2IMP40 EMPIRICAL METHODS IN SOFTWARE ENGINEERING

THINGS THAT CAN GO WRONG

ALEXANDER SEREBRENIK

https://upload.wikimedia.org/wikipedia/commons/3/3a/Trp-Sword-14226124129-v06.png

https://www.thebluediamondgallery.com/handwriting/images/validity.jpg

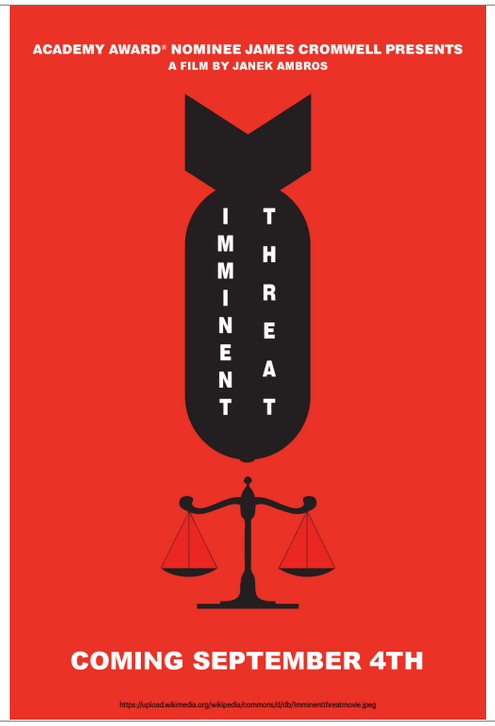
A fundamental question concerning results from an empirical study is how valid the results are. It is important to consider the question of validity already in the planning phase in order to plan for adequate validity of the experiment results. Adequate validity refers to that the results should be valid for the population of interest.

The text above comes from Wohlin’s book but it discusses experiments; to some extent this can be generalised to further qualitative empirical studies.

As I said adequate validity refers to that the results should be valid for the population of interest. But what does this mean? First of all, the results should be valid for the population from which the sample is drawn (for example, people of a certain city or country; developers of 25 projects we want to study). Secondly, it may be of interest to generalize the results to a broader population (people in general; open source developers in general). The results are said to have adequate validity if they are valid for the population to which we would like to generalize.

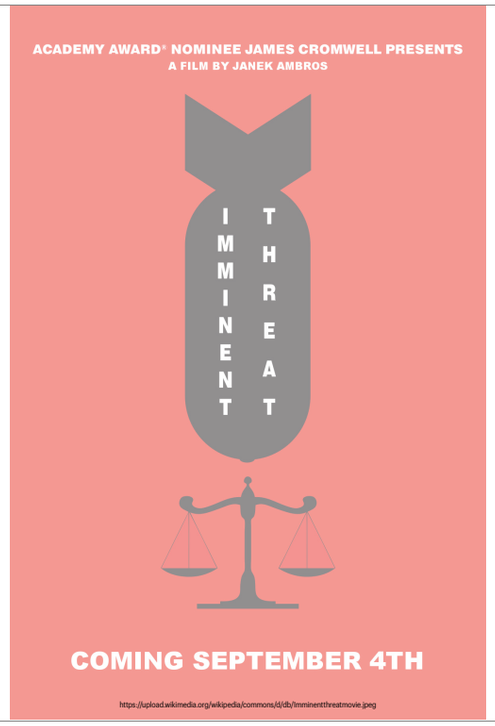
Adequate validity does not necessarily imply most general validity. An experiment conducted within an organization may be designed to answer some questions for that organization exclusively, and it is sufficient if the results are valid within that specific organization. On the other hand, if more general conclusions shall be drawn, the validity must cover a more general scope as well.

The text above comes from Wohlin’s book but it discusses experiments; this can be generalised to additional qualitative empirical studies.



So when discussing validity we need to think about things that can threaten validity of our study, things in our study that can go wrong, that can make the results based on a sample invalid for the population from which the sample is drawn, or non generalisable beyond this particular population to a much larger population to which we would like to generalize.

INTERNAL

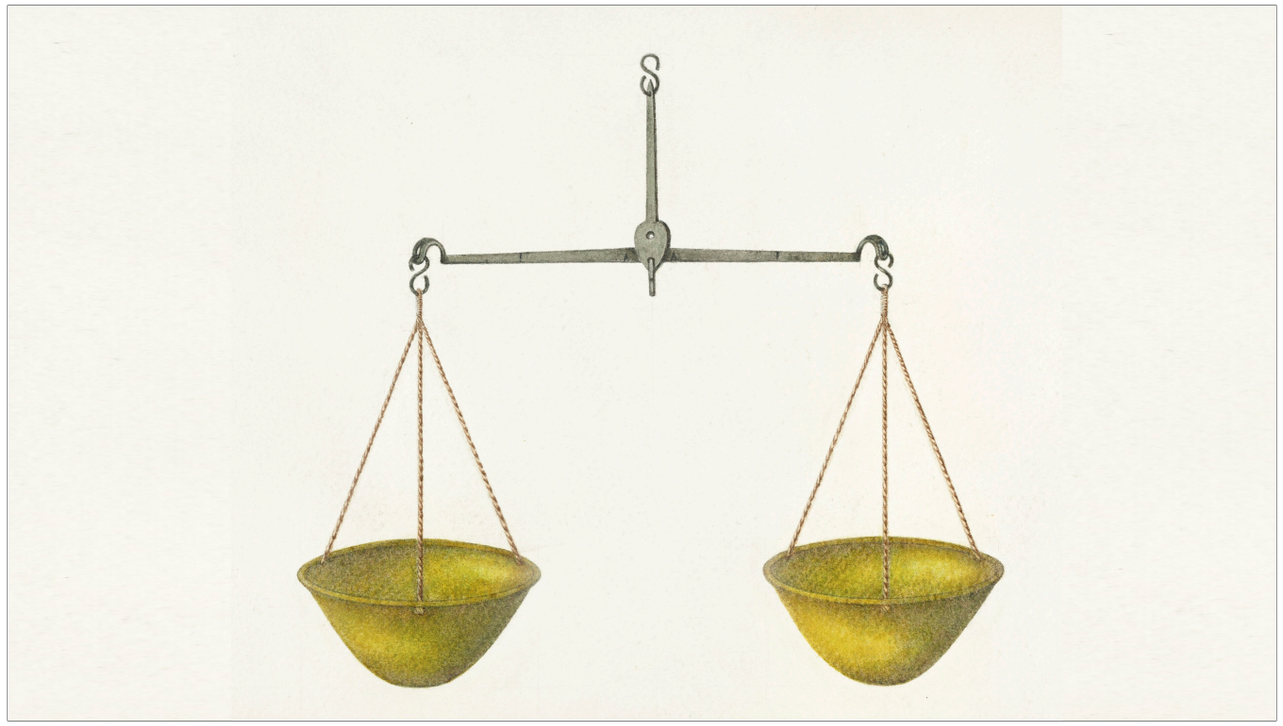


https://upload.wikimedia.org/wikipedia/commons/d/db/Imminentthreatmovie.jpeg

EXTERNAL

There have been several attempts to classify threats to validity. Such a classification is useful as a mechanism triggering reflection on what can go wrong in a specific way, as well as to communicate these concerns with other researchers.

The simplest distinction is between internal and external validity: roughly speaking internal validity is related to the conclusions derived for the study itself, external - to generalisability of its findings.



Keep in mind that balancing internal and external validity is not easy. Some decisions might favour internal validity (e.g., controlled experiments) while others - external validity (e.g., field studies). Gender: ask a question vs use an automatic tool - asking questions is strengthening internal validity, using tools - external.

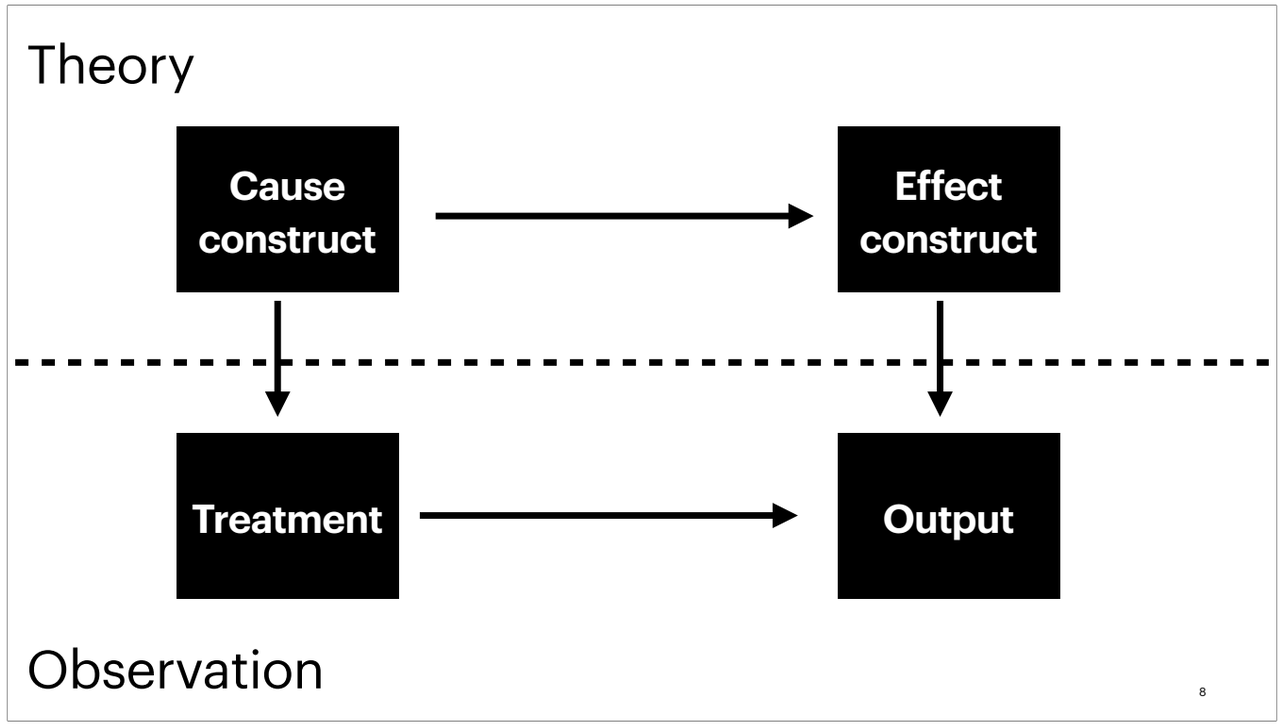
CONSTRUCT

INTERNAL

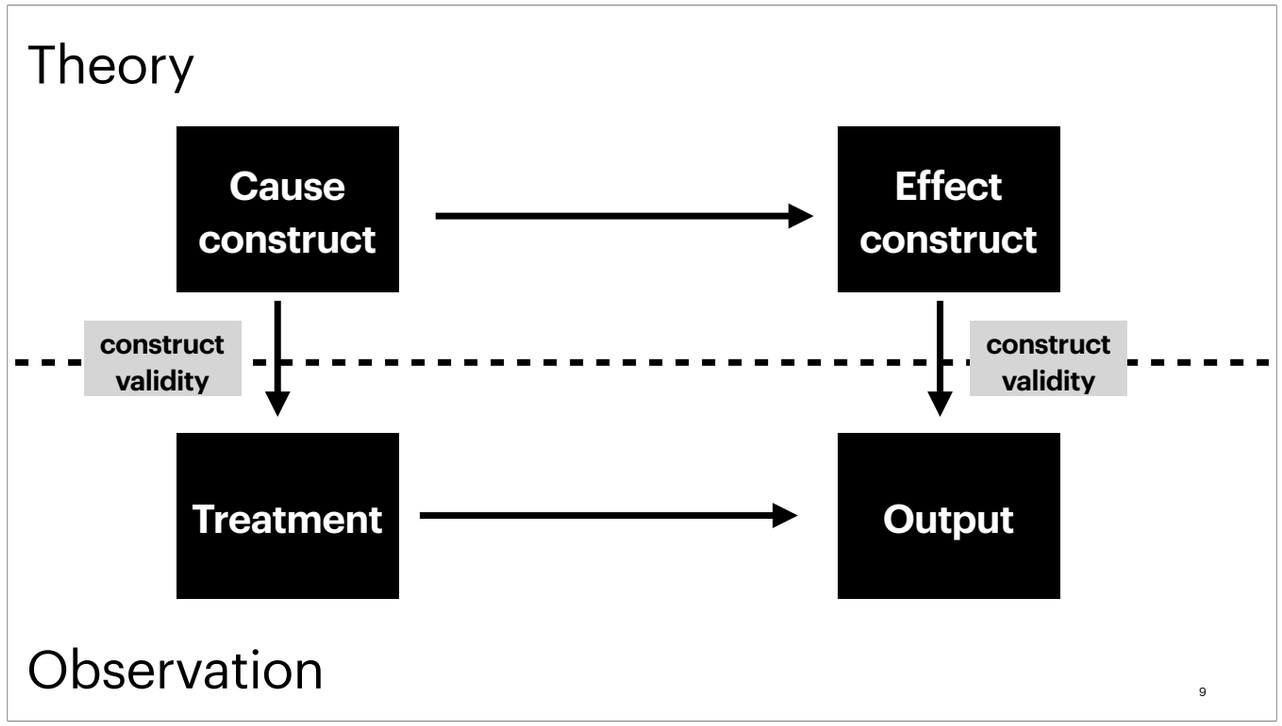
CONCLUSION

EXTERNAL

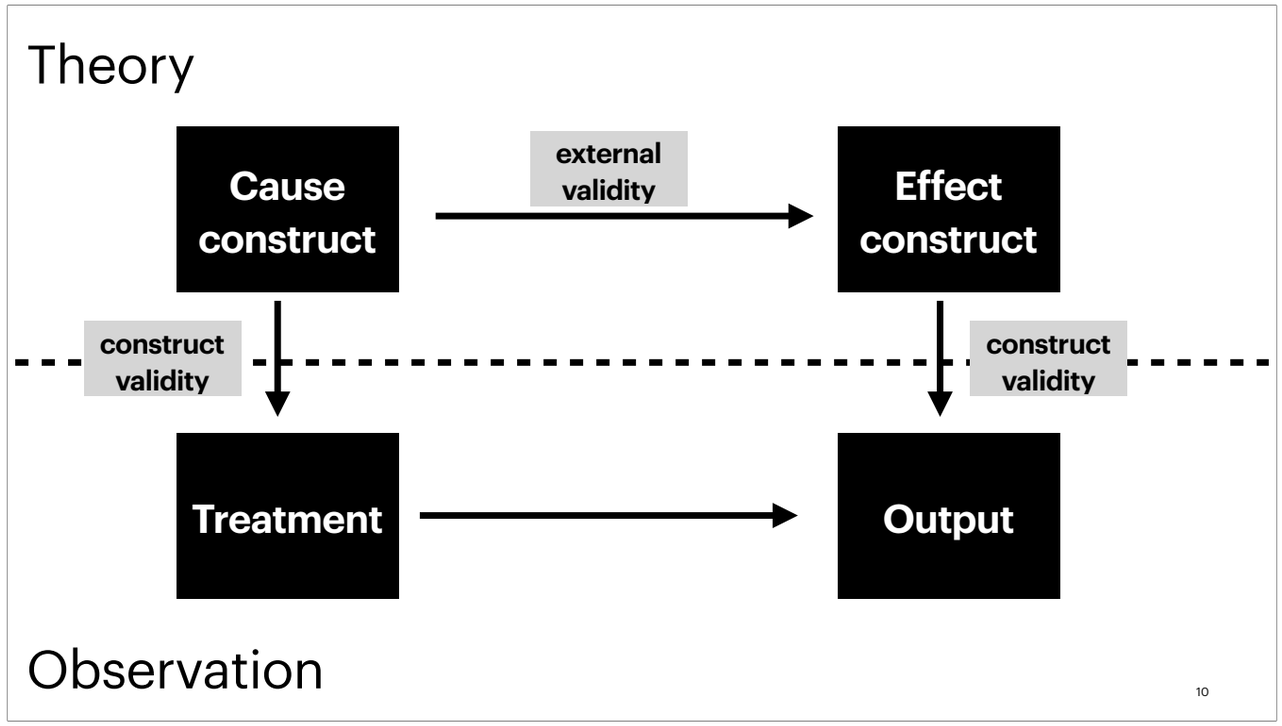
Wohlin et al. have proposed a more refined model that includes two more types of validity constructs, construct and conclusion validity. Wohlin et al. focus on experiments but quite some issues are relevant for other types of empirical research.



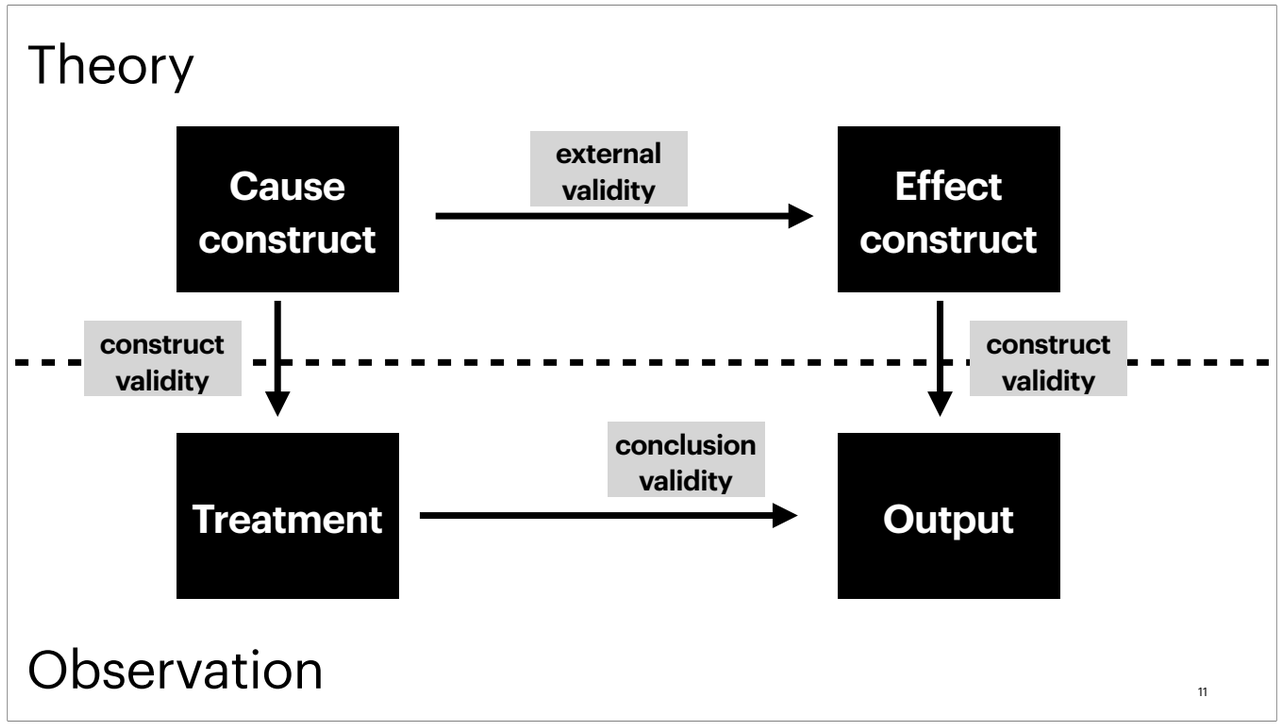
In experiments validity is related to relation between the following four components. For example, if our theory says that it takes more effort to maintain complex code, then the cause construct would be source code complexity, and the effect construct would be the maintenance effort. At the level of the observation, we need some kind of representation of both source code complexity and of the maintenance effort. For example, we can consider cyclomatic complexity or CK metrics as operationalisation of source code complexity, and time taken to perform a maintenance task as an operationalisation of the maintenance effort.



