University of Setif -2-
Faculty of Literature and Languages
Department of English Language and Literature
Introduction to Applied Linguistrics Research
Third Year
Instructor: Mrs Houda Khabcheche

Lecture 7 :

Internal & External Validity
(Controling Extraneous variables)

Introduction :

The central issues in thinking about any type of research depend on whether the research is logical and meaningful . After spending a great deal of time and effort designing a study, we want to make sure that the results of our study are valid. That is, we want them to reflect what we believe they reflect and that they are meaningful in the sense that they have significance not only to the population that was tested, but, at least for most experimental research, to a broader, relevant population. (how true and accurate the measurement is)

There are many types of validity, including content, face, construct, criterion-related, and predictive validity. We deal with each of these in turn after understanding internal and external validity, which are the most common areas of concern.

But you have first to distinguish between :

When evaluating a Study we discuss the Internal Validity & External Validity

When evaluating a Measure discuss the Reliability &Validity

Internal Validity : Refers to the extent to which the changes observed in the DV are caused by the IV. Or As noted by Campbell and Stanley (1963) It has to do with interpreting findings of research within study itself

External Validity : Refers to generalizability or representativeness of the findings. Or As noted by Campbell and Stanley (1963) It has to do with interpreting findings and generalising them beyond the study .

Internal Validity : As explained before , internal validity refers to what extent are the differences that have been found for the dependent variable directly related to the independent variable? A researcher must control for (i.e., rule out) all other possible factors that could potentially account for the results.

Imagine that you wished to replicate the study conducted by Ben-Zeev (1976) , which showed mixed results . In one part of the study she checked to see if bilingual and monolingual children had the same flexibility in recognising that there can be other words for concrete objects such as book , table , cat . In one sample she had 98 Hebrew-English bilingual and English monolingual children . In a second sample she had 188 Spanish-English bilingual and English monolingual children .The bilinguals outperformed the monolinguals in the first sample but in the second sample there was no difference . You want to see what would happen with a new sample .

To start the study , you collect data at a school , carefully checking that school records show which children are bilingual and which are monolingual English speakers . When you were on the

playground , though you notice that some of the «monolingual » children actually speak some spanish with each other . Your data are compromised (threatened) by poor subject selection , a major threat to internal validity . Not all children in the monolingual sample were truly monolingual . So you will not know how much confidence you can place in the results .

It is important to think through a design carefully to eliminate or at least minimize threats to internal validity. some of the most common and important ways include :

Participant Characteristics : The example provided in the previous section concerning true monolingual and bilingual participants is participant characteristic . Let us consider some relevant participant characteristics for second/foreign language research: language background, language learning experience, and proficiency level.

Language Background : In many studies, researchers want to compare one group of students with another group based on different treatments. For example, let us assume that a study on the role of attention in second language learning compared groups of students in a foreign language class who were exposed to a language structure with and without devices to ensure that they paid attention to that structure. It would be important that each group of students be relatively homogeneous. Otherwise , one could not be sure about the source of the results . For instance, let's further assume that one group of students had a large number of participants who were familiar with the target language (either through exposure at home or in the classroom). We then could not distinguish between the effects of the treatment and the effects of the background knowledge of the participants.

Language Learning Experience: Participants come to a language learning situation with a wide range of past experiences which may have importance for research. For example, many students in an ESL setting have had prior English instruction in their home countries, and this prior instruction may differ from one country to another. EFL students come to the university with different language experiences may be because they studied in different streams (science , lieterature , mathematics and foreign languages at high school)

If we wanted to conduct a study in which we compared implicit versus explicit methods of instruction, we might find that a group that received explicit instruction outperformed a group that received implicit instruction. If the two groups also differed in terms of prior learning experiences, we would be left with two variables: learning experience and instruction type. We would not, be able to distinguish between them. Are the results here due to the type of the instruction or to prior language learning experience .

Proficiency Level : In the area of foreign language research, there are some global proficiency measures such as the Oral Proficiency Interview (OPI) so that learners can be matched for proficiency. Or another common measure is to use placement in class level (first year versus second year versus third year, etc.) . In a foreign language environment, exposure is more or less limited to what occurs in the classroom . However, with second language learners, backgrounds and outside experiences are varied and there is typically unevenness (lacking consistency) in skill levels . It is therefore important to consider how this may bear on the specific research questions of the study .

