

## Internal and External Validity

Martin A. Kozloff

**Extraneous variables are variables that are not a *planned* part of an intervention** (e.g., a change in curriculum or instructional methods) whose effects are being tested. Extraneous variables may “interact with” independent (input, intervention) variables to produce an effect; or extraneous variables may produce an effect by themselves. Therefore, change (or lack of change) in dependent (outcome) variables (e.g., reading achievement) may be entirely or partly the result of extraneous variables, such as maturation, or other things happening outside of school (e.g., siblings teach some students to read) or measurement error (students appear to read better because observers at the outcome assessment failed to count many errors) or bias in selection (e.g., if the experimental group has many bright students and the control group doesn't, that difference---and not the curriculum---may account for differences in achievement). **Findings and conclusions are not credible or believable if researchers can't rule out strong possibility that OTHER factors account for findings.**

## Internal and External Validity

**"Internal validity"** refers to how accurately the data and the conclusions drawn from the data (e.g., Change in X causes change in Y) represent what really happened. For example, looking at pre-test and post-test scores, it may seem that a training program increased teachers' skills. However, some of the difference between pre- and post-test scores may be the result of **measurement error** (during the post-test, observers wrongly scored some sloppy teaching as “proficient”), of the effects of variables outside of the training program that affected some of the trainees **during** the training.

**"External validity"** refers to how accurately the data and your conclusions drawn from the data (e.g., Change in X causes change in Y) represent what goes on in **the larger population**. For instance, if a sample of teacher-trainees is biased in

some way (e.g., the sample contains a higher proportion of motivated trainees than is found in the general population of potential teacher-trainees), then findings from the sample may not be applicable to the general population.

Note that findings and inferences may have internal validity but not external validity. That is, findings and conclusions may accurately represent what was found in the sample studied, but may not apply to **other** samples. However, if findings and conclusions do not have **internal** validity, then they surely don't have external validity either.

The factors that can weaken internal and external validity are called "**extraneous variables.**" Maturation of study participants is an example. Change in children's skills during instruction may reflect maturation of the nervous system and muscles as well as the effects of instruction. So, if the research hypothesis is that instruction will increase children's skills, the "**plausible rival hypothesis**" is that maturation will increase children's skills. That's why it is important to identify possible extraneous variables (sources of "contamination"). You can then design research to weaken or eliminate the effects of these variables, or you can analyze the data to determine what effect the extraneous variables have had. For example, if you use an experimental and control group, and if you create the two groups using the method of **random allocation**, then you weaken the rival hypothesis of maturation (since children in both groups have an equal chance of improving as a result of maturation).

### **Extraneous Variables That Are Threats to *Internal* Validity**

**1. Instruments do not measure what they purport to measure.** In other words, the findings are not valid. For instance,

a. The dependent (outcome) variable is reading proficiency. However, that is NOT what the researcher is measuring. Instead, the researcher is measuring behavior such as turning pages, naming parts of a book, holding books properly, memorizing words, and guessing what words say. If the researcher is "testing" a method that teaches children to memorize and guess at words, then the method will appear to

be effective---but only **because the researcher is not measuring reading at all.** To prevent this, researchers must either use standardized validated methods and instruments, or must carefully define variables, and then develop measures based on these definitions.

- b. **The measurement method or instrument has *not* been tested for reliability;** that is, different observers or testers observing the same thing would get the same scores. If a method or instrument is NOT of known high reliability, then it is possible that **what appears to be high achievement in a group receiving an intervention is because the post-test scores were *wrong*.**

Therefore, researchers should use methods and instruments with known reliability, and should ensure that observers and testers produce reliable data before a study begins, and periodically during a study if repeated measurement is used.

- c. **Data that *should* be OBJECTIVE** (e.g., how often teachers properly correct student errors) **are subjective---opinions, impressions.** These data (“I learned a lot!” “Training was excellent.” “I am confident that I can properly teach the five reading skills.”) can’t be used to determine if change in input variables (e.g., teachers receive training in how to teach reading) is associated with/ followed by changes in outcome variables (such as increased teaching proficiency. Why? Because the opinions do not measure proficiency; they measure feelings. Also, opinions and feelings and impressions change—and therefore are not reliable indicators of hard facts of proficiency.

