used and their assumptions are met	1. Failure to meet test assumptions sometimes increases power (e.g., treating dependent units as independent), so you must know the relationship between assumption and power. 2. Transforming data to meet normality assumptions can improve power even though it may not affect Type I error rates much (McClelland, 2000). 3. Consider alternative statistical methods (e.g., Wilcox, 1996).

Threats

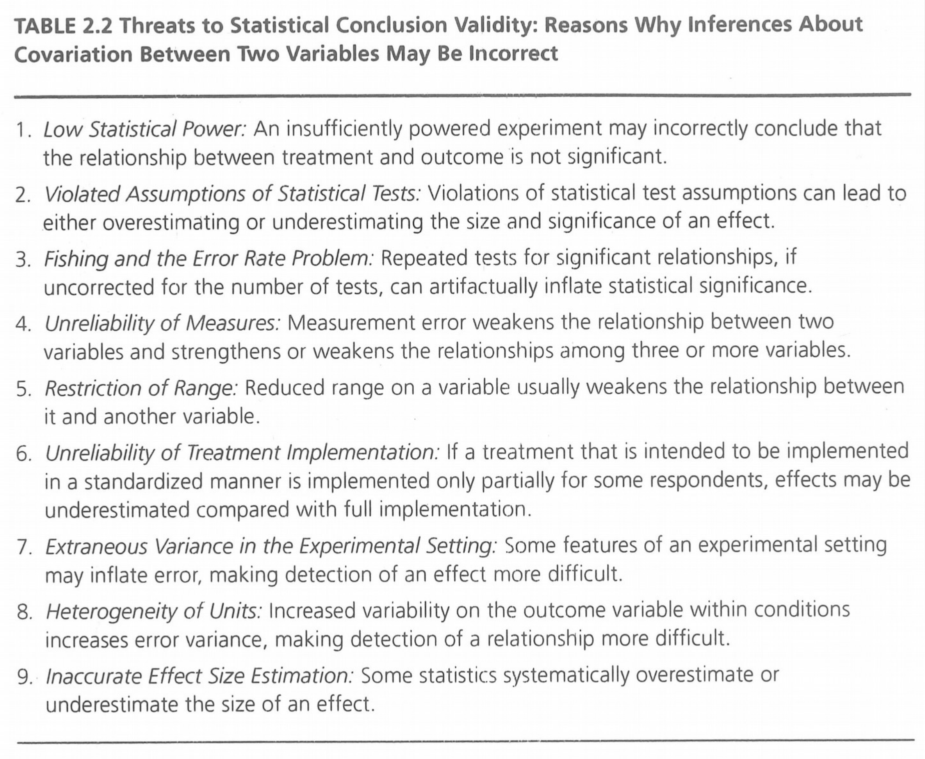


TABLE 2.2 Threats to Statistical Conclusion Validity: Reasons Why Inferences About Covariation Between Two Variables May Be Incorrect

1. *Low Statistical Power:* An insufficiently powered experiment may incorrectly conclude that the relationship between treatment and outcome is not significant.
2. *Violated Assumptions of Statistical Tests:* Violations of statistical test assumptions can lead to either overestimating or underestimating the size and significance of an effect.
3. *Fishing and the Error Rate Problem:* Repeated tests for significant relationships, if uncorrected for the number of tests, can artifactually inflate statistical significance.
4. *Unreliability of Measures:* Measurement error weakens the relationship between two variables and strengthens or weakens the relationships among three or more variables.
5. *Restriction of Range:* Reduced range on a variable usually weakens the relationship between it and another variable.
6. *Unreliability of Treatment Implementation:* If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation.
7. *Extraneous Variance in the Experimental Setting:* Some features of an experimental setting may inflate error, making detection of an effect more difficult.
8. *Heterogeneity of Units:* Increased variability on the outcome variable within conditions increases error variance, making detection of a relationship more difficult.
9. *Inaccurate Effect Size Estimation:* Some statistics systematically overestimate or underestimate the size of an effect.