Participant Inattention and Attitude : When we collect data from participants, we usually make the assumption that they are giving us their "best effort." In other words, we rely on the notion that the language data we are collecting are uncontaminated by the experiment itself. This may not always be true. One factor that might affect participant behavior is what is known as the Hawthorne effect, which refers to the positive impact that may occur simply because participants know that they are part of an

experiment and are, therefore, "different" from others. Participants may also try to please the researcher by giving the answers or responses they think are expected. This is known as the Halo effect.

Participating in a study also has potential negative effects. For example, researchers might want to consider factors such as **fatigue** and **boredom** when asking participants to perform tasks.

Whatever method is being used to gather data, one needs to think of the exhaustion and boredom factor. How much time can one reasonably ask a participant to perform without losing confidence in the results, especially if it is a repetitive and demanding task . A second factor is general inattentiveness, whether from the outset of the experiment or as a result of the experiment .

Gass (1994) gave participants the same task after a 1-week interval and noted that some participants provided diametrically opposed responses at the two time periods . one of the participants stated that his results from the two sessions differed because his mind was wandering given that he had two academic tests that week.

The researcher needs to be aware of this as a possible way of explaining what may appear to be divergent results .

Participant Maturation : Maturation is most relevant in longitudinal studies and particularly in those involving children. For example, a study that spans a year or longer will inevitably include participants who change in one way or another in addition to changes in language development. Adults may not change dramatically in a 1-year period, but children certainly do. Moreover, people who were comparable at the outset of the study may change in different ways due to different experiences over time. Thus, one must find a way to balance regular maturational factors against the requirements of the study. When maturation is a consideration, a control group not subjected to the treatment or intervention is appropriate wherever possible. The inclusion of a control group provides one way to test whether any changes occurred because of the experimental treatment or because of maturation.

Data Collection: Location and Collector: Some obvious concerns relate to the physical environment;for example, the environment for two groups given the same test might influence the results if one group is in a noisy or uncomfortable setting and the other is not . Another factor in some types of research relates to the person doing the data collection. For example , in a research concerning families being surveyed about their attitudes toward their children's learning of the target language .

Some obvious concerns relate to the physical environment; for example, the environment for two groups given the same test might influence the results if one group is in a noisy or uncomfortable setting and the other is not .Another factor in some types of research relates to the person doing the data collection. For example , in a research concerning families being surveyed about their attitudes toward their children's learning of the target language .

Instrumentation and Test Effects : The test instrument is quite clearly an important part of many research studies. In this section we discuss three factors that may affect internal validity: equivalence between pre- and posttests, giving the goal of the study away, and test instructions and questions .

The test instrument is quite clearly an important part of many research studies. In this section we discuss three factors that may affect internal validity: equivalence between pre- and posttests, giving the goal of the study away, and test instructions and questions .

Equivalence Between Pre- and Posttests: One serious design issue relates to the comparability of tests. A difficult pretest with an easier post-test will make it more likely for improvement to be apparent after a treatment. The opposite scenario will make it more likely for no improvement to be apparent following a treatment.

Giving the Goal of the Study Away: One of the problems in doing second language research is that one sometimes does not want participants to know the precise nature of the language area or behavior that is being tested. We might want to conceal the precise nature of the study because we want responses that reflect natural behavior rather than what participants think they should say or do (consent forms and how to strike a balance between not being deceptive and yet not revealing precisely what the study's focus is). This becomes particularly problematic when using a pretest because the pretest may in and of itself alert participants to the study's objective. One way of avoiding this problem is by conducting the pretest a few weeks before the study, the idea being that participants will not associate the pretest with the study itself .

Instructions/Questions : The researcher must make sure that the instructions are clear and appropriate to the developmental level of the participants in the study. We cannot rely on responses to questions when it is not clear whether the instructions have been adequately understood. In second language research, the instructions and questions should be appropriate to the level of linguistic and cultural knowledge of those who are taking the test or filling out a questionnaire.

External Validity : All research is conducted within a particular setting and using a specific set of characteristics (e.g., second year L1 English learners of French at X university). However, most quantitative research is concerned with broader implications that go beyond the confines of the research setting and participants. The participants chosen for any study form a research population. With external validity, we are concerned with the generalizability of our findings, or in other words, the extent to which the findings of the study are relevant not only to the research population, but also to the wider population of language learners. It is important to remember that a prerequisite of external validity is internal validity. If a study is not conducted with careful attention to internal validity, it clearly does not make sense to try to generalize the findings to a larger population.