**2. History.** History includes events in **addition** to the independent variables under study, that occur **between** one measurement and another (e.g., between a pre-test and post-test). For example, in testing the effects of an exercise program on psychological well-being following heart attack, some participants joined a church, or received additional social support, or changed jobs. These extraneous (history) variables may account for some of the differences between pre- and post-test scores.

Ways to weaken history as a rival hypothesis include using equivalent experimental and control groups (created by random allocation or matching). Since

the groups are, logically, likely to have the same historical variables happening between pre- and post-test, differences in the outcomes are NOT likely to be the result of history.

**3. Maturation.** Maturation refers to changes that ordinarily **occur with time** (e.g., strength, increasing knowledge). For instance, in an experimental intervention to decrease children's hyperactivity, some of the change in some of the children could reflect increased capacity to pay attention as a result of maturation of the nervous system. The rival hypothesis of maturation may be weakened by using equivalent comparison groups, or by using experimental designs in which the experimental group serves as its own control (e.g., the interrupted time-series design).

**4. Testing.** This refers to the effects of taking one test on the results of a later test. For instance, improvement in scores might reflect decreasing fear of being tested, or figuring out what kinds of answers are correct.

Testing can be controlled in part by using versions of the same tests and by using comparison groups in which one group does not receive a pre-test.

**5. Statistical regression.** A person's performance of any task can vary within a certain range. On the average, you may be able to do 10 pull-ups, but on a particular day you may do 8, 9, 11, or 12. In fact, there may be days when your performance is quite **unusual**--you can barely do 5 pull-ups, or somehow you manage to do 18. However, if you did pull-ups the next day, and the day after that, your performance would probably **regress (move) to the mean**, or your average performance.

In research, a group's pre-test performance might (by chance) be **unusually** high or low; some people had a good day or a bad day. On later testing, the group's performance regresses to the mean (i.e., is more usual). The researcher may mistakenly treat differences between pre- and post-test scores as the result of an intervention ("They improved.") or as the failure of an intervention ("They got worse!"), when in fact, the group merely turned in its average or usual performance.

The rival hypothesis of statistical regression can be partly controlled by using equivalent comparison groups, since the possibility of unusual scores applies equally to the groups.

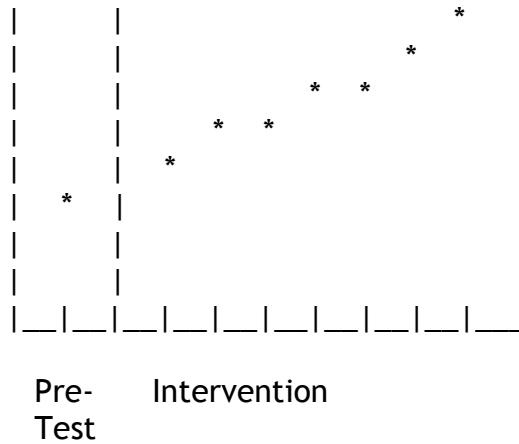
**6. Selection bias.** In research using comparison groups, some participants in one group may be different from those in the other group(s) in ways that affect performance. For instance, an experimental group may do much better on a post-test than the control group, *not* because the experimental intervention was effective but because more of the E group members figured out how to take the test (See number 3 above.). Similarly, the pre-test/post-test differences between the E and C group may be small, suggesting that the intervention did not work. However, in fact, the control group contained many people who **WERE** likely to change as a result of maturation or some historical factor.

This source of invalidity can be handled, in part, by random allocation of participants to comparison groups. This way, all possibly biasing factors have an equal chance of being in both groups.