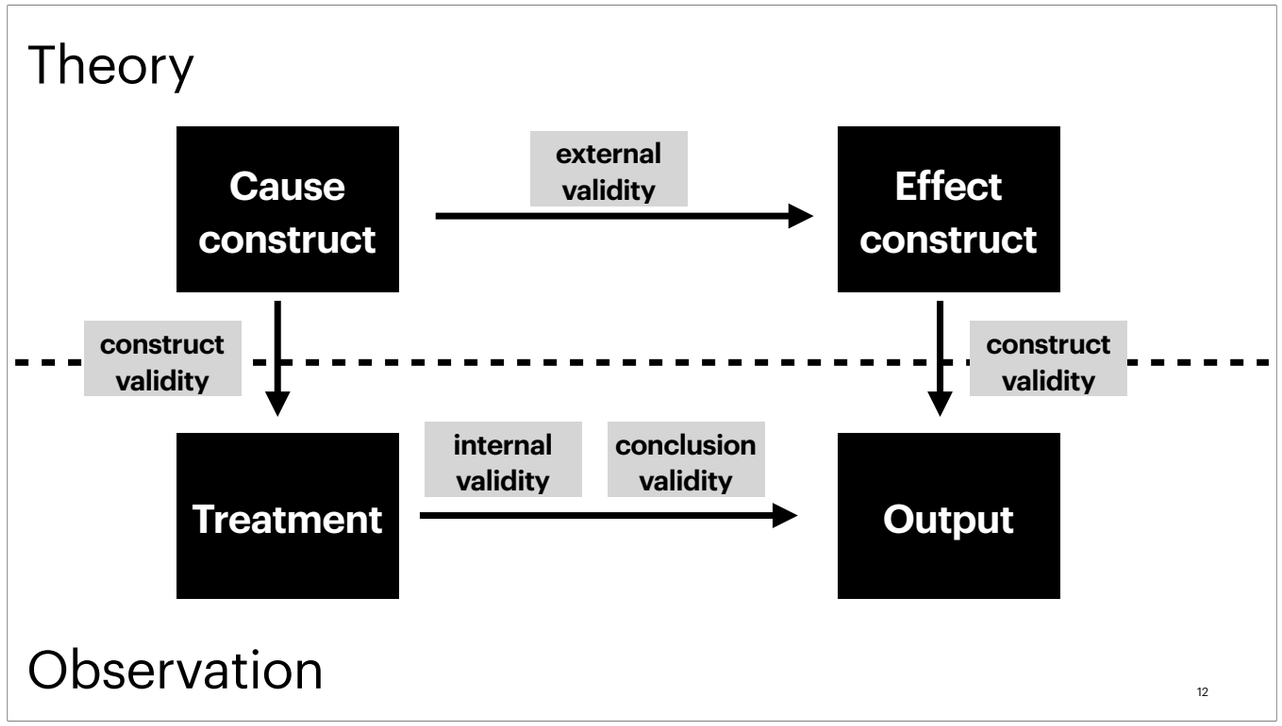
Construct validity. This validity is concerned with the relation between theory and observation. If the relationship between cause and effect is causal, we must ensure two things: (1) that the treatment reflects the construct of the cause well (see left part) and (2) that the outcome reflects the construct of the effect well (see right part). For example, operationalisation of maintenance effort as time induces threat to construct validity in an uncontrolled setting as developers might be taking a break.



External validity. The external validity is concerned with generalization. If there is a causal relationship between the construct of the cause, and the effect, can the result of the study be generalized outside the scope of our study? Is there a relation between the treatment and the outcome? So assume that our study has lead to concluding that indeed, source code complexity causes increase in the maintenance effort. However, external validity might be threatened if we have derived our conclusion from only very small open source systems and try to generalise this to *any* open source software systems.



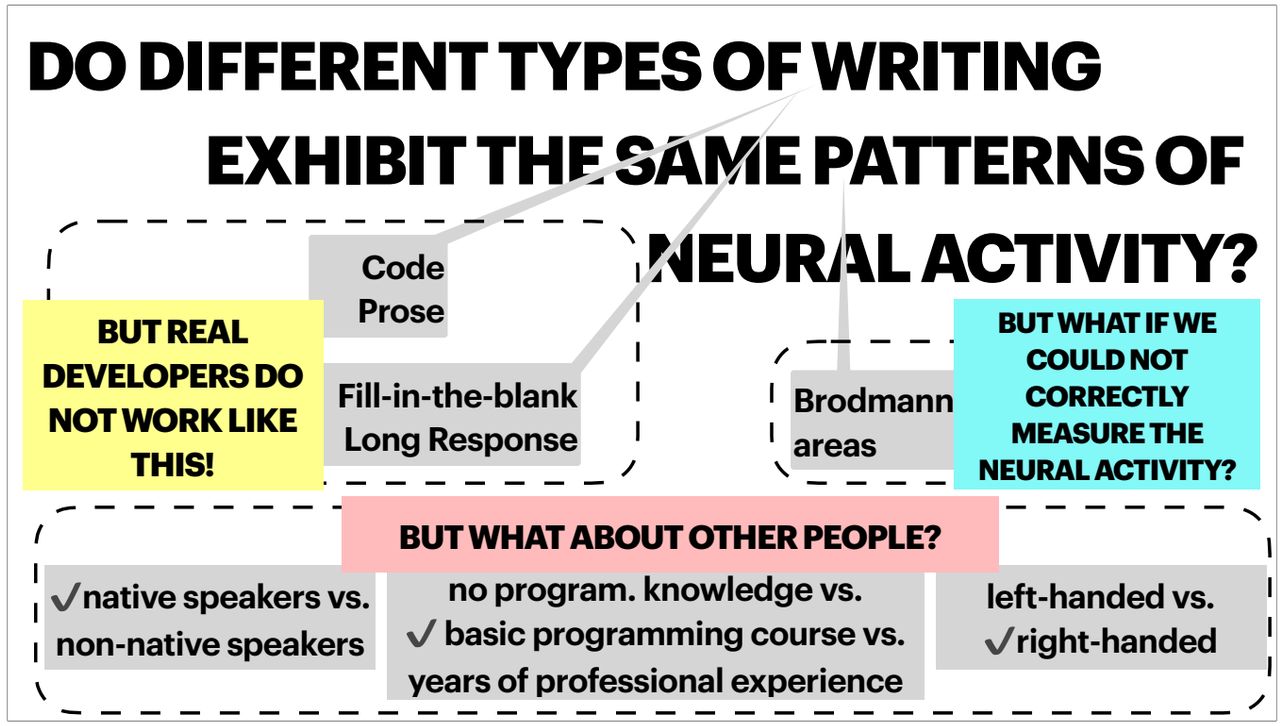
Conclusion validity. This validity is concerned with the relationship between the treatment and the outcome. We want to make sure that there is a statistical relationship, i.e. with a given significance. For example, violated assumption of statistical tests that we have discussed during one of the lectures would threaten the conclusion validity. This is often about **statistical analysis** (but not always).



Internal validity. If a relationship is observed between the treatment and the outcome, we must make sure that it is a causal relationship, and that it is not a result of a factor of which we have no control or have not measured. In other words that the treatment causes the outcome (the effect). This is much more about the **design of the experiment** itself. Factors that impact on the internal validity are how the subjects are selected and divided into different classes, how the subjects are treated and compensated during the experiment, if special events occur during the experiment etc. All these factors can make the experiment show a behaviour that is not due to the treatment but to the disturbing factor.

RYAN KRUEGER, YU HUANG, XINYU LIU, TYLER SANTANDER, WESTLEY WEIMER, KEVIN LEACH. NEUROLOGICAL DIVIDE - AN FMRI STUDY OF PROSE AND CODE WRITING. INT CONF SOFTWARE ENGINEERING 2020

You might remember this example from the first lecture. The authors used fMRI to study whether code writing and prose writing are similar in terms of brain area activation.

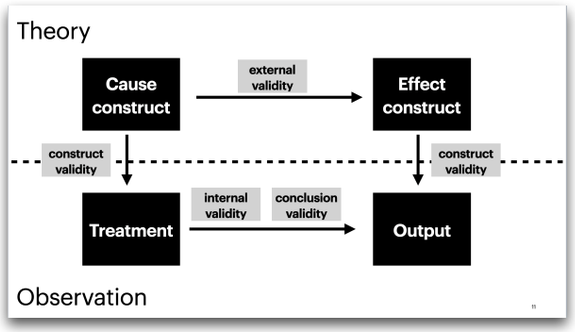


In this study we can identify at least three different threats to validity. On the left we see an issue related to the organisation of the task. In practice developers do not feel the blanks and usually do not write small coding exercise such as “is_sorted” that returns true if the input vector is sorted. We can see this is as a threat to the relation between the abstract notion of “writing code” and the way it has been translated to the experimental tasks. This is a threat to construct validity.

On the right you see a concern related to correctness of measuring the neural activity. For example, the stimuli included written instructions that participants read before typing their responses. This design decision introduces the possibility that the measured brain activity goes beyond strictly writing responses and also measures the neural activity related to reading. This has impacted the way the authors have designed their experiment: their fMRI analyses are subtractive such that the effects of reading the prompt cancel out, leaving only the differences between prose writing and code writing. This is an example of a threat to the internal validity, and specifically to instrumentation.

Finally, the threat in the middle is related to generalisation of the findings beyond the participants of the study: this is a threat to external validity.

QUESTION



Researchers have asked six experts to evaluate *modularity* of several software systems. Discussion revealed that experts A, B and C interpreted modularity as *coupling*, while experts D, E and F interpreted modularity as *cohesion*. This threatens

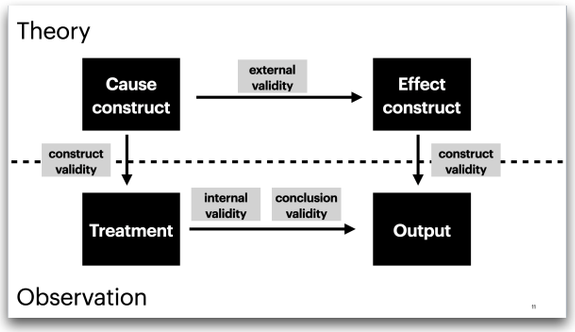
(A) CONSTRUCT VALIDITY
(B) EXTERNAL VALIDITY

(C) CONCLUSION VALIDITY
(D) INTERNAL VALIDITY

YANJA DAJSUREN, MARK G. J. VAN DEN BRAND, ALEXANDER SEREBRENIK, SERGUEI A. ROUBTSOV: SIMULINK MODELS ARE ALSO SOFTWARE: MODULARITY ASSESSMENT. QOSA 2013: 99-106

A - the “modularity” construct was not sufficiently clarified.
Inadequate pre-operational explication of constructs. The constructs are not sufficiently defined, before they are translated into measures or treatments. The theory is not clear enough, and hence the experiment cannot be sufficiently clear. For example, if two inspection methods are compared and it is not clearly enough stated what being ‘better’ means. Does it mean to find most faults, most faults per hour, or most serious faults?

QUESTION

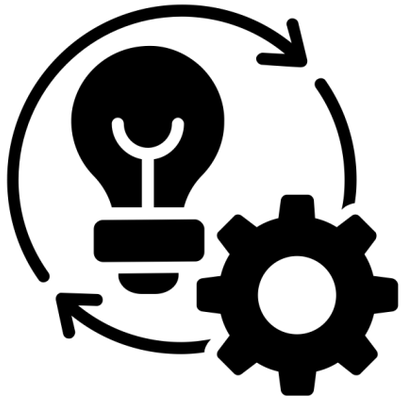


Researchers have applied Student's t-test but did not verify whether the samples compared are likely to have been drawn from a normal distribution. This threatens

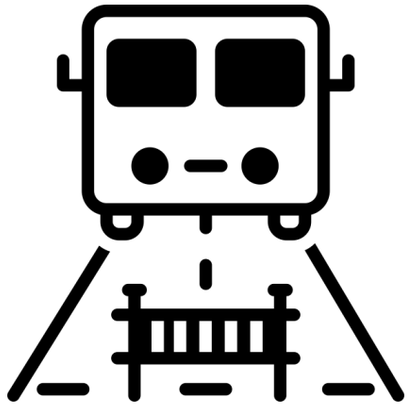
(A) CONSTRUCT VALIDITY
(B) EXTERNAL VALIDITY
(C) CONCLUSION VALIDITY
(D) INTERNAL VALIDITY

YANJA DAJSUREN, MARK G. J. VAN DEN BRAND, ALEXANDER SEREBRENIK, SERGUEI A. ROUBTSOV: SIMULINK MODELS ARE ALSO SOFTWARE: MODULARITY ASSESSMENT. QOSA 2013: 99-106

C - assumptions of the statistical test might have been violated.



Created by ProSymbols
from Noun Project



Created by priyanka
from Noun Project

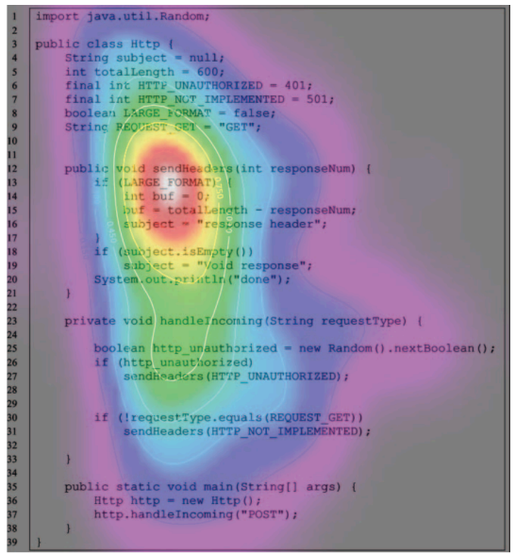
IMPROVE THE STUDY DESIGN

UNDERSTAND THE LIMITATIONS

Next we are going to take a closer look at the threats to validity, and in particular in the context of the research methods that we have discussed during the previous lectures. However, first of all we need to answer an important question: why do we even need to think about the threats to validity?

- reflect on the threats and improve the experiment to reduce the threats
- be aware of the inherent limitations

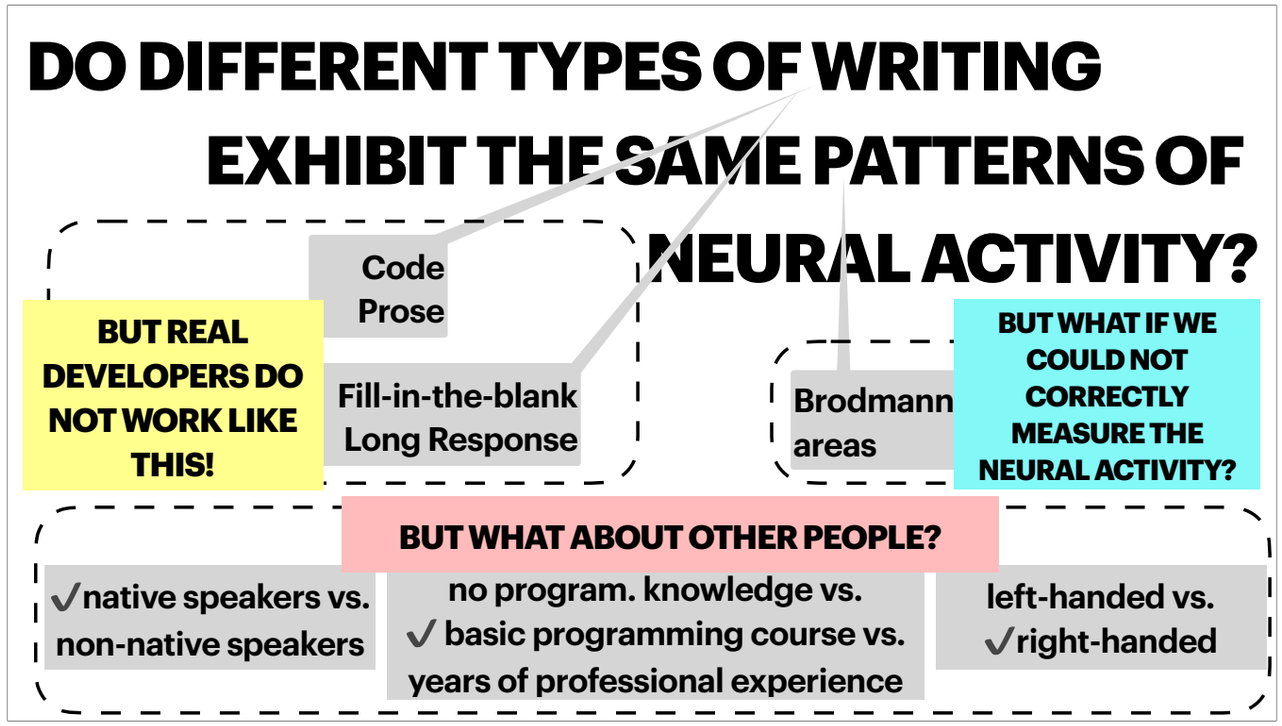
NOT AN EXCUSE!



JEAN MELO, FABRICIO BATISTA NARCIZO, DAN WITZNER HANSEN, CLAUS BRABRAND, ANDRZEJ WASOWSKI: VARIABILITY THROUGH THE EYES OF THE PROGRAMMER. ICPC 2017: 34-44

It is important to understand that threats to validity are not an excuse not to improve the design as much as possible. For example, recall that operationalisation of maintenance effort as time induces threat to construct validity in an uncontrolled setting as developers might be taking a break. Instead one could, for example, consider moving to a controlled setting when the participant cannot take a break, or at least the experimenter knows what happens. Alternatively, one can consider a different way of measuring the maintenance effort by recording the eye-movement.

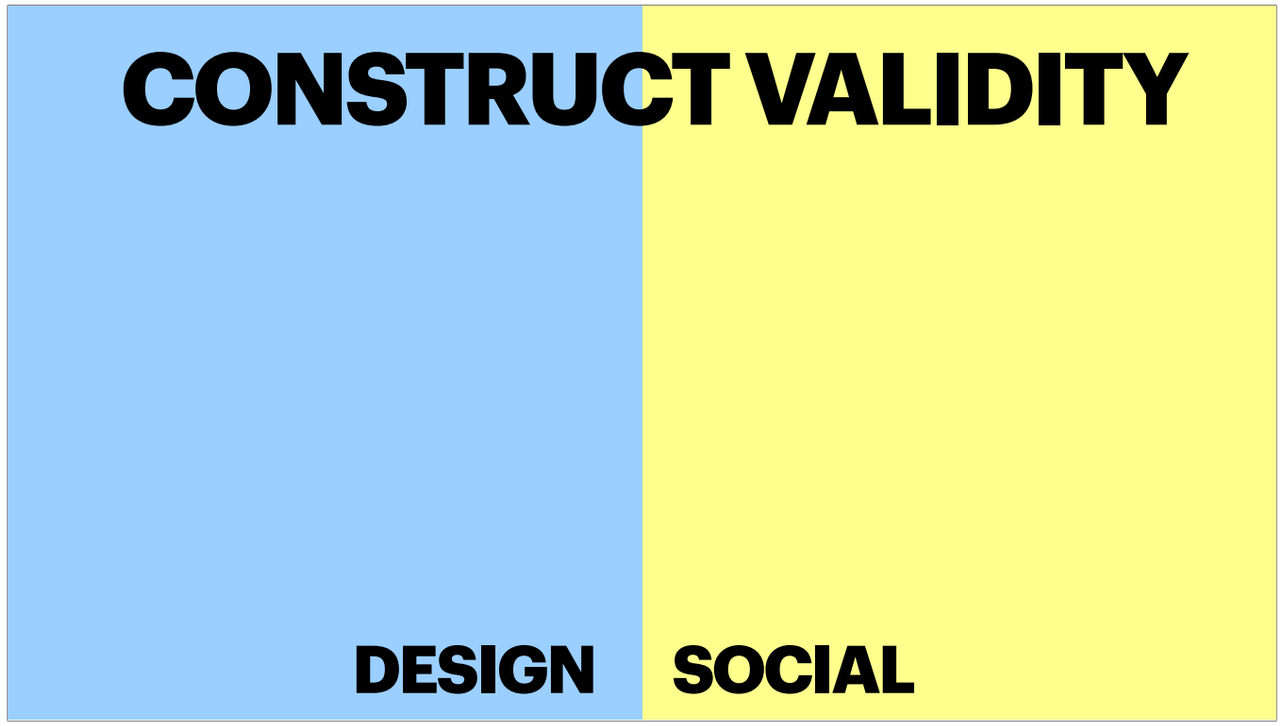
The meaning of fixations is context-dependent. A higher fixation rate on a specific AOI may indicate greater interest in its content, such as when reading some statements in a source code file. However, a cluster of fixations may also indicate effort/difficulties in understanding.