Generalizability is usually defined as the degree to which the results of a study based on a sample can be said to represent the results that would be obtained from the entire population from which the sample was drawn. In other words, generalizability depends on the degree to which the particular sample in question can be said to be representative of the population.

A *population* is the entire group of people that a particular study is interested in. For example,

Properly sampled data should represent what would result if data for the entire population were used. In other words, the results of the study should be representative of results that would occur if the researcher were able to investigate the entire population.

Therefore a number of strategies are used to accomplish this representativeness, but the two most common ones are called **random samples** and **stratified random samples.**

Random Sampling : refers to the selection of participants from the general population that the sample will represent. In most second language studies, the population is the group of all language learners, perhaps in a particular context. Quite clearly, second language researchers do not have access to the entire population (e.g., all learners of English at Algerian universities), so they have to select an accessible sample thatis representative of the entire population .

There are two common types of random sampling: **simple random** (e.g., putting all names in a hat and drawing from that pool or using a dice) . Or the researcher assigns a number to each participant and use a table of random numbers to choose as many subjects as are needed . The use of such a list eliminates biases in the researcher's choice of subjects . Simple random sampling is generally believed to be the best way to obtain a sample that is representative of the population, especially as the sample size gets larger. However, simple random sampling is not used when researchers wish to ensure the representative presence of particular subgroups of the population under study (e.g., male versus female or particular language groups). In that case, **stratified random sampling** (e.g., random sampling based on categories) is used . In stratified random sampling, the proportions of the subgroups in the population are first determined, and then participants are randomly selected from within each stratum according to the established proportions.

Stratified random sampling provides precision in terms of the representativeness of the sample and allows preselected characteristics to be used as variables . In some types of second language research it might be necessary, for example, to balance the number of learners from particular L1 backgrounds in experimental groups. For other sorts of second language questions it might be important to include equal numbers of males and females in experimental groups, or to include learners who are roughly equivalent in terms of amount and type of prior instruction or length of residence in the country where the research is being conducted. There is yet another approach to sampling, called cluster random sampling. Cluster random sampling is the selection of groups (e.g., intact second language classes) rather than individuals as the objects of study. It is more effective if larger numbers of clusters are involved. In larger-scale second language research, for example, it might be important to ensure that roughly equal numbers of morning and evening classes receive the same treatments; however, as with any method, the research question should always drive the sampling choice.

Non-Random Sampling : Nonrandom sampling methods are also common in second language research. Common nonrandom methods include systematic, convenience, and purposive sampling. Systematic sampling is the choice of every nth individual in a population list (where the list should not be ordered systematically). For example, in organizing a new class where learners have seated themselves randomly in small groups (although one must be sure that the seating was truly random rather than in groups of friends/acquaintances), teachers often ask learners to count themselves off as As, Bs, and Cs, putting all the As into one group and so on. In a second language study, researchers could do the same for group assignments, although it would be important that the learners were seated randomly.

Convenience sampling is the selection of individuals who happen to be available for study. For instance, a researcher who wanted to compare the performance of two classes after using different review materials might select the two classes that require the review materials based on the curriculum. The obvious disadvantage to convenience sampling is that it is likely to be biased and should not be taken to be representative of the population.

However, samples of convenience are quite common in second language research. For example, researchers may select a time and a place for a study, announce this to a pool of potential participants, and then use those who show up as participants. These learners will show up depending on their motivation to participate and the match between the timetable for the research and their own schedules and other commitments.

In a **purposive sample,** researchers knowingly select individuals based on their knowledge of the population and in order to elicit data in which they are interested. The sample may or may not be

intended to be representative. For example, teachers may choose to compare two each of their top-, middle-, and lower-scoring students based on their results on a test, or based on how forthcoming these students are when answering questions about classroom processes. Likewise, a researcher may decide to pull out and present in-depth data on particular learners who did and did not develop as a result of some experimental treatment in order to illustrate the different pathways of learners in a study.

Representativeness and Generalizability : In addition to the representativeness of the sample , it is important to describe the setting. A study conducted in a university setting may not be generalizable to a private language school setting. It is often the case that to protect the anonymity of participants, one makes a statement such as the following about the location of the study: "Data were collected from 35 students enrolled in a second-year Japanese class at a large U.S. university." It is important to minimally include this information so that one can determine generalizability.