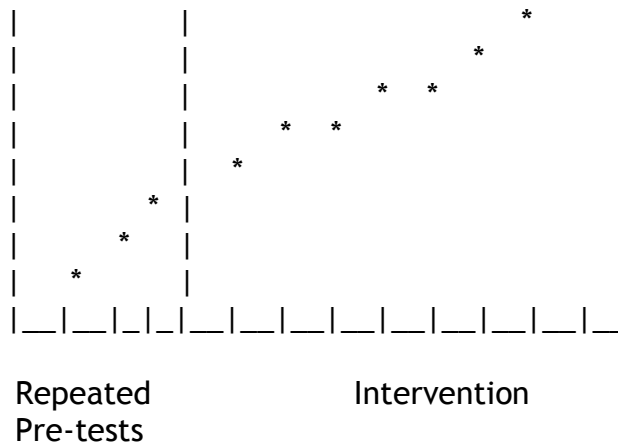
**7. Experimental mortality.** This refers to the differential loss of participants from comparison groups. For example, an experimental intervention may appear to work only because participants with whom it **was not going to work dropped out**. Similarly, an intervention may appear to work no better than nothing at all, only because people in the control group who would have gotten **WORSE** over time dropped out, leaving people in the control group who improved. Thus, the control group scores about the same as the experimental group.

The rival hypothesis of experimental mortality can be partly controlled by using equivalent comparison groups, since the chances of dropping out should be about equal in the two groups.

**9. Causal time order.** Here, participants began to change *prior* to an intervention, **but the researcher does not know this**. It only appears that the intervention is the cause of the change.



This is what is really happening.



A partial solution is an extended series of baseline or pre-intervention observations, to assess the stability of performance before an intervention. If the “baseline” or pre-test scores are stable (mostly a straight line), and scores ONLY rise AFTER the intervention begins, you have evidence that the intervention is having an effect.

**10. Diffusion or imitation.** Here, some part of an intervention given to an experimental group is used by members of the **control group**. Thus, the intervention does not appear to make much of a difference, because both groups have changed.

For example, families in a training program lend materials to friends in the control group.

One way to try to control this is to make sure that members of the comparison groups do not know one another. Another method is not to tell participants what group they are in---a single blind study. However, this may pose ethical problems. Still another method is to use delayed-intervention control groups (so that members of the control group may be more willing to wait).

**11. Compensatory rivalry.** Knowing they are in a control group, some participants try to change on their own. Improvement in the control group may be mistaken as a lack of effect of the intervention.

One method of handling this is NOT to tell participants which group they are in. This is called a “single blind” study.

**12. Demoralization.** Knowing they are in a control group, and not receiving an intervention that they want, some members of the control group look worse over time than they otherwise would. This may result in differences between the E and C group being mistaken for the effects of the intervention. (Imagine the effects on their life expectancy if people with aides knew that they were in the control group of a drug experiment.)

A partial solution is to use a delayed-treatment design (rather than no-treatment design). Also, one could use alternative treatment groups rather than a control group.

### **Extraneous Variables That Are Threats to External Validity**

Keep in mind that all of the threats to internal validity are also threats to external validity. Additional threats to external validity include the following.

**13. Reactive or interactive effects of testing** "Reactive" effects of testing means that a pre-test alone influences post-test performance. "Interactive" effects of testing

means that a pre-test influences how people are affected by an intervention. If the performance of an experimental group after an intervention has been influenced by the pre-test, the findings (e.g., amount of beneficial change resulting from treatment) may not apply to the general population which is **not likely to receive a pre-test**.

Therefore, it may be important to assess the effects of pre-testing. An experimental design called the Solomon Four-group Design is an effort to control this source of invalidity.

**14. Interaction of selection bias and X (intervention)** Here, a bias in the selection of the E group has resulted in enough members of the E group being especially likely to be affected (or not affected) by X, so that the E group's post-test scores are higher (or not higher) than scores of the C group. But since samples in the general population are NOT likely to have this bias, the results of the intervention with other samples may be less than in the experiment.

One way to handle selection bias is to use random sampling so that study samples are equivalent to the general population.

**15. Interactive effects of experimental arrangements.** If the performance of people in an experimental group was affected (positively or negatively) by certain features of the experiment, or by the fact that it was seen by them as an experiment, findings from the E group may not apply to samples from the general population who will receive the intervention in a **nonexperimental** setting. For instance, teachers in an experimental training program (which gives them a sense of being special) may change more than later trainees who simply receive a course on the same material. There is no way getting around this one. The more you control a situation so that you get valid data, the LESS the situation is like real life, and therefore, the results you got in the contrived setting may not happen outside of it. However, you CAN TEST THAT very hypothesis by replicating the research in more and more natural settings, and see if the results remain about the same.