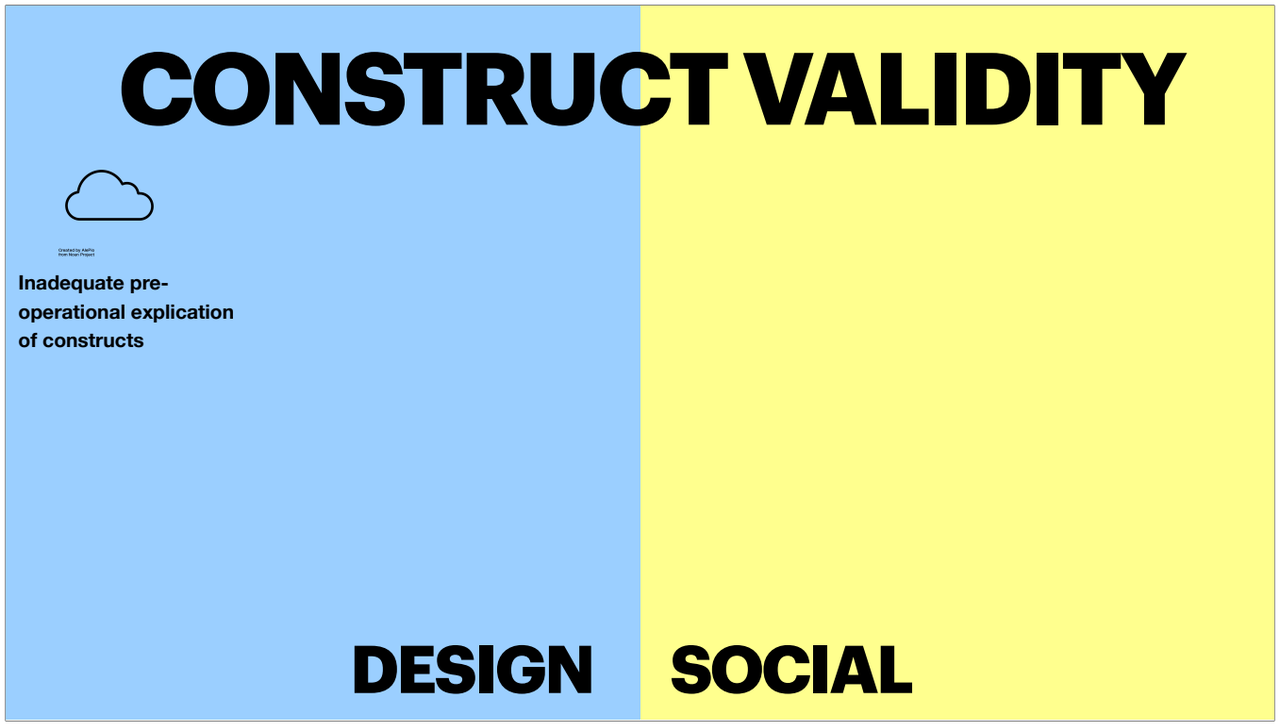
How can we improve the design? Or what alternatives could we consider?

- listening the experimenter reading the task description instead of reading it themselves
- more realistic setting: people performing tasks at their desks but then fMRI is not an option but an EEG, for example, is still possible, but is it good enough? For small tasks (= lab) we still might find volunteers CHECK THE LITERATURE but it will be hard to implement with professionals in their normal settings (= field)
- more people and different people but the participants need to be paid (= expenses are higher)

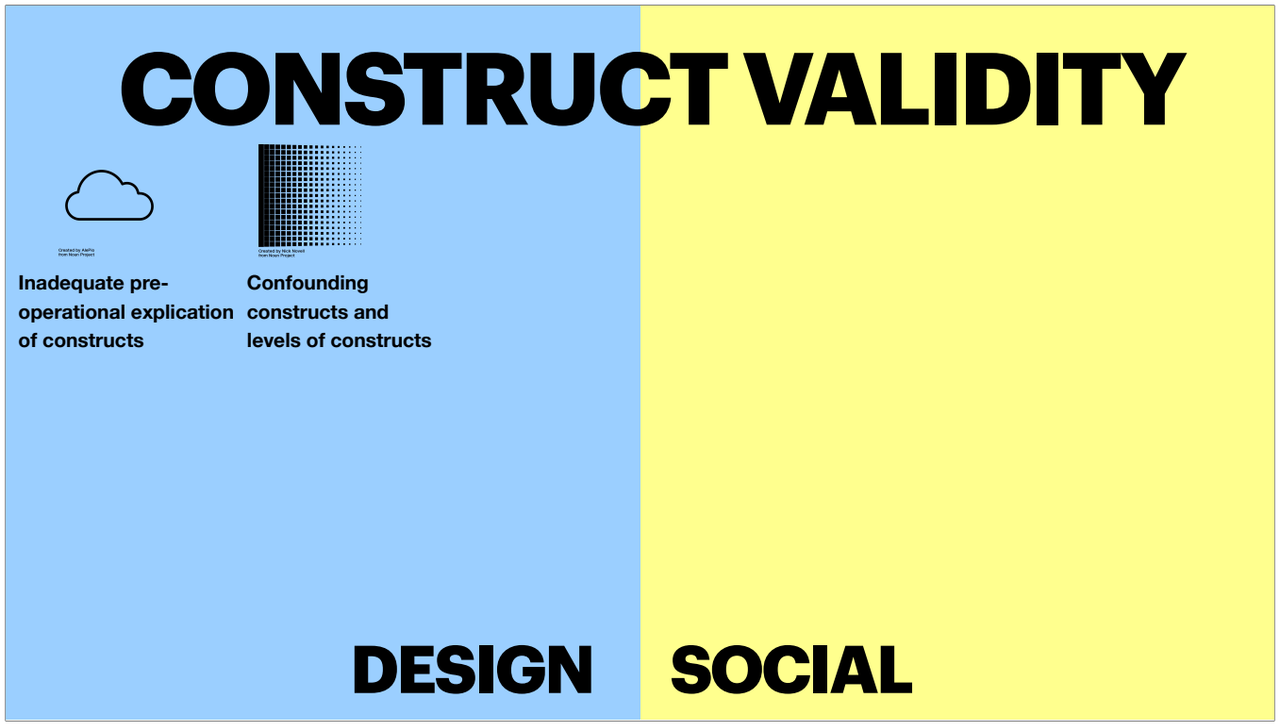
Feasibility constraints => Engineering as finding the best solution under the given constraints



The **design threats** to construct validity cover issues that are related to the design of the experiment and its ability to reflect the construct to be studied.
The **social threats** are concerned with issues related to behavior of the subjects and the experimenters. They may, based on the fact that they are part of an experiment, act differently than they do otherwise, which gives false results from the experiment.



Inadequate pre-operational explication of constructs: we have already seen it when discussing the modularity evaluation experiment. The constructs are vague, are insufficiently defined, before they are translated into measures or treatments. We have seen an example of this before: a problem with operationalisation of modularity.



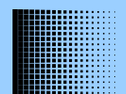
Confounding constructs and levels of constructs. In some relations it is not primarily the presence or absence of a construct, but the level of the construct which is of importance to the outcome, i.e., not merely a distinction between black and white but between different shades.

For example, the presence or absence of prior knowledge in a programming language may not explain the causes in an experiment, but the difference may depend on if the subjects have 1, 3 or 5 years of experience with the current language.

CONSTRUCT VALIDITY

Created by Adrific
from Noun Project

Inadequate pre-operational explication of constructs



Created by Nick Street
from Noun Project

Confounding constructs and levels of constructs

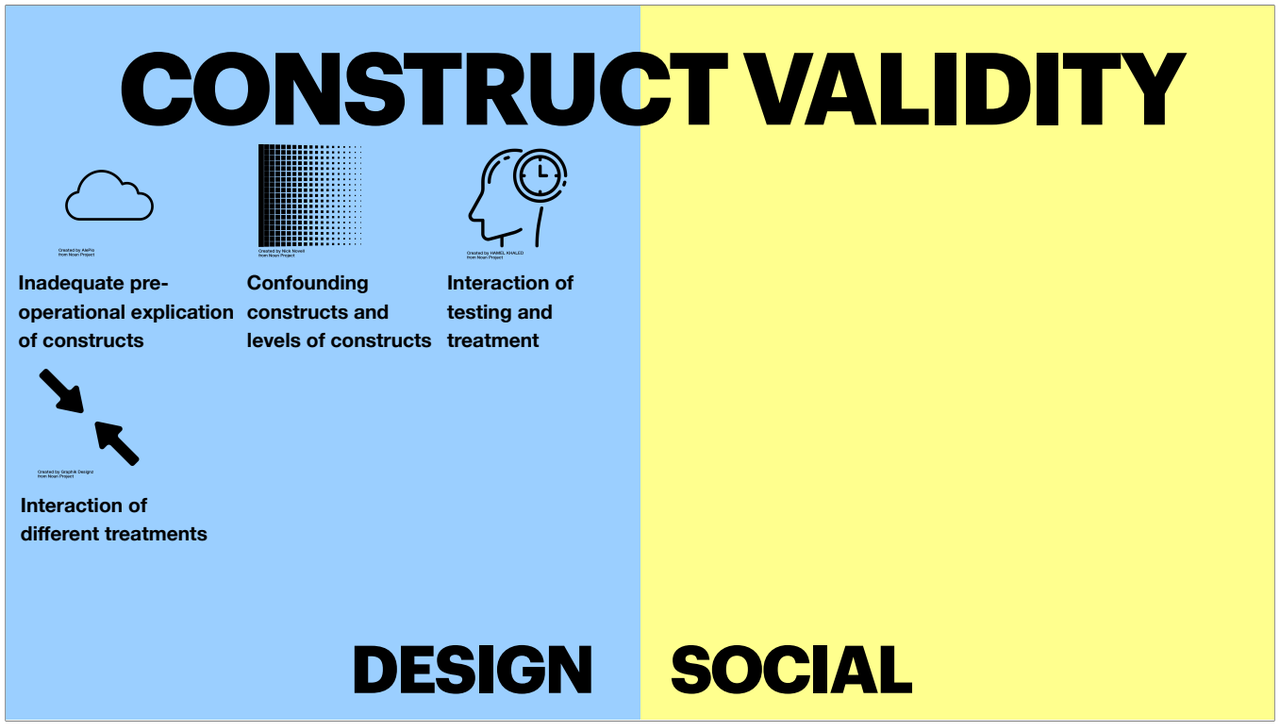
Created by HAMEL KHALED
from Noun Project

Interaction of testing and treatment

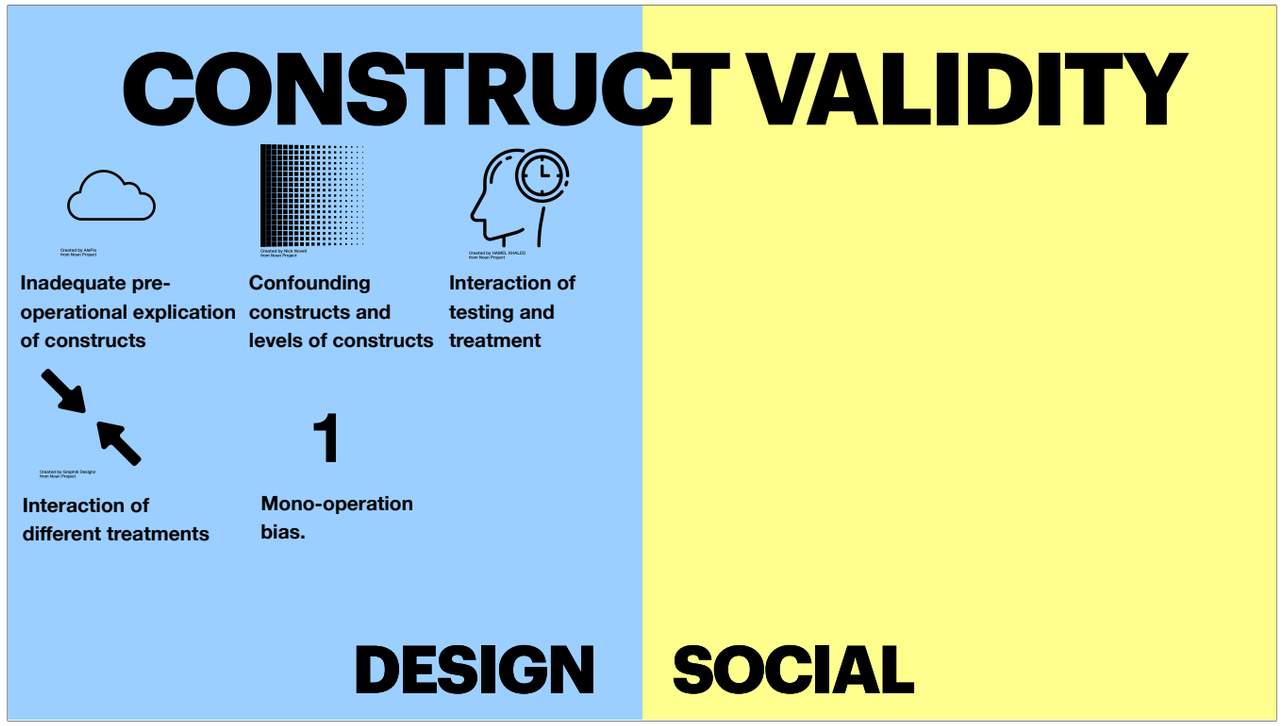
DESIGN

SOCIAL

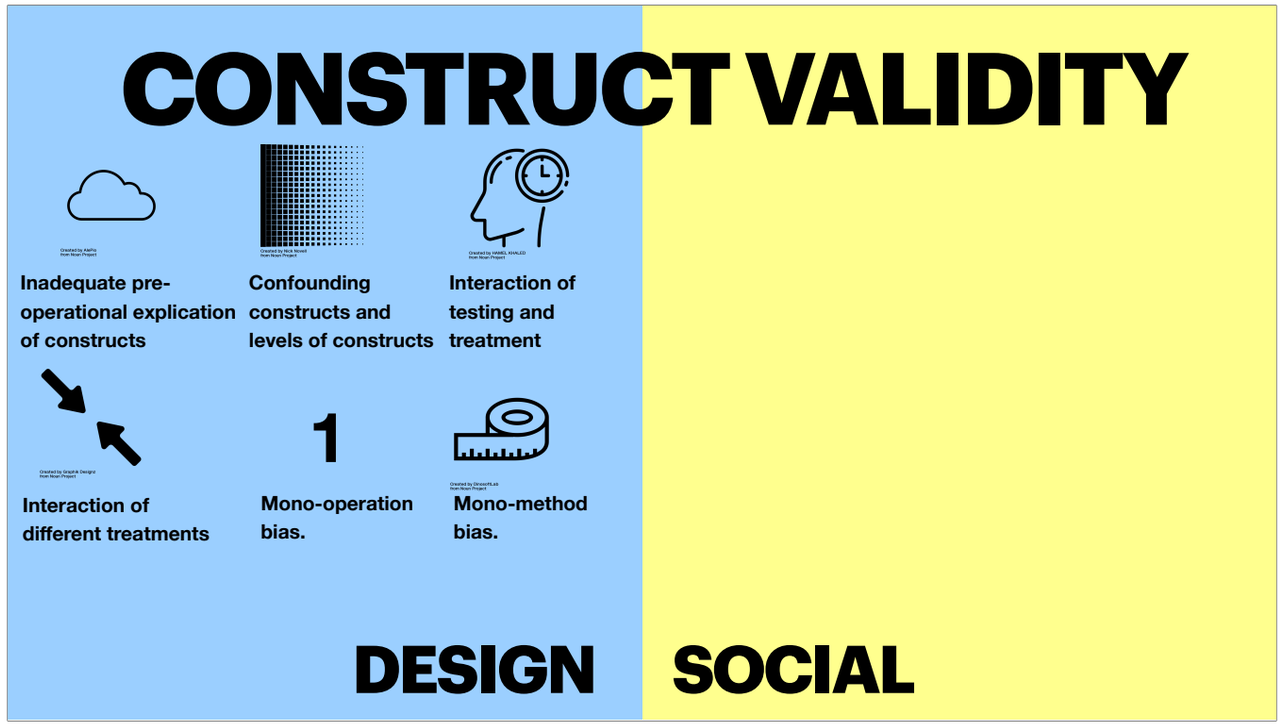
Interaction of testing and treatment. The testing itself, i.e. the application of treatments, may make the subjects more sensitive or receptive to the treatment. Then the testing is a part of the treatment. For example, if the testing involves measuring the time, then the subjects will feel pressured and try to be as fast as possible; similarly, if we measurlthe number of errors made in coding, then the subjects will be more aware of their errors made, and thus try to reduce them.



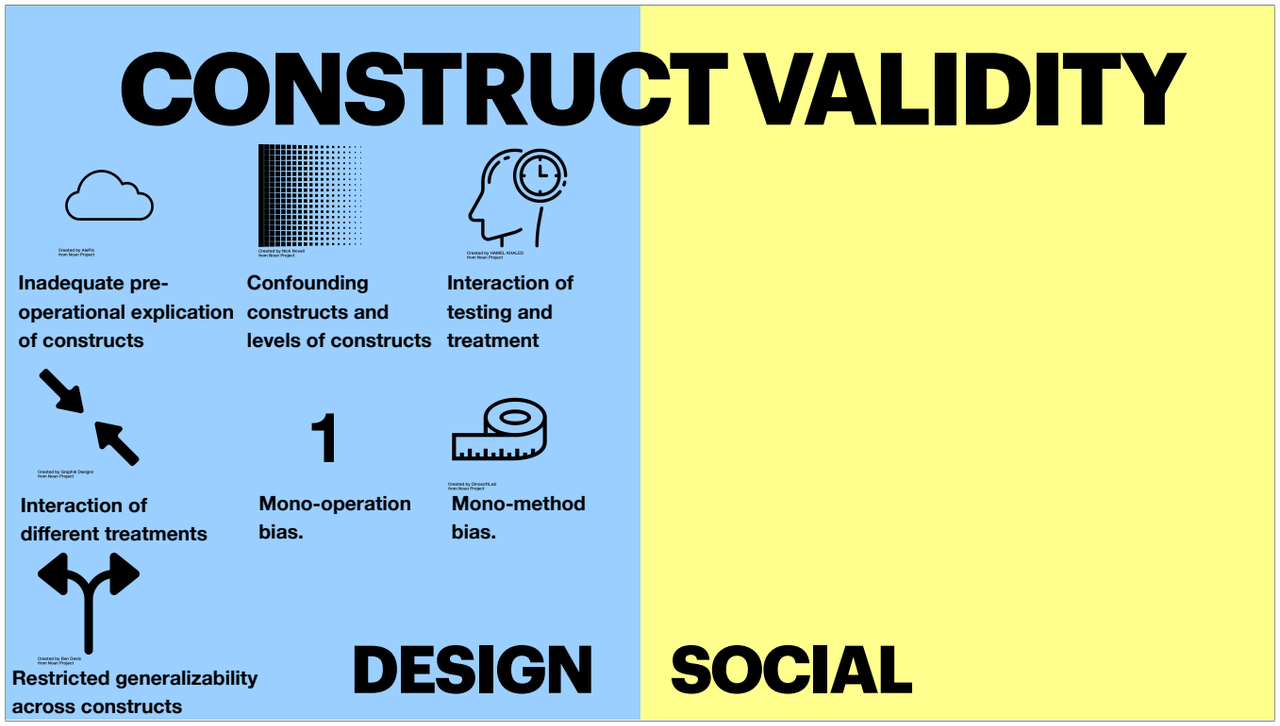
Interaction of different treatments. If the subject is involved in more than one study, treatments from the different studies may interact. Then you cannot conclude whether the effect is due to either of the treatments or of a combination of treatments. To make this more tangible: if one wants to study the impact of continuous integration on pull request resolution one might need to exclude projects that have introduced code review bots since both interventions can have an impact on pull request resolution.