Private language school students may be different from students at large universities, who may in turn be different from students at other types of institutions. When choosing a sample, the goal is usually that the sample be of sufficient size to allow for generalization of results . Novice researchers often wonder how many learners are "enough“ for each group or for their study overall.In second language research for instance , Fraenkel and Wallen (2003) provided the following minimum sample numbers as a guideline: 100 for descriptive studies, 50 for correlational studies, and 15 to 30 per group in experimental studies. Finally , If random sampling is not feasible, there are two possible solutions: First, thoroughly describe the sample studied so that others can judge to whom and in what circumstances the results may be meaningful. Second, as we also discussed in lecture 1, conduct replication studies (and encourage the same of others) wherever possible, using different groups of participants and different situations so that the results, if confirmed, may later be generalized.

Collecting Biodata Information :When reporting research, it is important to include sufficient information to allow the reader to determine the extent to which the results of your study are indeed generalizable to a new context. For this reason, the collection of biodata information is an integral part of one's database. The major consideration is how much information to collect and report with respect to the participants themselves. In general, it is recommended that the researcher include enough information for the study to be replicable (American Psychological Association, 2001)

In reporting information about participants, the researcher must balance two concerns. The first is the privacy and anonymity of the participants; the second is the need to report sufficient data about the participants to allow future researchers to both evaluate and replicate the study. There are no strict rules or even guidelines about what information should be obtained in the second language field; because of this, exactly what and how much detail is obtained will depend on the research questions and will vary for individual researchers.

It is generally recommended that major demographic characteristics such as gender, age, and race/ethnicity be reported (American Psychological Association, 2001), as well as information relevant to the study itself (e.g., the participants' first languages, previous academic experience, and level of L2 proficiency). Additional information that might be important for a study on second language learning could include the frequency and context of L2 use outside the classroom, amount of travel or experience in countries where the L2 is spoken, learners' self-assessment of their knowledge of the target language, and the participants' familiarity with other languages.

- Therefore , we have pointed out that it is often difficult to ensure external validity but have shown ways to minimize threats to external validity. Following is a summary of ways in which one can deal with such threats:
- Random sampling.
- • Stratified random selection.
- • Systematic, convenience, and purposive sampling.
- • Sufficient descriptive information about participants.
- • Description of setting.
- • Replication of study in a variety of settings.

Lecture 8

Validity& Reliability of Measurement

Now we are going to learn about reliability and validity of a **measure**

- Validity asks
  - if an instrument measures what it is supposed to
  - how "true" or accurate the measurement is

There are many types of validity :

Content Validity Content validity refers to the representativeness of our measurement regarding the phenomenon about which we want information. If we are interested in the acquisition of relative clauses in general and plan to present learners with an acceptability judgment task, we need to make sure that all relative clause types are included.

Face Validity is closely related to the notion of content validity and refers to the familiarity of our instrument and how easy it is to convince others that there is content validity to it. If, for example, learners are presented with reasoning tasks to carry out in an experiment and are already familiar with these sorts of tasks because they have carried them out in their classrooms, we can say that the task has face validity for the learners. Face validity thus hinges on the participants' perceptions of the research treatments and tests. If the participants do not perceive a connection between the research activities and other educational or second language activities, they may be less likely to take the experiment seriously.

Construct Validity

This is perhaps the most complex of the validity types discussed so far. Construct validity is an essential topic in second language acquisition research precisely because many of the variables investigated are not easily or directly defined. In second language research, variables such as language proficiency, aptitude, exposure to input, and linguistic representations are of interest. However, these constructs are not directly measurable in the way that height, weight, or age are. In research, construct validity refers to the degree to which the research adequately captures the construct of interest. Construct validity can be enhanced when multiple estimates of a construct are used. For example, in the hypothetical study discussed earlier that was seeking to link exposure to input with accuracy in

identifying final consonants, the construct validity of the measurement of "amount of input" might be enhanced if multiple factors such as length of residence, language instruction, and the language used in the participants' formal education were considered together.

Criterion-Related Validity

Criterion-related validity refers to the extent to which tests used in a research study are comparable to other well-established tests of the construct in question. For example, many language programs attempt to measure global proficiency either for placement into their own program or to determine the extent to which a student might meet a particular language requirement. For the sake of convenience, these programs often develop their own internal tests, but there may be little external evidence that these tests are measuring what the programs assume they are measuring. One could measure the performance of a group of students on the local test and a well-established test (e.g., TOEFL in the case of English, or in the case of other languages, another recognized standard test). Should there be a good correlation , one can then say that the local test has been demonstrated to
have criterion-related validity.