(2) observations are not independent (same class→ background, SES) → errors are not independently distributed

(3) α-inflation with repeated testing (3: 0.143; 20: 0.642; 50: 0.923) → Bonferroni-correction (but: overlooking small effects)

(4) unreliability attenuates bivariate relationship (multivariate relations are less predictable) → increase the SNR by increasing the number of measurements (more items, more raters), improving the quality of measures (better items, better instruction / training of raters) or statistical techniques (e.g., latent variable modelling)

(5) range: avoid floor or ceiling effects

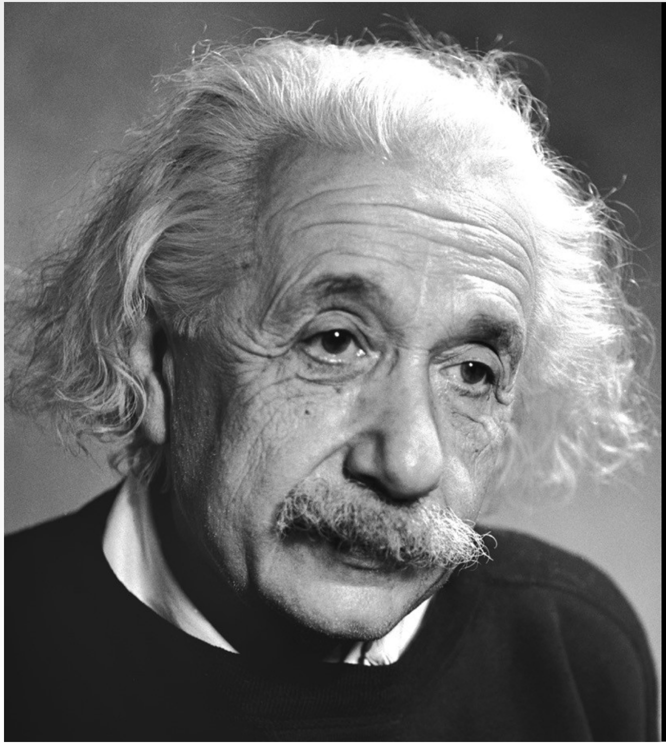
(6) standardization of instructions and the implementation of an experiment / intervention

(7) reduce extraneous variance (distracting noises, temperature fluctuations, circadian rhythm)

(8) homogenize sample (but: may reduce external validity)

Construct validity

Definition



„Thinking without the positing of categories and concepts in general would be as impossible as breathing in a vacuuum.“

(Einstein, 1949, p. 673-674)

Definition

- constructs are central means for **connecting** the **operations** used **in an experiment to pertinent theory** and language

  construct labels carry social, political and economic implications (shape perceptions, frame debates, and elicit support and criticism)

  creation and defense of constructs is a fundamental task of all science
- construct validity is forstered by: (1) clear explication of the person, settings, treatments and outcome constructs of interest; (2) carefully selecting instances to match those constructs; (3) assessing the match between the instances and constructs; (4) revising construct descriptions accordingly → continous process