Mono-operation bias. If the experiment includes a single **independent** variable, case, subject or treatment, the experiment may under-represent the construct and thus not give the full picture of the theory. => Alternative design should consider multiple variables to represent the construct. For example, for the construct “complexity” we can take into account traditional complexity metrics (e.g., cyclomatic complexity), OO-metrics (e.g., depth of inheritance tree), adherence of variable names to the Java standard etc

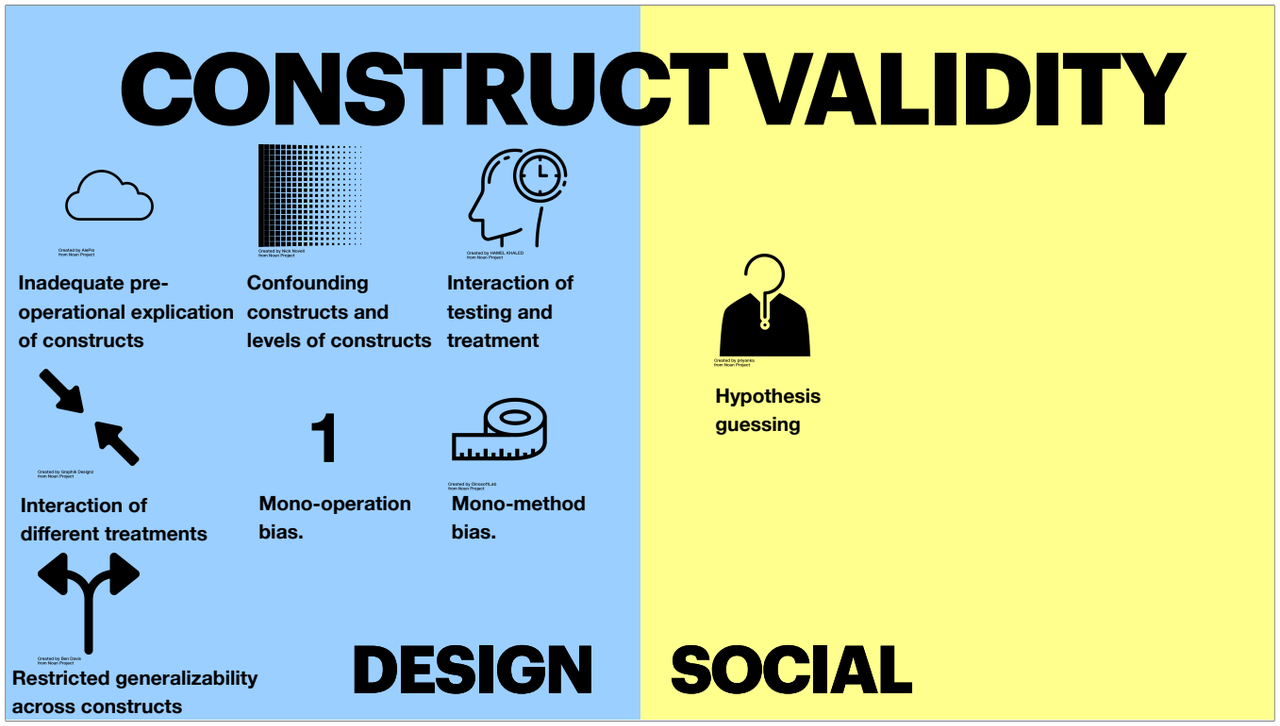


Mono-method bias. Using a single type of measures or observations involves a risk that if this measure or observation gives a measurement bias, then the experiment will be misleading. By involving different types of measures and observations they can be cross-checked against each other. For example, if the number of faults found is measured in an inspection experiment, where fault classification is based on subjective judgement, the relations cannot be sufficiently explained. The experimenter may bias the measures.



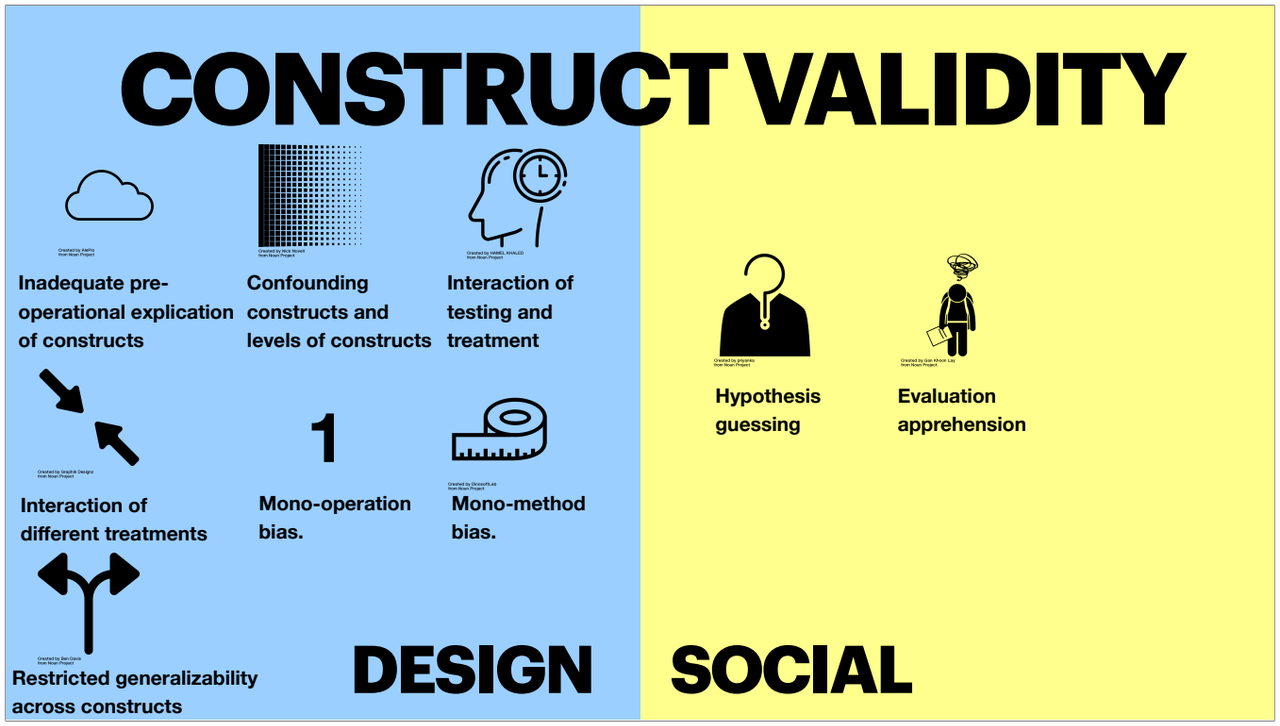
Restricted generalizability across constructs. The treatment may affect the studied construct positively, but unintenionally affect other constructs negatively. This threat makes the result hard to generalize into other potential outcomes.

For example, a comparative study concludes that improved productivity is achieved with a new method. However, it can be observed that the new method reduces the maintainability, which is an unintended side effect. If the maintainability is not measured or observed, there is a risk that conclusions are drawn based on the productivity attribute, ignoring the maintainability.



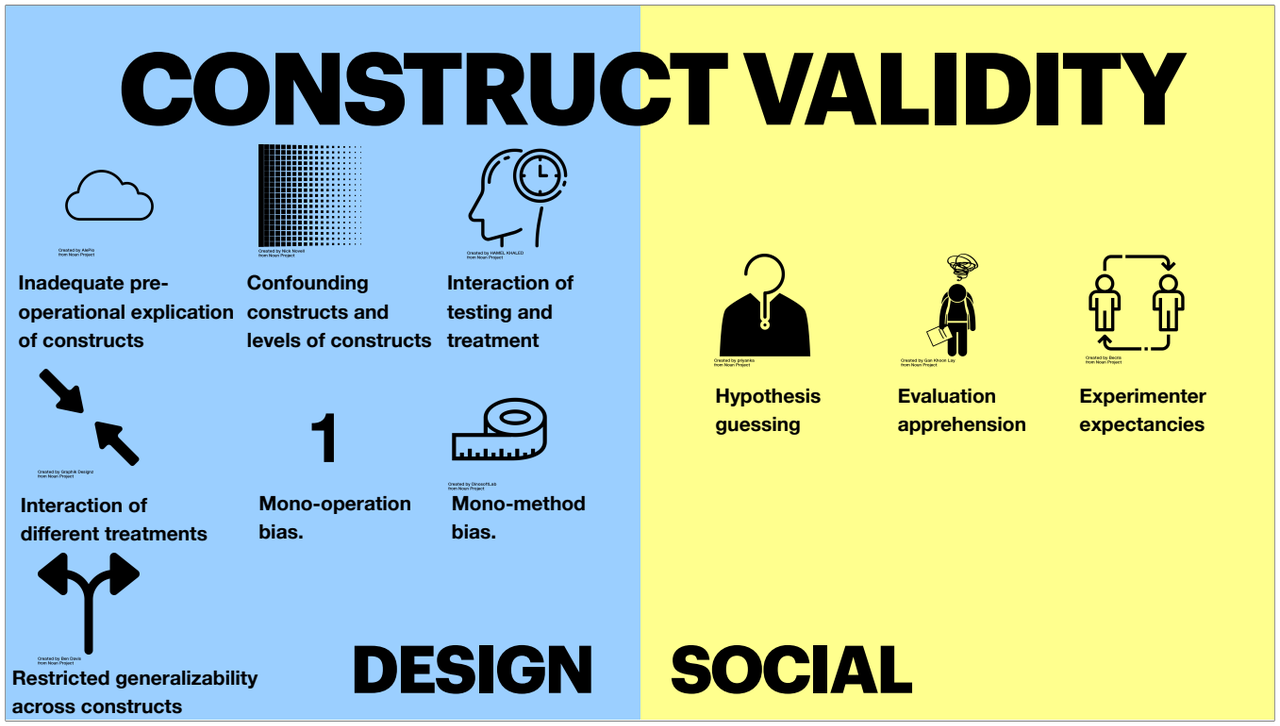
Reminder: the **social threats** are concerned with issues related to behaviour of the subjects and the experimenters. They may, based on the fact that they are part of an experiment, act differently than they do otherwise, which gives false results from the experiment.

Hypothesis guessing. When people take part in an experiment they might try to figure out what the purpose and intended result of the experiment is. Then they are likely to base their behaviour on their guesses about the hypotheses, either positively or negatively, depending on their attitude to the anticipated hypothesis. Some researchers introduce extra measures to make the goal of the study less apparent or explicitly deceive participants. Deception is not prohibited but requires a very careful argument and an approval of the ethical review board.

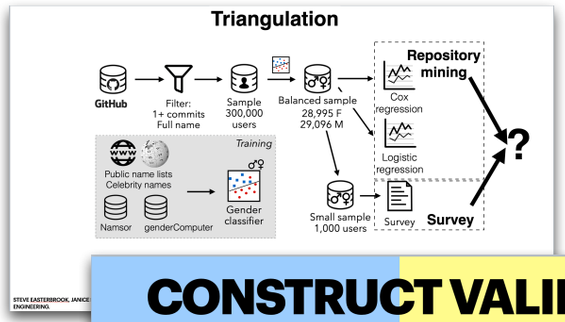


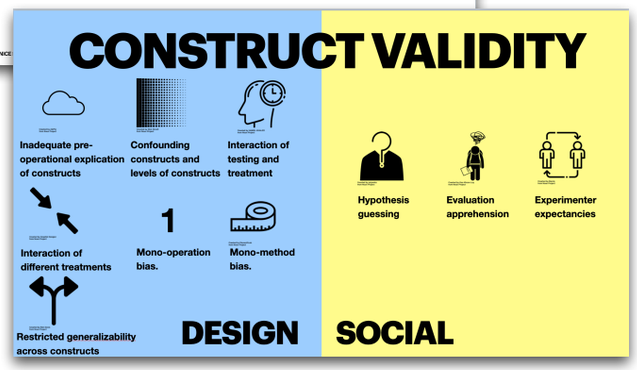
Evaluation apprehension. Some people are afraid of being evaluated. A form of human tendency is to try to look better when being evaluated which is confounded to the outcome of the experiment.

For example, if different estimation models are compared, people may not report their true deviations between estimate and outcome, but some false but ‘better’ values.



Experimenter expectancies. The experimenters can bias the results of a study both consciously and unconsciously based on what they expect from the experiment. The threat can be reduced by involving different people which have no or different expectations to the experiment. For example, questions can be raised in different ways in order to give the answers you want.



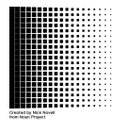


QUESTION

Which one of these threats can be addressed by **triangulation**?

(A) 

Hypothesis guessing

(C) 

Confounding constructs and levels of constructs

(B) 

Interaction of different treatments

(D) 

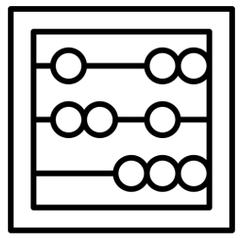
Mono-method bias.

D - mono-method bias

EXTERNAL VALIDITY

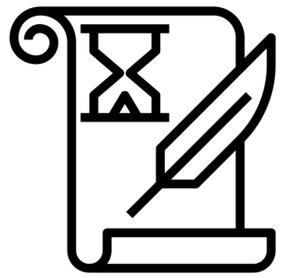
Created by Vectors Point
from Noun Project

Interaction of <u>selection</u> and treatment



Created by Made x Made
from Noun Project

Interaction of <u>setting</u> and treatment



Created by Becris
from Noun Project

Interaction of <u>history</u> and treatment

Threats to external validity are conditions that limit our ability to generalize the results of our experiment.

Interaction of selection and treatment. This is an effect of having a subject population, not representative of the population we want to generalize to, i.e. the wrong people participate in the experiment. Studies conducted on students aiming to generalise for professionals.

Interaction of setting and treatment. This is the effect of not having the experimental setting or material representative of, for example, industrial practice. An example is using old-fashioned tools in an experiment when up-to-date tools are common in industry. Another example is conducting experiment on toy problems. This means wrong ‘place’ or environment.

Interaction of history and treatment. This is the effect of that the experiment is conducted on a special time or day which affects the results. If, for example, a questionnaire is conducted on safety-critical systems a few days after a big software-related crash, people tend to answer differently than a few days before, or some weeks or months later.

REALISM/GENERALISABILITY ⇒ EXTERNAL VALIDITY

Do you remember this picture?

The threats to external validity are reduced by making the experimental environment as realistic as possible. At the same time it should not be too specific. Reality is not homogenous.

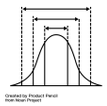
CONCLUSION VALIDITY

Threats to the conclusion validity are concerned with issues that affect the ability to draw the correct conclusion about relations between the treatment and the outcome of an experiment.

CONCLUSION VALIDITY

STATISTICS

Low statistical power



Violated assumptions of statistical tests

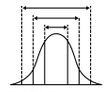
Fishing and the error rate

Low statistical power. The power of a statistical test is the ability of the test to reveal a true pattern in the data. If the power is low, there is a high risk that an erroneous conclusion is drawn.
Violated assumptions of statistical tests. Certain tests have assumptions on, for example, normally distributed and independent samples. Violating the assumptions may lead to wrong conclusions.
Fishing and the error rate. This threat contains two separate parts. Searching or ‘fishing’ for a specific result is a threat, since the analyses are no longer independent and the researchers may influence the result by looking for a specific outcome. The error rate is concerned with the actual significance level. The error rate (i.e. significance level) should thus be adjusted when conducting multiple analyses.