Predictive Validity Predictive validity deals with the use that one might eventually want to make of a particular measure. Does it predict performance on some other measure? Considering the earlier example of a local language test, if the test predicts performance on some other dimension (class grades), the test
can be said to have predictive validity.

Reliability: in its simplest definition refers to consistency, often meaning instrument consistency. For example, one could ask whether an individual who takes a particular test would get a similar score on two administrations of the same test. If a person takes a written driving test and receives a high score, it would be expected that the individual would also receive a high score if he or she took the same written test again. We could then say the test is reliable. This differs from validity, which measures the extent to which the test is an indication of what it purports to be (in this case, knowledge of the rules of the road). Thus, if someone leaves the licensing bureau having received a high
score on the test and runs a red light not knowing that a red light indicates "stop," we would say that the test is probably not a valid measure of knowledge of the rules of the road. Or, to take another example, if we want to weigh ourselves on scales and with two successive weighings find that there is a 10-pound difference, we would say that the scales are not reliable.
There are many ways that one can determine **rater reliability** as well as **instrument reliability**.

Rater Reliability

The main defining characteristic of rater reliability is that scores by two or more raters or between one rater at Time X and that same rater at Time Y are consistent.

Interrater and Intrarater Reliability : In many instances, test scores are objective and there is little judgment involved. However, it is also common in second language research for researchers to make judgments about data. **Interrater reliability** begins with a well-defined construct. It is a measure of whether two or more raters judge the same set of data in the same way. If there is strong reliability, one can then assume with reasonable confidence that raters are judging the same set of data as representing the same phenomenon.
Intrarater reliability is similar, but considers one researcher's evaluations of data, attempting to ensure that the researcher would judge the data the same way at different times—for example, at Time

1 and at Time 2, or even from the beginning of the data set to the end of the data set. To do this, one essentially uses a test-retest method ; two sets of ratings are produced by one individual at two times or for different parts of the data. Similar to interrater reliability, if the result is high, then we can be confident in our own consistency

Instrument Reliability Not only do we have to make sure that our raters are judging what they believe they are judging in a consistent manner, we also need to ensure that our instrument is reliable. In this section, we consider three types of reliability testing: test-retest, equivalence of forms of a test (e.g., pretest and posttest), and internal consistency.

Test-Retest. In a test-retest method of determining reliability, the same test is given to the same group of individuals at two points in time. One must carefully determine the appropriate time interval between test administrations. This is particularly important in second language research given the likelihood that performance on a test at one time can differ from performance on that same test 2 months later, because participants are often in the process of learning (i.e., do not have static knowledge). There is also the possibility of practice effects, and the question of whether such effects impact all participants equally. In order to arrive at a score by which reliability can be established, one determines the correlation coefficient5 between the two test administrations.

Equivalence of Forms. There are times when it is necessary to determine the equivalence of two tests, as, for example, in a pretest and a posttest. Quite clearly, it would be inappropriate to have one version of a test be easier than the other because the results of gains based on treatment would be artificially high or artificially low, as discussed earlier. In this method of determining reliability, two versions of a test are administered to the same individuals and a correlation coefficient is calculated.

Internal Consistency. It is not always possible or feasible to administer tests twice to the same group of individuals (whether the same test or two different versions). Nonetheless, when that is the case, there are statistical methods to determine reliability; split-half, Kuder-Richardson 20 and 21, and Cronbach's a are common ones. We provide a brief description of each. Split-half procedure is determined by obtaining a correlation coefficient by comparing the performance on half of a test with performance on the other half. This is most frequently done by correlating even-numbered items with odd-numbered items. A statistical adjustment (Spearman-Brown prophecy formula) is generally made to determine the reliability of the test as a whole. If the correlation coefficient is high, it suggests that there is internal consistency to the test.

Kuder-Richardson 20 and 21 are two approaches that are also used. Although Kuder-Richardson 21 requires equal difficulty of the test items, Kuder-Richardson 20 does not. Both are calculated using information consisting of the number of items, the mean, and the standard deviation . These are best used with large numbers of items. Cronbach's a is similar to the Kuder-Richardson 20, but is used when the number of possible answers is more than two. Unlike Kuder-Richardson, Cronbach's a can be applied to ordinal data.