Threats

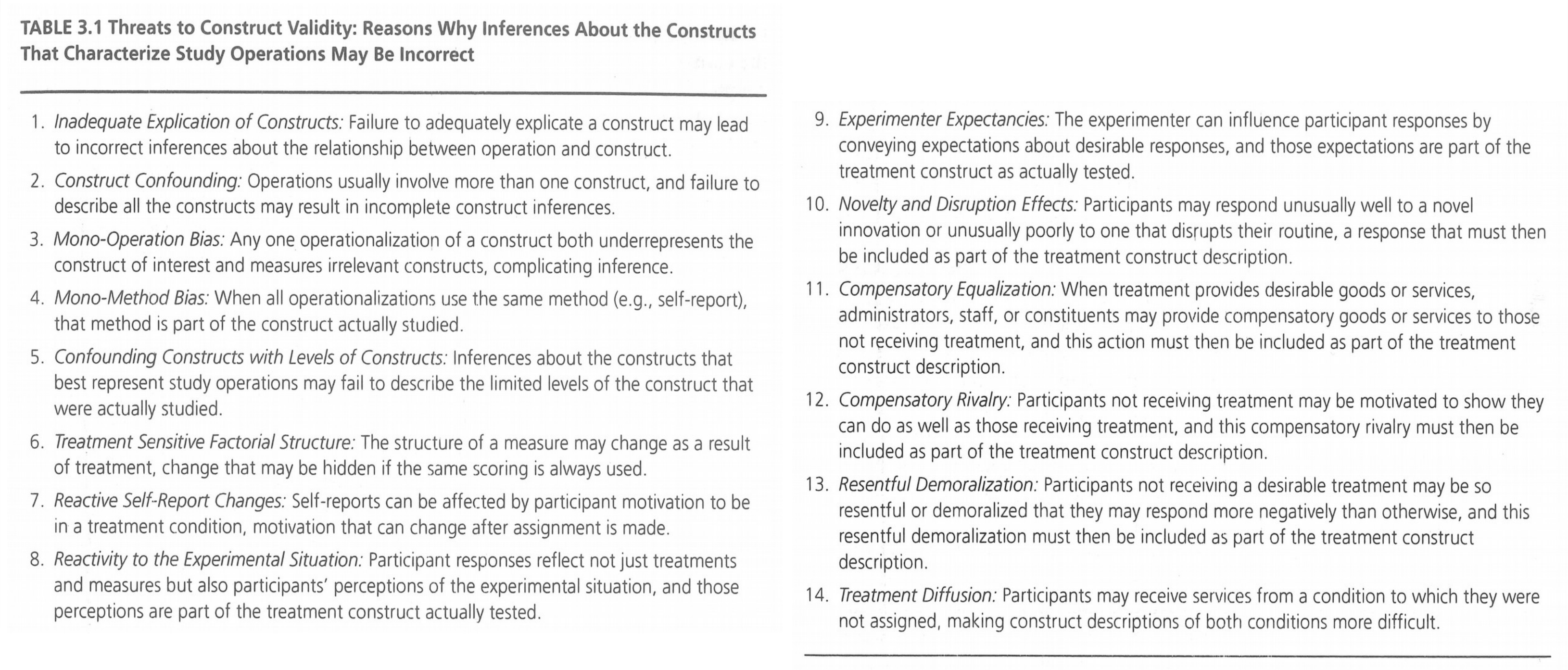


TABLE 3.1 Threats to Construct Validity: Reasons Why Inferences About the Constructs That Characterize Study Operations May Be Incorrect

1. *Inadequate Explication of Constructs:* Failure to adequately explicate a construct may lead to incorrect inferences about the relationship between operation and construct.
2. *Construct Confounding:* Operations usually involve more than one construct, and failure to describe all the constructs may result in incomplete construct inferences.
3. *Mono-Operation Bias:* Any one operationalization of a construct both underrepresents the construct of interest and measures irrelevant constructs, complicating inference.
4. *Mono-Method Bias:* When all operationalizations use the same method (e.g., self-report), that method is part of the construct actually studied.
5. *Confounding Constructs with Levels of Constructs:* Inferences about the constructs that best represent study operations may fail to describe the limited levels of the construct that were actually studied.
6. *Treatment Sensitive Factorial Structure:* The structure of a measure may change as a result of treatment, change that may be hidden if the same scoring is always used.
7. *Reactive Self-Report Changes:* Self-reports can be affected by participant motivation to be in a treatment condition, motivation that can change after assignment is made.
8. *Reactivity to the Experimental Situation:* Participant responses reflect not just treatments and measures but also participants' perceptions of the experimental situation, and those perceptions are part of the treatment construct actually tested.
9. *Experimenter Expectancies:* The experimenter can influence participant responses by conveying expectations about desirable responses, and those expectations are part of the treatment construct as actually tested.
10. *Novelty and Disruption Effects:* Participants may respond unusually well to a novel innovation or unusually poorly to one that disrupts their routine, a response that must then be included as part of the treatment construct description.
11. *Compensatory Equalization:* When treatment provides desirable goods or services, administrators, staff, or constituents may provide compensatory goods or services to those not receiving treatment, and this action must then be included as part of the treatment construct description.
12. *Compensatory Rivalry:* Participants not receiving treatment may be motivated to show they can do as well as those receiving treatment, and this compensatory rivalry must then be included as part of the treatment construct description.
13. *Resentful Demoralization:* Participants not receiving a desirable treatment may be so resentful or demoralized that they may respond more negatively than otherwise, and this resentful demoralization must then be included as part of the treatment construct description.
14. *Treatment Diffusion:* Participants may receive services from a condition to which they were not assigned, making construct descriptions of both conditions more difficult.

External validity

Definition

- extent to which a causal relationship holds from persons, settings, treatments and outcomes who where in the experiment to those who where not
- targets of generalization:
  (1) narrow to broad: from the persons, settings, treatments and outcomes in an experiment to a larger population
  (2) broad to narrow: from the experimental sample to a smaller group or even an individual (especially for therapy)
  (3) at a similar level: from the experimental sample to another sample (e.g., when implementing a welfare measure)
  (4) to a similar (from a male job applicant in Seattle to male applicants in the US) or different kind (from a afroamerican male in NJ to a hispanic females in TX)
  (5) from a random sample to population members
- goal to design experiments that are more valid externally (e.g., by testing whether treatment effects hold over different outcomes / measures or different kinds of persons); but: heterogenous range of persons, treatments, outcomes and settings requires large samples to obtain adequate power