CONCLUSION VALIDITY

STATISTICS

Low statistical power



Violated assumptions of statistical tests

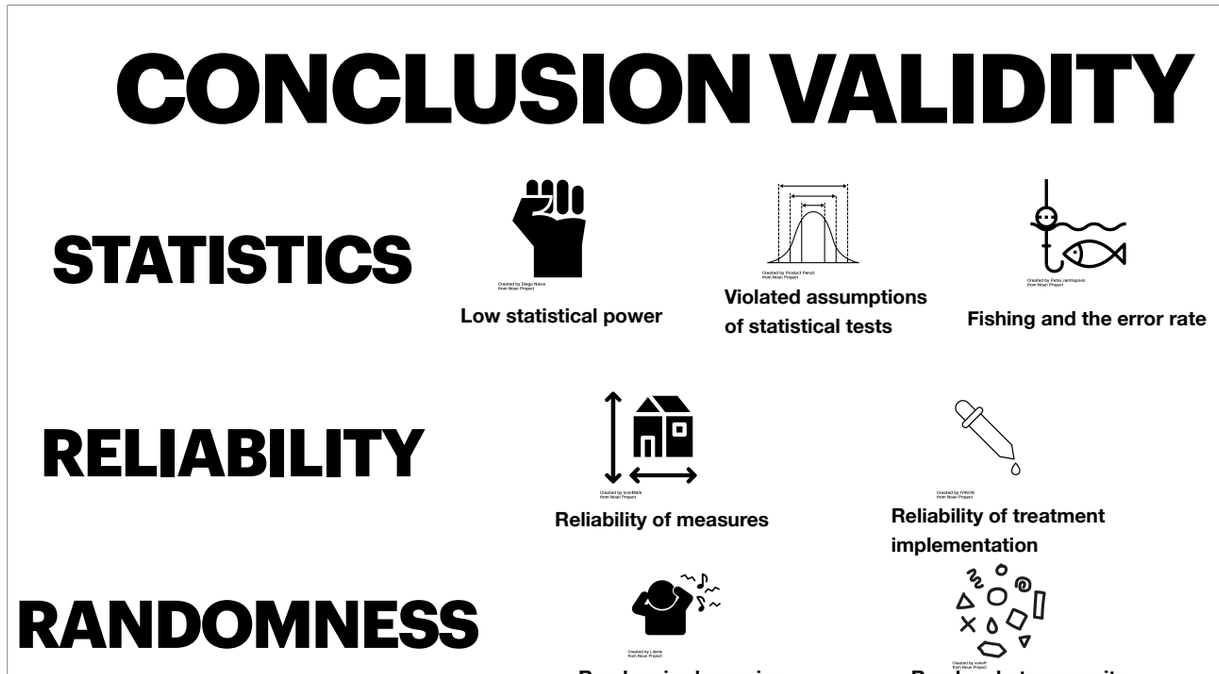
Fishing and the error rate

RELIABILITY

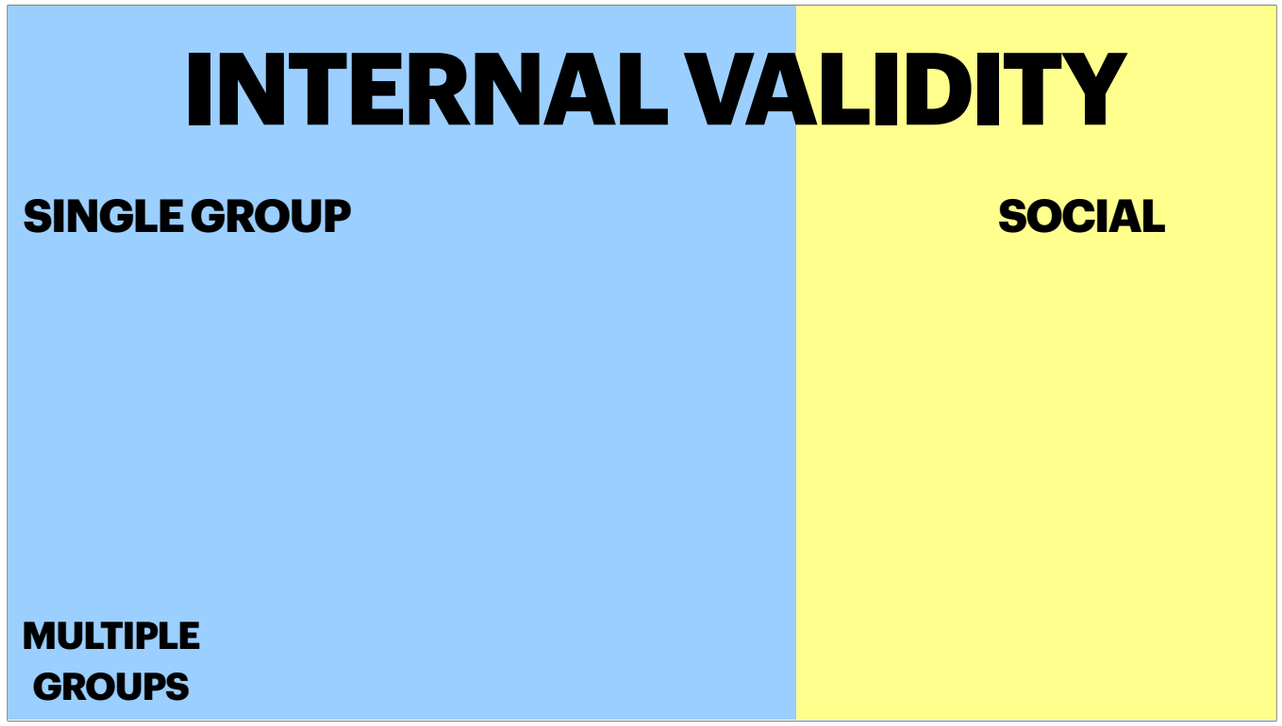
Reliability of measures

Reliability of treatment implementation

Reliability of measures. The validity of an experiment is highly dependent on the reliability of the measures. This in turn may depend on many different factors, like poor question wording, bad instrumentation or bad instrument layout. The basic principle is that when you measure a phenomenon twice, the outcome shall be the same. For example, lines of code are more reliable than function points since it does not involve human judgement.
Reliability of treatment implementation. The implementation of the treatment means the application of treatments to subjects. There is a risk that the implementation is not similar between different persons applying the treatment or between different occasions. The implementation should hence be as standard as possible over different subjects and occasions.



Random irrelevancies in experimental setting. Elements outside the experimental setting may disturb the results, such as noise outside the room or a sudden interrupt in the experiment.
Random heterogeneity of subjects. There is always heterogeneity in a study group. If the group is very heterogeneous, there is a risk that the variation due to individual differences is larger than due to the treatment. Choosing more homogeneous groups will on the other hand affect the external validity, see below. For example, an experiment with undergraduate students reduces the heterogeneity, since they have more similar knowledge and background, but also reduces the external validity of the experiment, since the subjects are not selected from a general enough population.

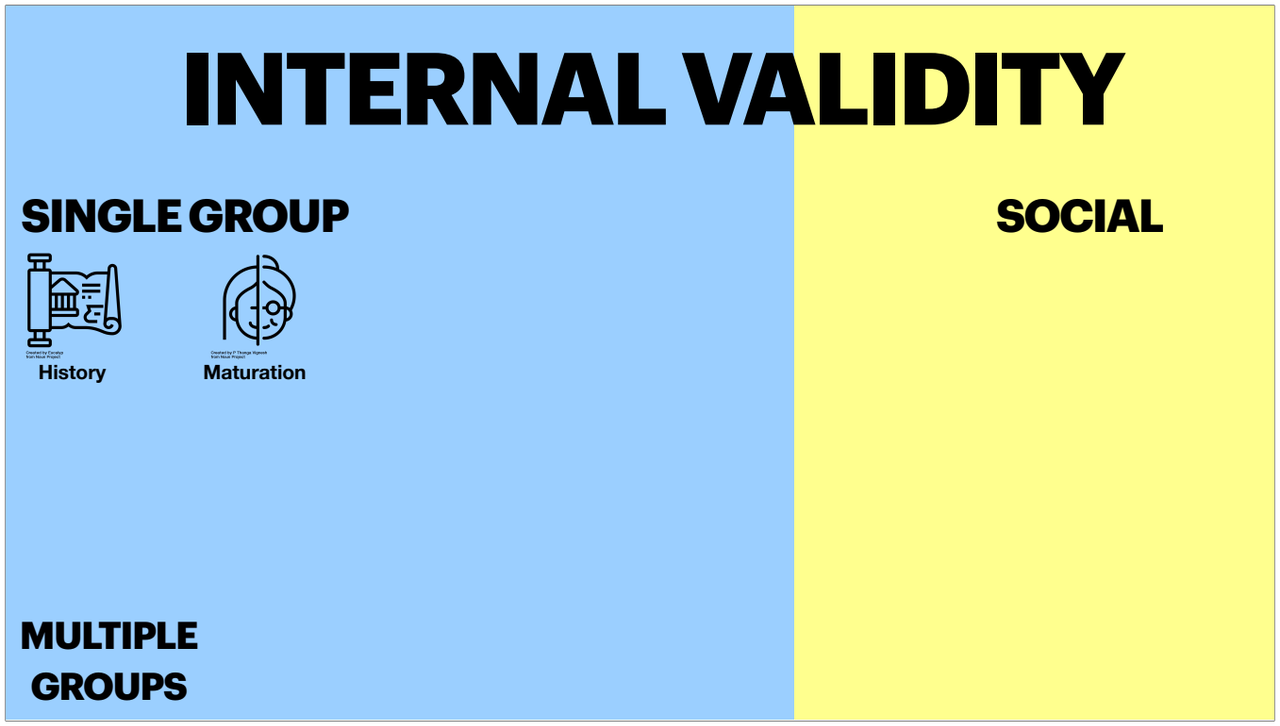


Threats to **internal** validity are influences that can affect the independent variable with respect to causality, without the researcher's knowledge. Thus they threat the conclusion about a possible causal relationship between treatment and outcome.

Single group threats. These threats apply to experiments with single groups. We have no control group to which we do not apply the treatment. Hence, there are problems in determining if the treatment or another factor caused the observed effect.

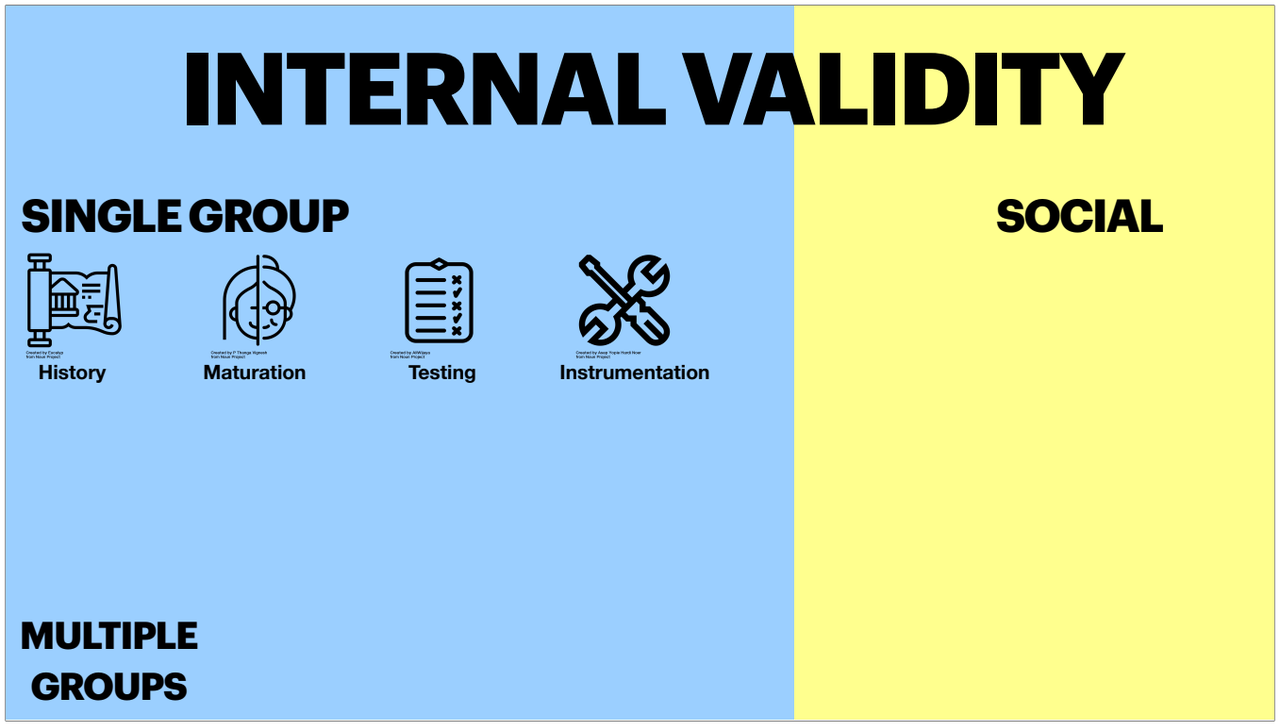
Multiple groups threats. In a multiple groups experiment, different groups are studied. The threat to such studies is that the control group and the selected experiment groups may be affected differently by the single group threats as defined above. Thus there are interactions with the selection.

Social threats to internal validity. These threats are applicable to single group and multiple group experiments, and are related to human participants being autonomous actors and, hence, not necessarily behaving as the researchers would expect.



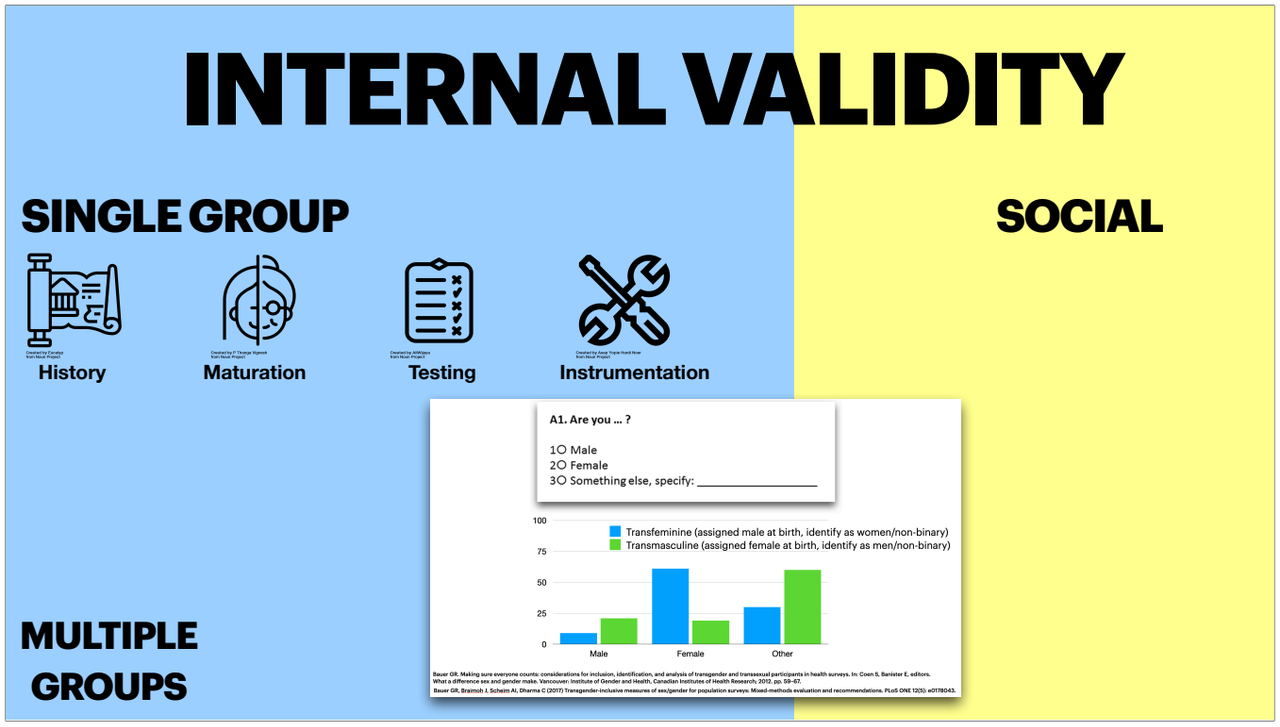
History. In an experiment, different treatments may be applied to the same object at different times. Then there is a risk that the history affects the experimental results, since the circumstances are not the same on both occasions. For example if one of the experiment occasions is on the first day after a holiday or on a day when a very rare event takes place, and the other occasion is on a normal day.

Maturation. This is the effect of that the subjects react differently as time passes. Examples are when the subjects are affected negatively (tired or bored) during the experiment, or positively (learning) during the course of the experiment. One way to counteract the maturation effects would be to perform so called counter-balancing when group 1 first performs task 1 and then task 2, and group 2 first performs task 2 and then task 1 with groups being very similar.

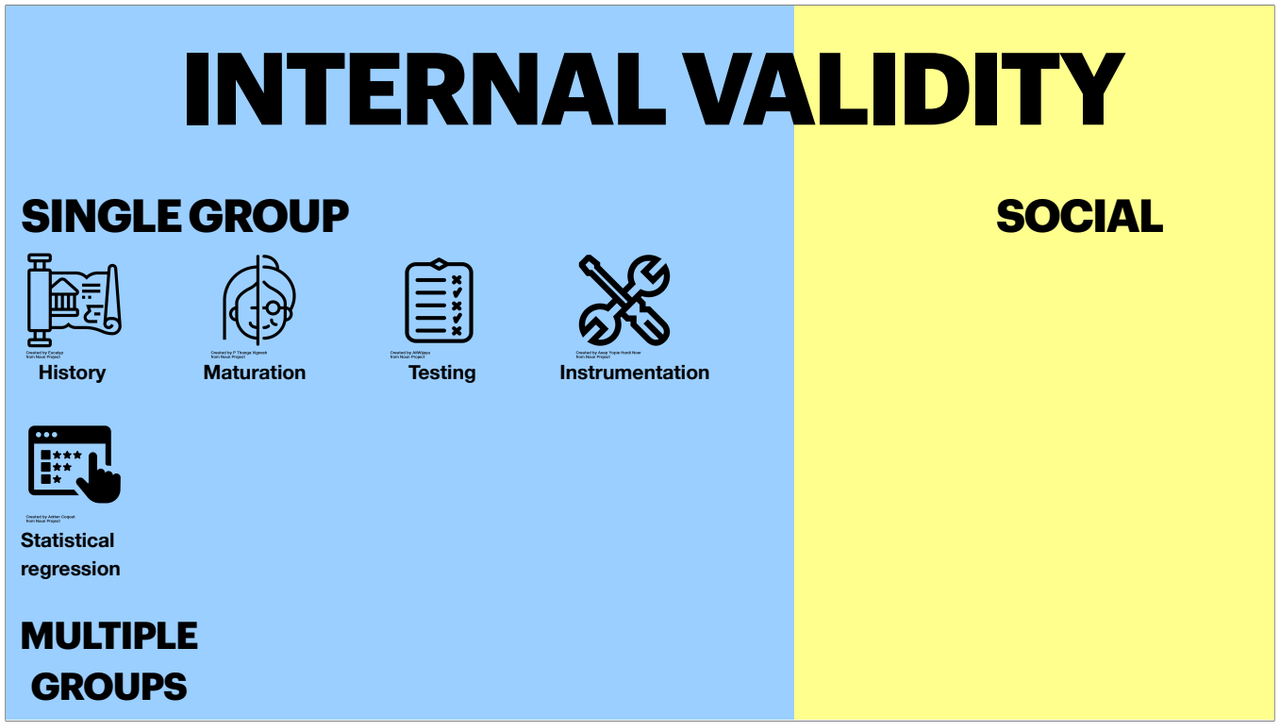


Testing. If the test is repeated, the subjects may respond differently at different times since they know how the test is conducted. If there is a need for familiarisation to the tests, it is important that the results of the test are not fed back to the subject, in order not to support unintended learning.

Instrumentation. This is the effect caused by the artefacts used for experiment execution, such as data collection forms, document to be inspected in an inspection experiment etc. If these are badly designed, the experiment is affected negatively. You might recall that when we have been discussing surveys we have talked extensively about how one can ask about gender and experience.

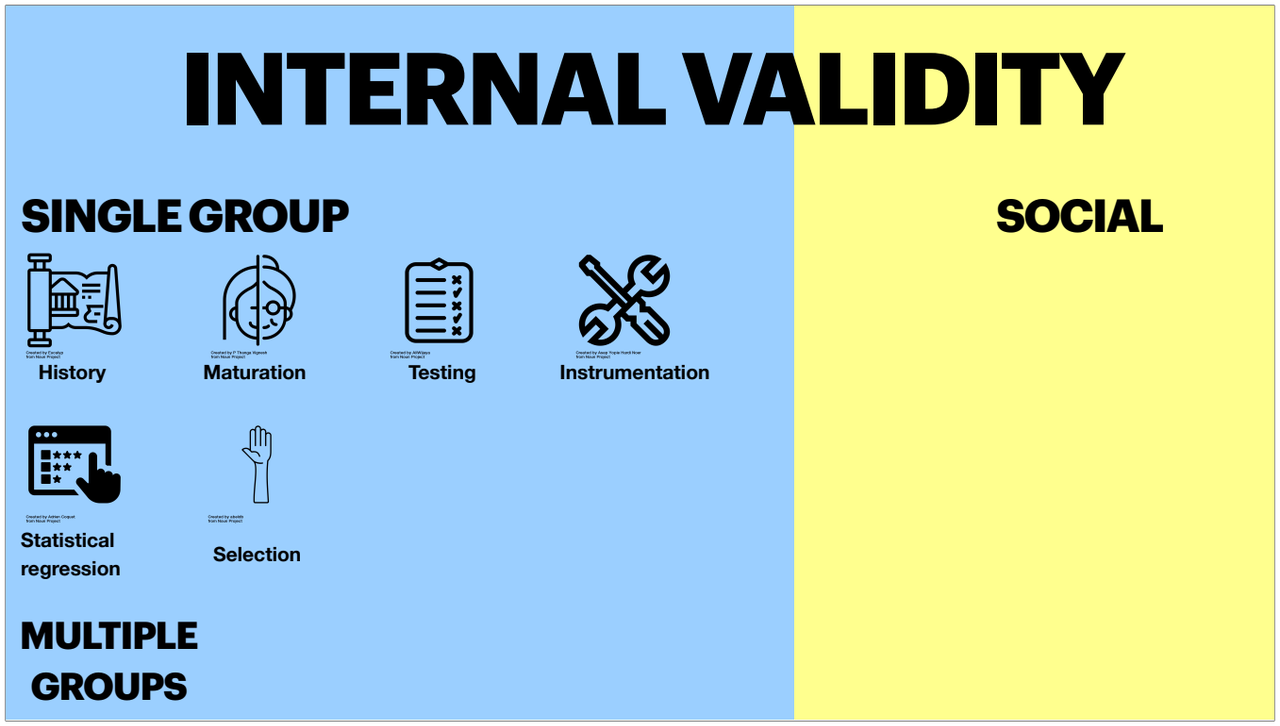


You might recall that when we have been discussing surveys we have talked extensively about how one can ask about gender and experience. One of the possible questions phrasings was proposed by Greta Bauer - see the slide. However, while this question was clear and easily answered by cisgender participants, it did not clearly identify birth-assigned sex or gender identity. In the interviews this item was cognitively taxing for trans interview participants, who tried to figure out exactly what the researchers were asking, and reached different conclusions. This is an example of problematic choice of an instrument affecting internal validity of the study.

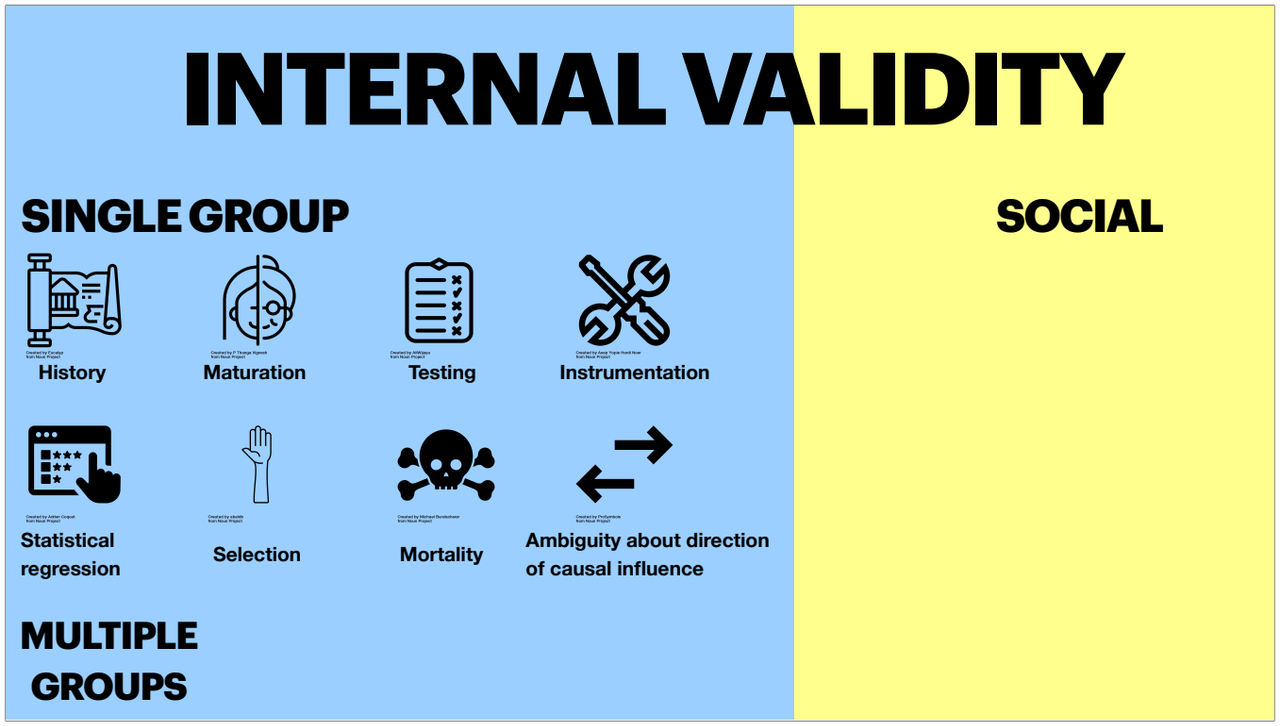


The next threat is related to statistical regressionn. We are still talking about a single group.

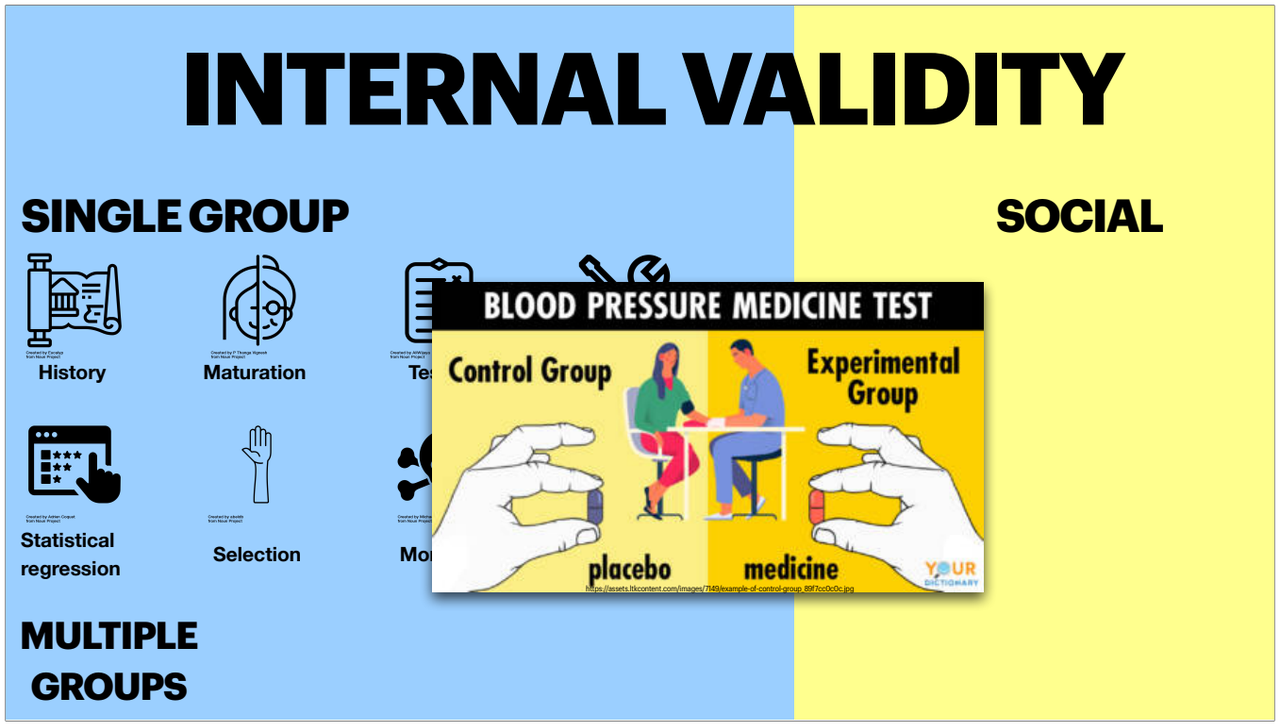
Statistical regression is a threat when the subjects are classified into experimental groups based on a previous experiment or case study, for example top-ten or bottom-ten. Think for example, about participants that have been given a software development task, then they have been subject to a training, and then another development task has been given. In this case there might be an increase or improvement, even if no treatment is applied at all. For example if the bottom-ten in an experiment are selected as subjects based on a previous experiment, all of them will probably not be among the bottom-ten in the new experiment due to pure random variation. The bottom-ten cannot be worse than remain among the bottom-ten, and hence the only possible change is to the better, relatively the larger population from which they are selected.



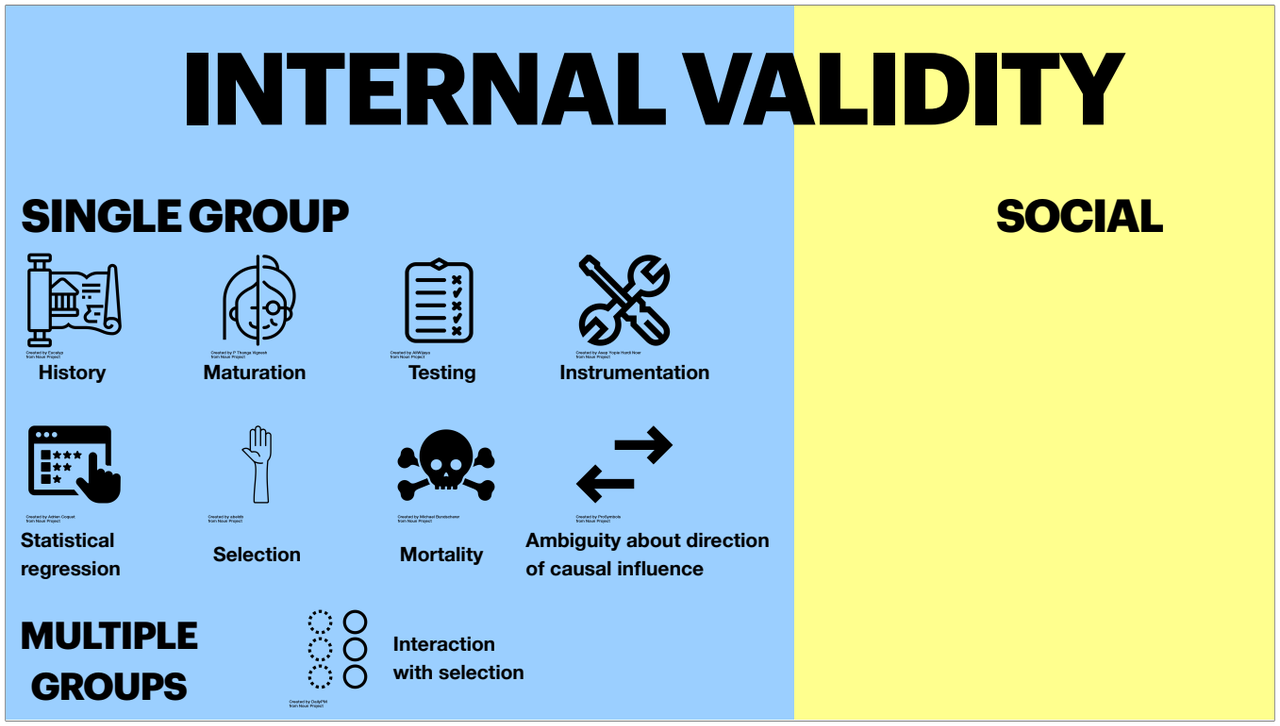
Selection. This is the effect of natural variation in human performance. Depending on how the subjects are selected from a larger group, the selection effects can vary. Furthermore, the effect of letting volunteers take part in an experiment may influence the results. Volunteers are generally more motivated and suited for a new task than the whole population. Hence the selected group is not representative for
the whole population. This is a typical threat for interviews/surveys in the open source world since the participants are volunteering to participate, and researchers usually have no insights in opinions of non-respondents.



Mortality. This effect is due to the different kinds of persons who drop out from the experiment. It is important to characterize the dropouts in order to check if they are representative of the total sample. If subjects of a specific category drop out, for example, all the senior reviewers, the validity of the experiment is highly affected.
Ambiguity about direction of causal influence. This is the question of whether A causes B, B causes A or even X causes A and B. An example is if a correlation between source code complexity and error rate is observed. The question is if high source code complexity causes high error rate, or vice versa, or if high complexity of the problem to be solved causes both.

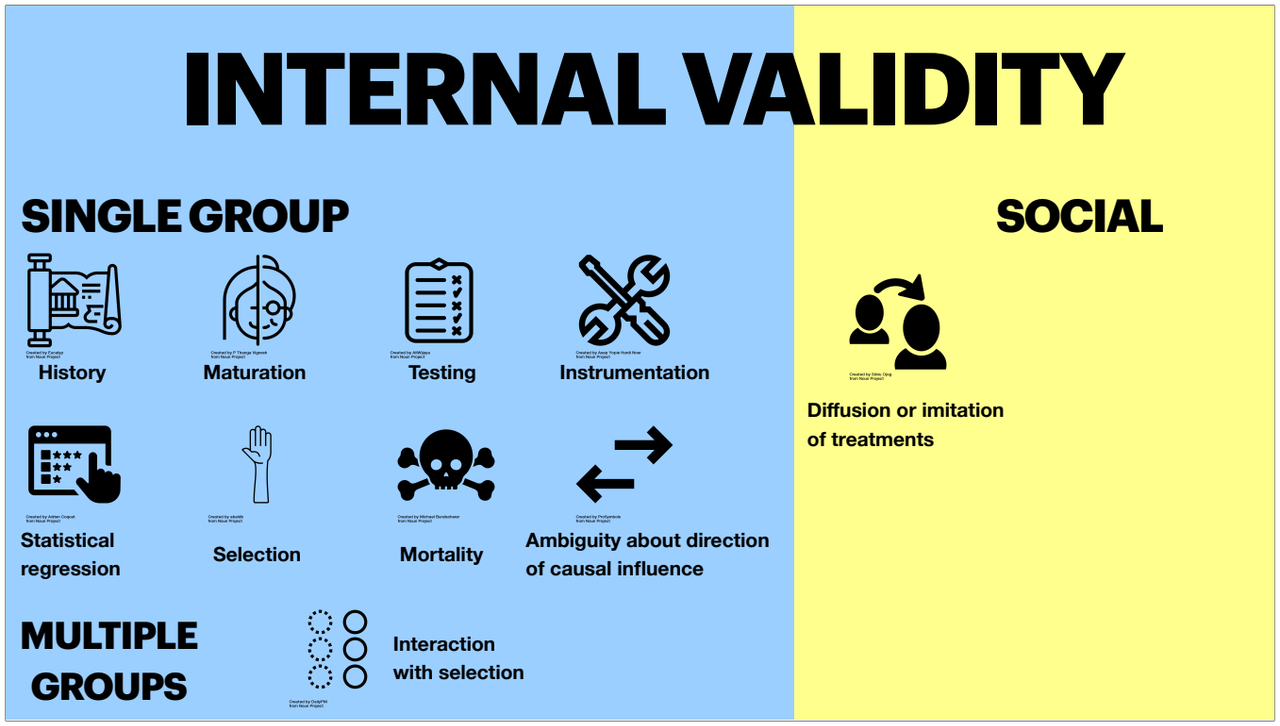


Most of the threats to internal validity can be addressed through the experiment design: for example, counter-balancing helps to address the maturation threat and random selection of interviewees/survey participants to address the threats related to selection. Alternatively, we can introduce a control group, i.e., a group that is not subject to treatment: we can control what happens with the top/bottom 10% (statistical regression). However, introduction of a control group might induce multiple group threats and social threats we are going to discuss next.



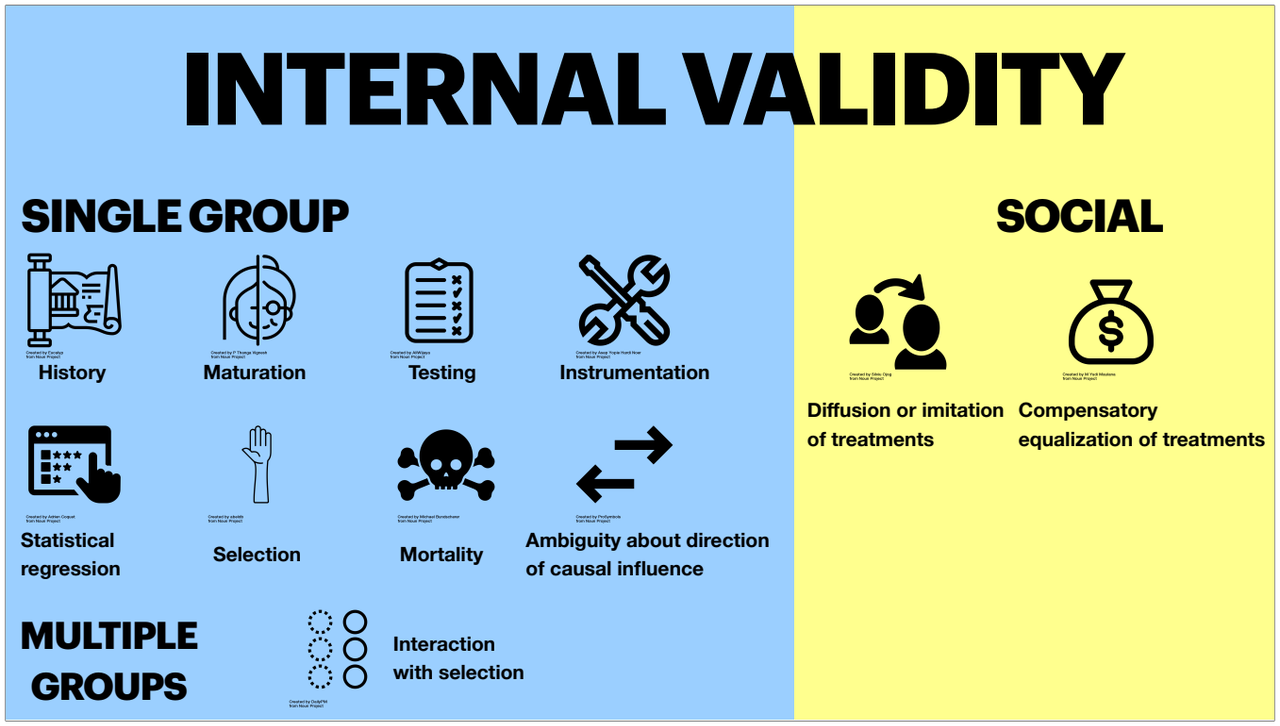
In a multiple groups experiment, different groups are studied. The threat to such studies is that the control group and the selected experiment groups may be affected differently by the single group threats as defined above. Thus there are interactions with the selection.

The interactions with selection are due to **different behavior in different groups**. For example, the selection-maturation interaction means that different groups mature at different speed, for example if two groups apply one new method each. If one group learns its new method faster than the other, due to its learning ability, does, the selected groups mature differently. Selection-history means that different groups are affected by history differently, etc.