Threats

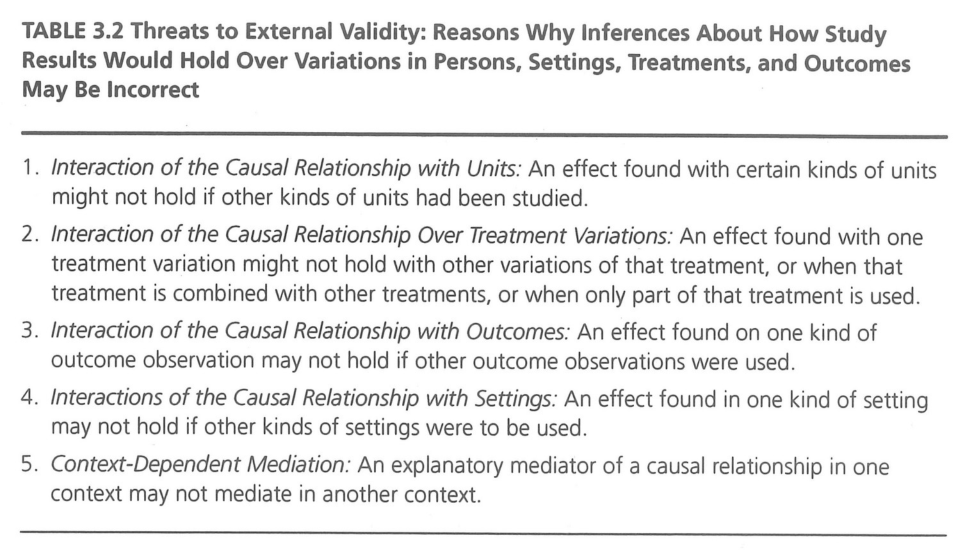


TABLE 3.2 Threats to External Validity: Reasons Why Inferences About How Study Results Would Hold Over Variations in Persons, Settings, Treatments, and Outcomes May Be Incorrect

1. *Interaction of the Causal Relationship with Units:* An effect found with certain kinds of units might not hold if other kinds of units had been studied.
2. *Interaction of the Causal Relationship Over Treatment Variations:* An effect found with one treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of that treatment is used.
3. *Interaction of the Causal Relationship with Outcomes:* An effect found on one kind of outcome observation may not hold if other outcome observations were used.
4. *Interactions of the Causal Relationship with Settings:* An effect found in one kind of setting may not hold if other kinds of settings were to be used.
5. *Context-Dependent Mediation:* An explanatory mediator of a causal relationship in one context may not mediate in another context.

(1) Self-selection

(2) interaction of drug effects (e.g., antibiotics and milk products); drug only working as a combination (AIDS)

(3) cancer: QoL, five year metastasis-free survival, overall survival; job-training programm might also train other skills and therefore be beneficial

Further aspects of external validity

- meta-analyses: constant effect sizes are rare but often the causal direction remains
- ***random sampling*** eliminates possible interactions between the causal relationship and the class of persons (or settings) who were studied to those who where not studied **vs.** ***purposive sampling*** of heterogeneous instances: persons, settings, treatments and outcomes deliberately chosen to be diverse → effect occus despite the heterogenity (variance and effect size)

Tradeoffs and priorities

Tradeoffs and priorities

- list of threats to the validity of generalized causal inference are heuristic devices
- no experiment can successfully avoid all of them; but: raise consciousness about priorities and tradeoffs and the choice which validity type should be emphasized
- what is emphasized also varies between basic (construct) and applied resaerchers (external)
- internal vs. external validity as sine qua non?

Summary

- experiments as a device for exploring cause-effect-relationships and to derive causal inference
- how to ensure validity of causal inference?
- types of validity: (1) internal validity, (2) statistical conclusion validity, (3) construct validity, (4) external validity
- tradeoffs and priorities in dealing with threats to validity

Literature

Shadish, W. R, Cook, T. D., Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Cengage Learning. (esp. Chap. 1-3)

Field, A., Hole, G. J. (2003). *How to design and report experiments*. London, UK: Sage Publications.

Popper, K. R. (1952). *The logic of scientific discovery*. London, UK: Routledge.

Rosenthal, R., Rosnow, R. L.(1969). *Artifacts in behavioral research*. New York, NY: Academic Press.

Valentine, E. R. (1992). *Conceptual issues in psychology*. Hove, East Sussex: Routledge.

Banyard, P., Grayson, A. (1996). *Introducing psychological research: Sixty studies that shape psychology*. New York, NY: New York University Press.

Thank you for your patience and (hopefully) your interest!

UNIVERSITY OF BERGEN