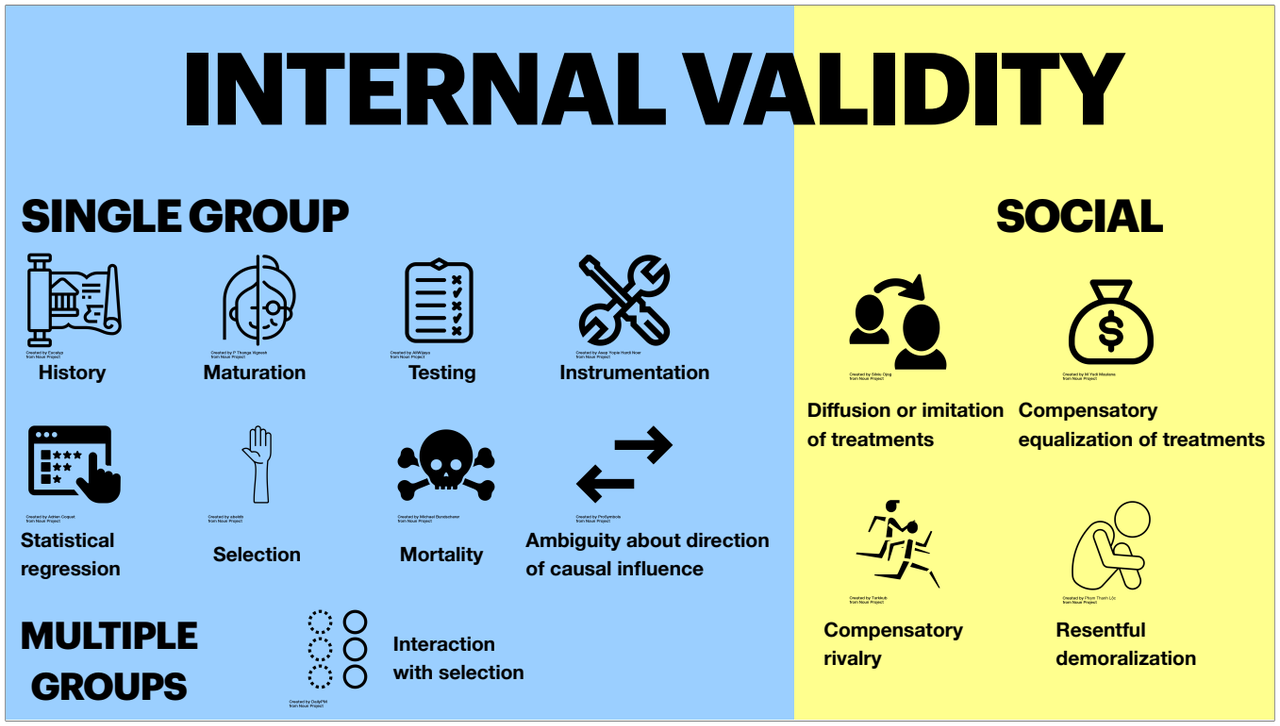


Finally, we need to keep in mind that human participants have their own free will and that they do not necessarily behave in the ways researchers expect. In particular, this is the case if multiple groups are studied.

Diffusion or imitation of treatments. This effect occurs when a control group learns about the treatment from the group in the experiment study or they try to imitate the behaviour of the group in the study. For example, if a control group uses a traditional software development approach and the experiment group uses test-driven development, the former group may hear about the test-driven development and “just decide” to start with writing tests.



Compensatory equalization of treatments. If a control group is given compensation for being a control group, as a substitute for that they do not get treatments; this may affect the outcome of the experiment. If the control group is taught another new method as a compensation for not being taught the experimental method, their performance may be affected by that method. For example, Beatriz Bernárdez has studied whether Software Engineering & Information Systems students enhance their conceptual modelling skills after the continued daily practice of mindfulness during four weeks. The students were divided into two groups: one group practised mindfulness, and the other---the control group---were trained in public speaking. One might wonder whether public speaking helped the participants to structure their thoughts and hence inadvertently improve their conceptual modelling skills.



Compensatory rivalry. A subject receiving less desirable treatments may, as the natural underdog, be motivated to reduce or reverse the expected outcome of the experiment. The group using the traditional method may do their very best to show that the old method is competitive.

Resentful demoralization. This is the opposite of the previous threat. A subject receiving less desirable treatments may give up and not perform as good as it generally does. The group using the traditional method is not motivated to do a good job, while learning something new inspires the group using the new method.

QUESTION

“Depending on their preferences expressed in an interest questionnaire, the participants were divided into two groups: the mindfulness group (G1, 38 subjects) and the control group attending the placebo public speaking workshop (G2, 37 subjects).” This decision induces internal validity threats related to

(A)

Maturation

(B)

Diffusion or imitation of treatments

(C)

Resentful demoralization

(D)

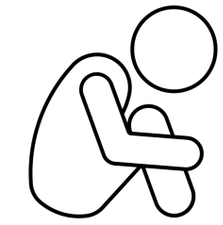
Selection

BEATRIZ BERNÁRDEZ, AMADOR DURÁN, JOSÉ ANTONIO PAREJO, ANTONIO RUIZ CORTÉS: A CONTROLLED EXPERIMENT TO EVALUATE THE EFFECTS OF MINDFULNESS IN SOFTWARE ENGINEERING. ESEM 2014: 17:1-17:10

Correct answer: D. There is no discussion of passing of time (A), or participants talking to each other (B). Moreover, participants have been allocated to the groups based on their interests, so it is unlikely that some of them will be demotivated (C).

QUESTION

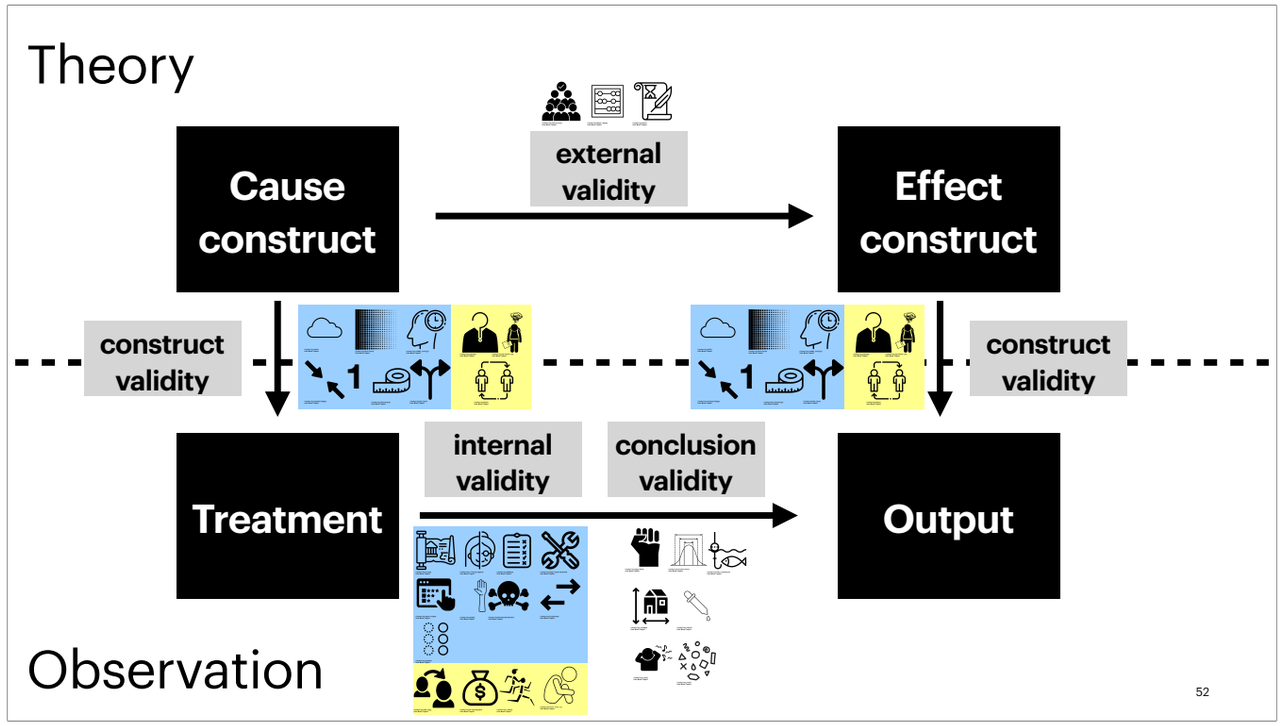
How would you design an experiment to avoid it?



Created by Pham Thanh Lộc
from Noun Project

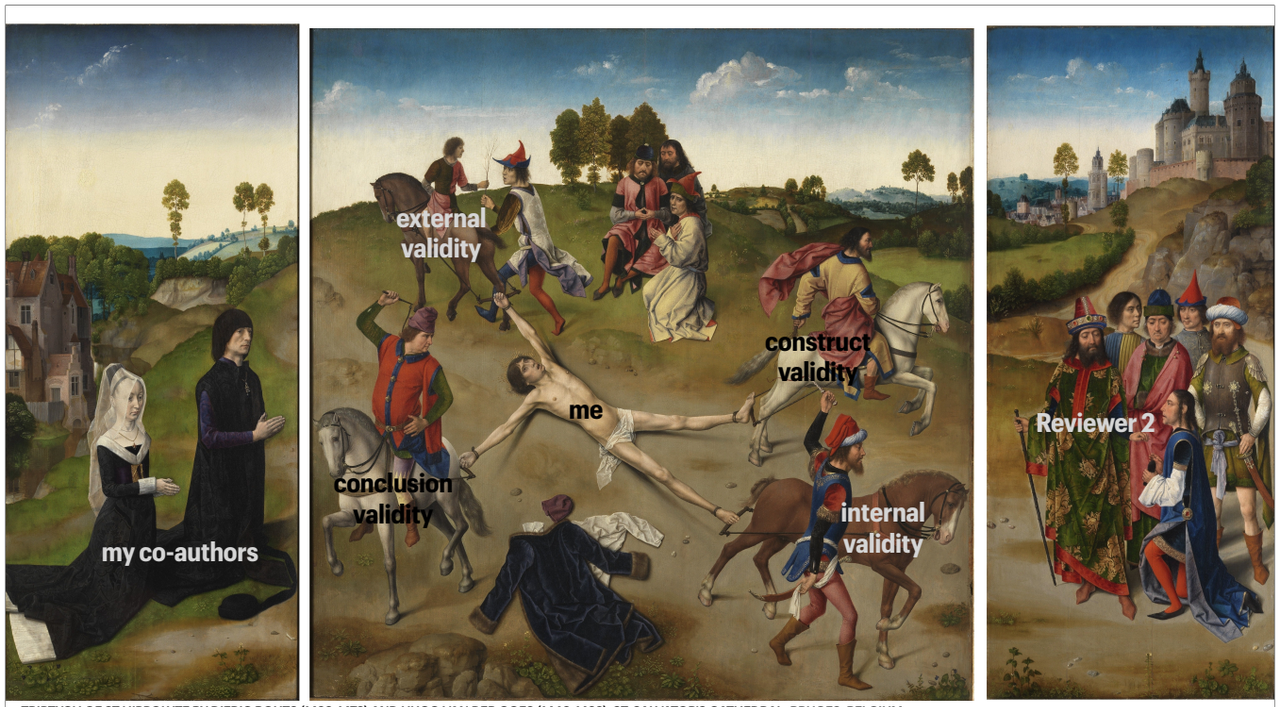
Resentful demoralization

Offer the participants different options with none of them being “obviously” better.
Ensure that the participants do not know what treatment they/others receive (not always possible)



This is an overview of different kinds of threats to validity: blue is related to design, yellow to social.

External: Interaction of selection and treatment, Interaction of setting and treatment, Interaction of history and treatment
Construct validity: Design: Inadequate pre-operational explication of constructs, Mono-operation bias, Mono-method bias, Confounding constructs and levels of constructs, Interaction of different treatments, Interaction of testing and treatment, Restricted generalizability across constructs; Social: Hypothesis guessing, Evaluation apprehension, Experimenter expectancies
Internal validity: Single group: History, Maturation, Testing, Instrumentation, Statistical regression, Selection, Mortality, Ambiguity about direction of causal influence, Multiple groups Interactions with selection, Social Diffusion of imitation of treatments, Compensatory equalization of treatments, Compensatory rivalry, Resentful demoralization
Conclusion validity Statistics: Low statistical power, Violated assumption of statistical tests, Fishing and the error rate, Reliability Reliability of measures, Reliability of treatment implementation, Randomness Random irrelevancies in experimental setting, Random heterogeneity of subjects

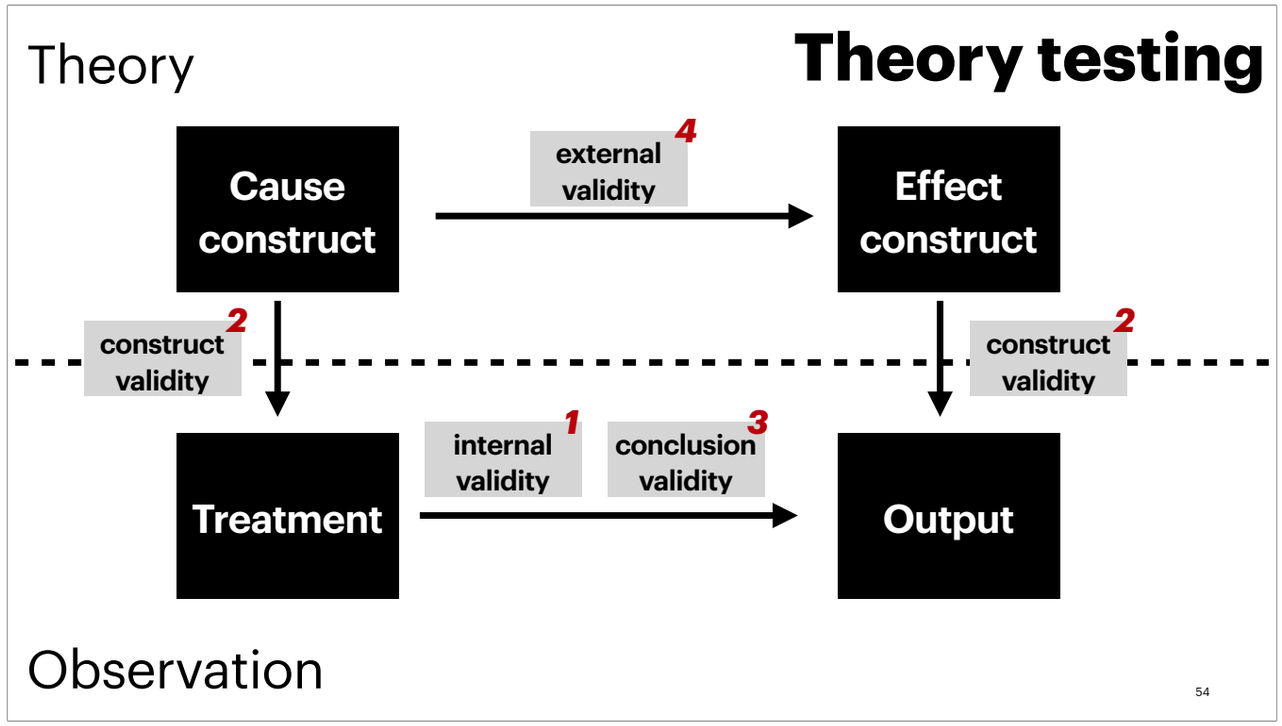


TRIPTYCH OF ST HIPPOLYTE BY DIERIC BOUTS (1420-1472) AND HUGO VAN DER GOES (1440-1482). ST. SALVATOR'S CATHEDRAL. BRUGES. BELGIUM

Ideally, we would like to reduce all kinds of threats.

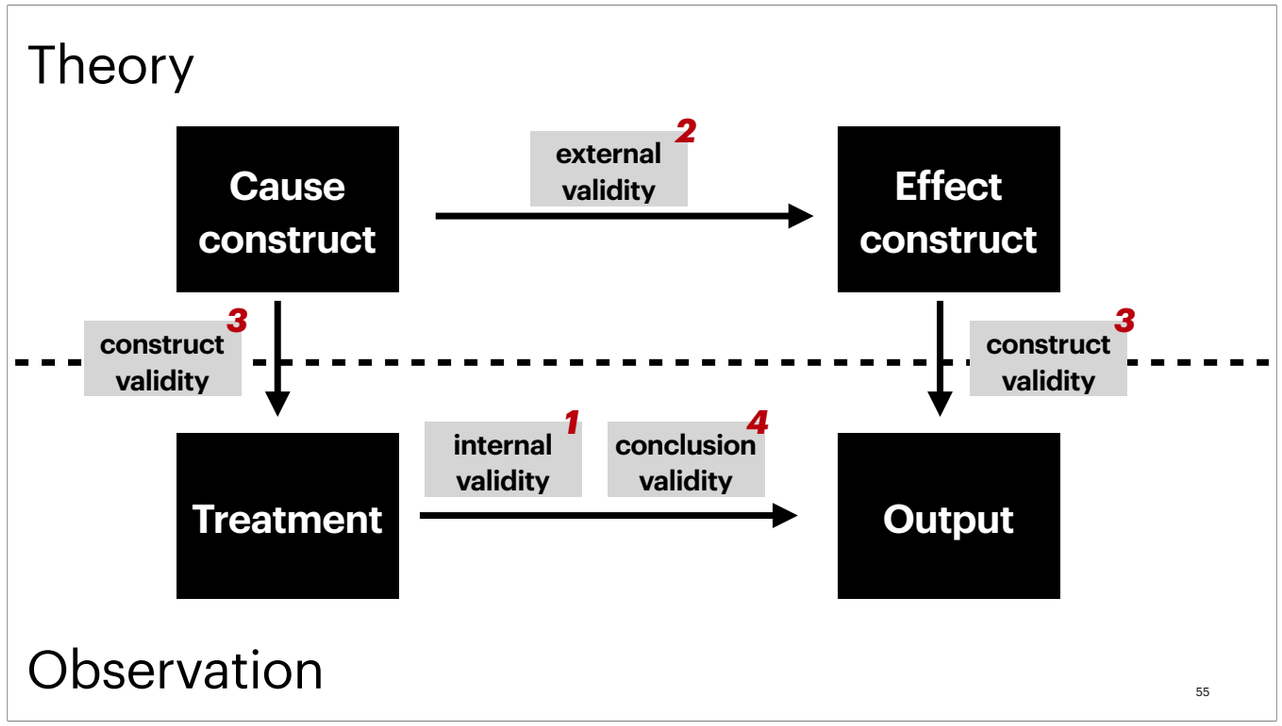
But as we have seen generalisability is needed to ensure external validity, control is needed to ensure internal validity, and generalisability and control don't always go together! RECAP: This is all about the choices and the trade-offs.

We need to make a choice depending on the context of the study.

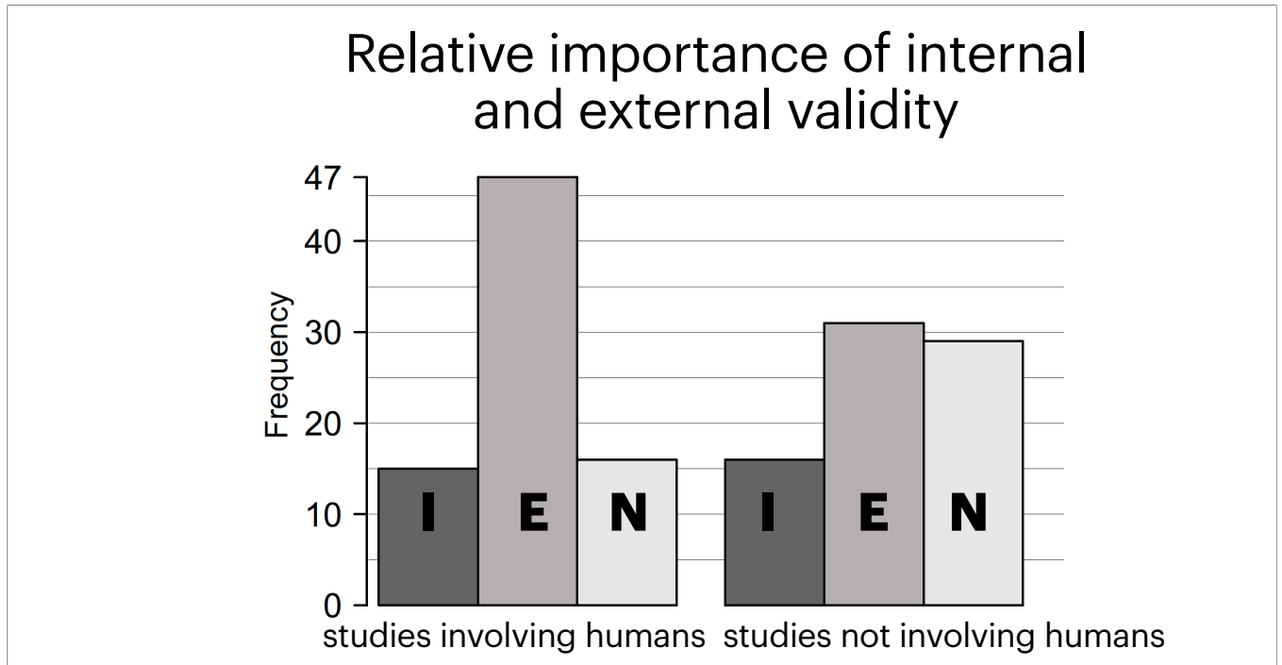


Wohlin et al. recommend different order of importance for different goals.

Theory testing. In theory testing, it is most important to show that there is a casual relationship (internal validity) and that the variables in the experiment represent the constructs of the theory (construct validity). Adding to the experiment size can generally solve the issues of statistical significance (conclusion validity). Theories are seldom related to specific settings, population or times to which the results should be generalized. Hence there is little need for external validity issues. The priorities for experiments in theory testing are in decreasing order: internal, construct, conclusion and external.



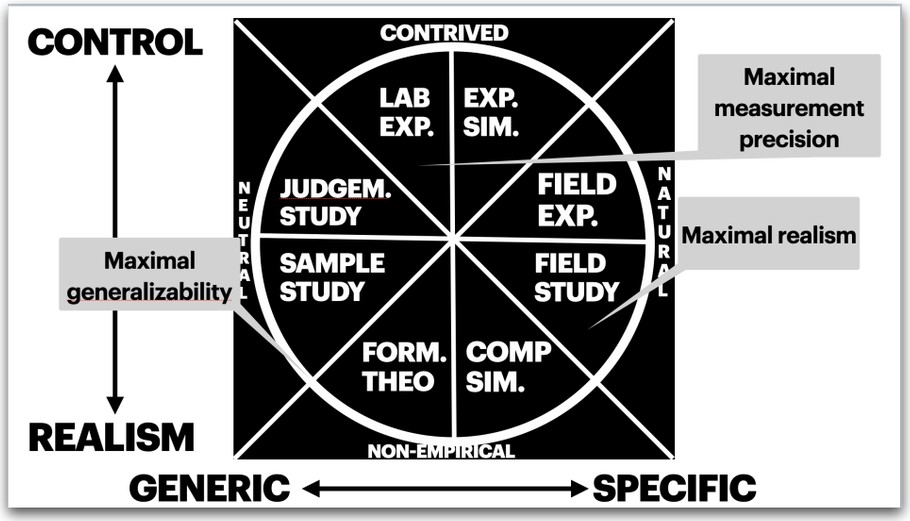
Applied research. In applied research, which is the target area for most of the software engineering experiments, the priorities are different. Again, the relationships under study are of highest priority (internal validity) since the key goal of the experiment is to study relationships between causes and effects. In applied research, the generalization – from the context in which the experiment is conducted to a wider context – is of high priority (external validity). For a researcher, it is not so interesting to show a particular result for company X, but rather that the result is valid for companies of a particular size or application domain. Third, according to Wohlin et al. the applied researcher is relatively less interested in which of the components in a complex treatment that really causes the effect (construct validity). For example, in a reading experiment, it is not so interesting to know if it is the increased understanding in general by the reviewer, or it is the specific reading procedure that helps the readers to find more faults. The main interest is in the effect itself. Finally, in practical settings it is hard to get sufficient size of data sets, hence the statistical conclusions may be drawn with less significance (conclusion validity). The priorities for experiments in applied research are in decreasing order: internal, external, construct and conclusions.



JANET SIEGMUND, NORBERT SIEGMUND, SVEN APEL: VIEWS ON INTERNAL AND EXTERNAL VALIDITY IN EMPIRICAL SOFTWARE ENGINEERING. ICSE (1) 2015: 9-19

The importance of external validity is also highlighted by Janet Siegmund and her co-authors. The letters mean that according to the survey respondents Internal validity is more important than external, External validity is more important than internal, or they do Not have a preference.

“Leave the ivory tower. If actual insights for people’s lives are supposed to be the outcome of research, it better be applied to such problems.” These and further statements indicate that external validity and practical relevance are seen as equivalent. However, this is not entirely true: some studies are inherently non-generalisable.



Do you recall this image? Different methods are associated with different validity threats. For example, maximal precision of laboratory experiments can reduce threats to internal validity, while sample studies aiming at generalisability can reduce threats to external validity.

Since each research strategy has its own associated threats to validity, a better way would be to combine different strategies.

In this paper, we use a mixed-methods approach to identify contribution barriers females face in online communities.

Through 22 semi-structured interviews with a spectrum of female users ranging from non-contributors to a top 100 ranked user of all time, we identified 14 barriers preventing them from contributing to Stack Overflow.

We then conducted a survey with 1470 female and male developers to confirm which barriers are gender related or general problems for everyone. Females ranked five barriers significantly higher than males.

DENAE FORD, JUSTIN SMITH, PHILIP J. GUO, CHRIS PARNIN: PARADISE UNPLUGGED: IDENTIFYING BARRIERS FOR FEMALE PARTICIPATION ON STACK OVERFLOW. SIGSOFT FSE 2016: 846-857

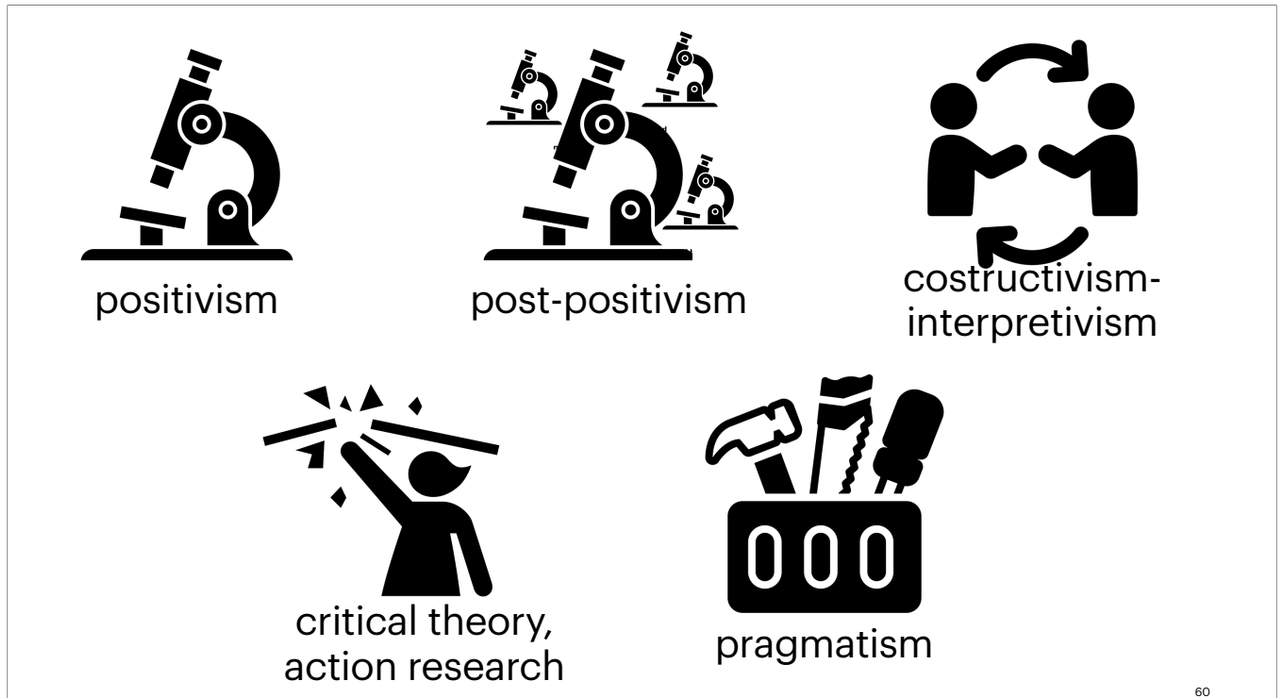
In this paper, we use a mixed-methods approach to identify contribution barriers females face in online communities.

Through 22 semi-structured interviews with a spectrum of female users ranging from non-contributors to a top 100 ranked user of all time, we identified 14 barriers preventing them from contributing to Stack Overflow.

We then conducted a survey with 1470 female and male developers to confirm which barriers are gender related or general problems for everyone. Females ranked five barriers significantly higher than males.

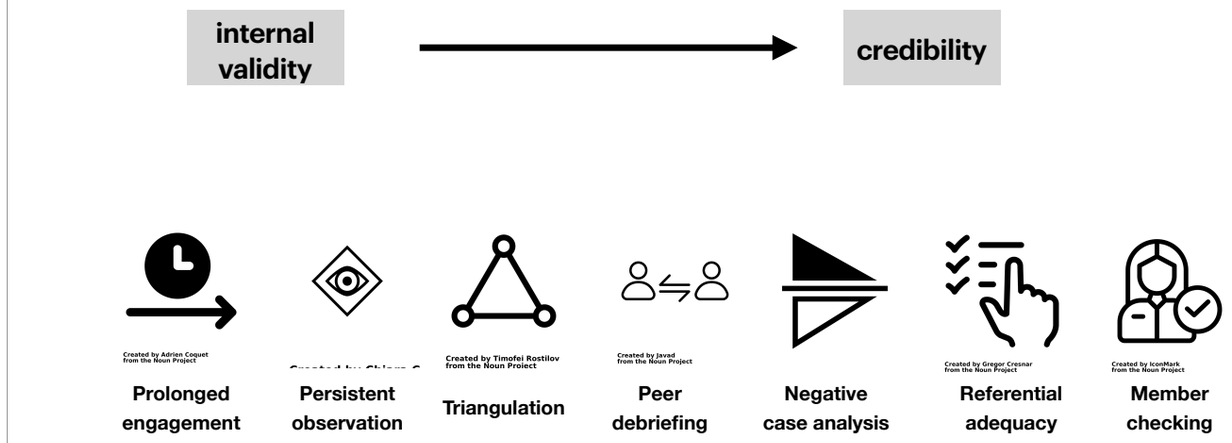
DENAE FORD, JUSTIN SMITH, PHILIP J. GUO, CHRIS PARNIN: PARADISE UNPLUGGED: IDENTIFYING BARRIERS FOR FEMALE PARTICIPATION ON STACK OVERFLOW. SIGSOFT FSE 2016: 846-857

The fragment on the green background focuses on interviews: they can be associated with high internal and conclusion validity. However, the generalisability - or external validity - of this step is low. We do not know whether the barriers identified by means of interviews are shared by other people, and this is why the authors conducted a follow up study on the red background.



Do you remember?

The previous discussion is strongly (post-)positivist in nature. For example, discussion of construct validity is based on the assumption that the underlying construct can be measured but our measurement instruments are imperfect; discussion of threats to external validity reflects the desire to generalise the findings while constructivism-interpretivism stresses that the reality is built based on the interpretation from those that experience it and that there are multiple and equally valid realities. How would one expect findings to generalise to all these realities?



YVONNE S. LINCOLN, EGON G. GUBA. NATURALISTIC INQUIRY. CHAPTER 11. ESTABLISHING TRUSTWORTHINESS

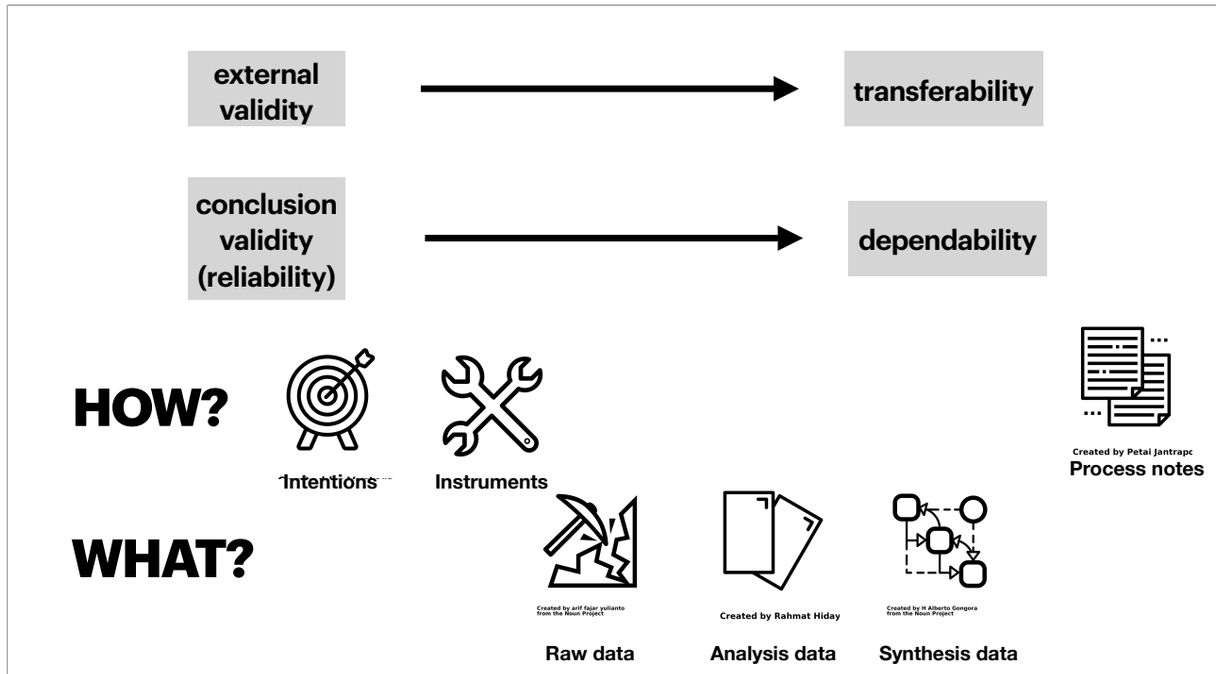
Of course, this does not mean that any qualitative study produces valid results or no qualitative studies can produce valid results. Rather this means that we need to look at different criteria. For example, instead of internal validity we need to consider credibility. To ensure credibility the researcher has “to carry out the inquiry in such a way that the probability that the findings will be found to be credible is enhanced and, second, to demonstrate the credibility of the findings by having them approved by the constructors of the multiple realities being studied.”

* prolonged engagement - to ensure that the researcher understands the community they are studying. For example, when we have started studying Stack Overflow and devRant I have registered on these platforms and tried to hang out there for some time. Threat: “going native”
* persistent observation - it is not enough to merely be there, but one has to be acutely aware of what is going on.

“We shall suggest five major techniques: activities that make it more likely that credible findings and interpretations will be produced (prolonged engagement, persistent observation, and triangulation); an activity that provides an external check on the inquiry process (peer debriefing); an activity aimed at refining -working hypotheses as more and more information becomes available (negative case analysis); an activity that makes possible checking preliminary findings and interpretations against archived "raw data" (referential adequacy); and an activity providing for the direct test of findings and interpretations with the human sources from which they have come-the constructors of the multiple realities being studied (member checking).”

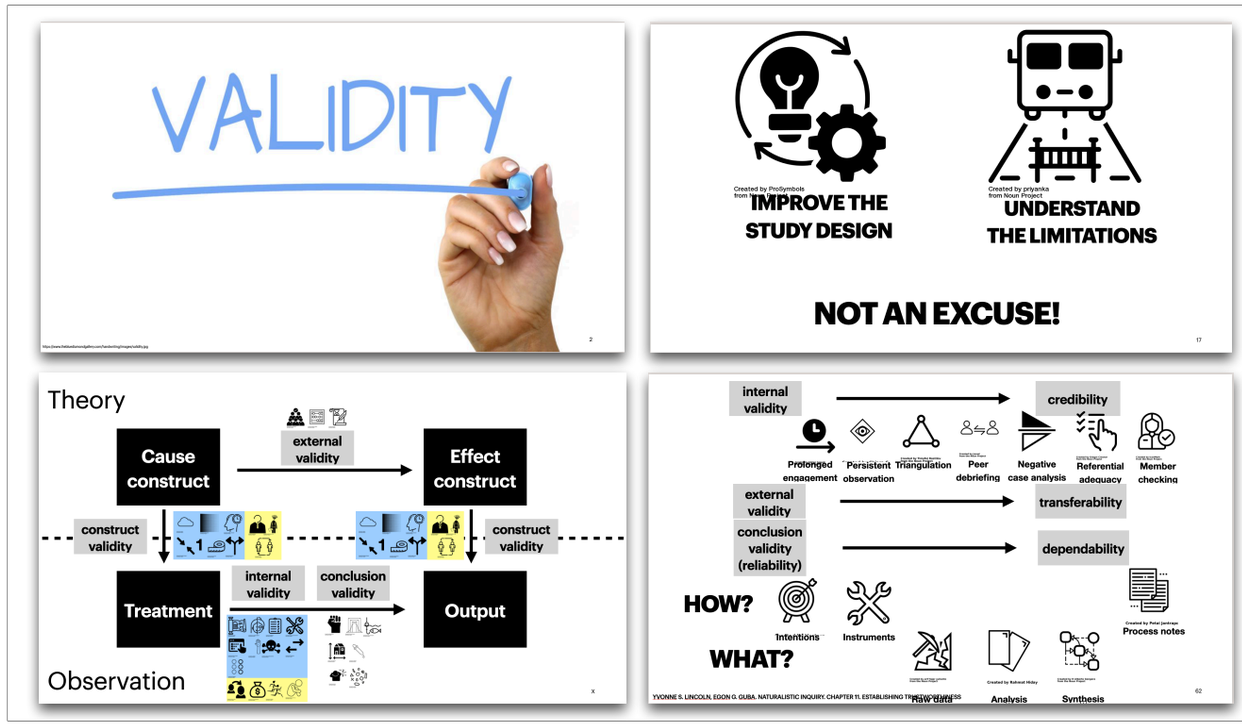
YVONNE S. LINCOLN, EGON G. GUBA. NATURALISTIC INQUIRY. CHAPTER 11. ESTABLISHING TRUSTWORTHINESS

The primary difference between external validity and transferability is in the focus. External validity aims at generalisability for *any* possible scenario. For constructivists this is impossible since there might be multiple realities and they are equally valid. Hence, it is highly likely that the insights cannot hold universally. At the same time, it make sense to provide detailed description of the process (so called *thick description*) such that the person who might be interested in transferring the study insights deciding whether they might be applicable to their situation.



When discussing conclusion validity we have seen three elements: statistics, randomness and reliability. Statistics is not applicable since we focus on qualitative analysis; randomness as in random interferences is not seen as a problem; reliability becomes dependability, i.e., ability to audit the process. To this end the same techniques can be used as before: peer debriefing, triangulation, thick descriptions as well as the **audit trail**. Audit trail should support the full chain reasoning from the “raw” data that has been collected through data reduction and analysis products (such as brief summaries), data reconstruction and synthesis products (such as theory models), process notes (codebooks and examples), materials relating to intentions and dispositions, including the inquiry proposal; instrument development.

We start with intentions, based on the intentions we develop such instruments as interview guides, survey questionnaires or repository mining tools, then we obtain the raw data, analyse it, and synthesise it. All the way through we keep the process notes to record what has been done.



To conclude the last technical lecture of the course… let us summarise what we have talked about today. First of all, whatever data collection and data analysis technique you might use, it is important to think about validity of the conclusions and what issues might have threatened it. This can be used both to improve the study design and to understand its limitations. Remember that reflection on the limitations of the study is not an excuse not to do a proper job!
If we are conducting a quantitative investigation we have the framework of threats to validity including construct validity, internal validity, external validity and conclusion validity. If we are conducting a qualitative study, then we need to think about credibility, transferability and dependability, and there are several ways of ensuring these properties of the study.