Generalization

When we look at a study or set of related studies, we would like to know how general the results are.

1. Do they apply to other populations or are the results limited by age, sex or gender, culture, species?

2. Would other experimenters and laboratories obtain the same results?

3. Are laboratory and real world results comparable? How replicable are the results?

4. How general are the concepts? Does changing details of the experiment alter the results? Can we change operational definitions and get similar results?

I. Generalization to Other Populations

A) The populations that we routinely use

1. For human studies, we use *college students*, largely from introductory psychology courses.

College freshmen and sophomores do differ systematically from other groups of humans. They are 17-21 years old (late adolescent, still forming political and social views, groups of friends and relationships still relatively unstable, above average intellectually).

2. For animal studies, *rats* are commonly used. Rats are easy to rear, relatively inexpensive to maintain, hardy.

3. *Volunteers*. People who volunteer for research differ systematically from non-volunteers. They tend to be more highly educated, are higher in need of social approval, have wider networks of friends ("more social"), and have higher socio-economic status.

Furthermore, different kinds of people volunteer for different kinds of experiments. For example, Hood & Black (1971) found that when requests for volunteers were posted on a bulletin board, different kinds of people signed up for different studies based on the titles of the studies.

4. *Gender and race*. For many years (up to the 1970s), white males were over-represented as participants in research. In medical studies, research focused on disorders and disease in males. In research in social psychology, many studies used just white males to make their sample more homogeneous (reducing variability due to individual differences).

B) Factors limiting generalization

If the routine populations listed above are different from other groups, this may limit the generality of our results.

1. Gender & race – In the last 30+ years, there has been a push to broaden medical research to focus on women and minorities. Similarly, behavioral research is done with a broader sample of the population. As our society has become more integrated, the Introductory Psychology Subject Pool has come to look like a cross-section of society.

2. Locale – Results may differ from location to location (e.g. Los Angeles versus Minneapolis). Separate from gender, race and ethnicity, there are cultural differences between different parts of the country or even between different neighborhoods.

3. Age – Different ages are parts of different *cohorts*. They differ due to maturation, history, and cultural differences. Even "simple", "stable" phenomena change. Hearing, eyesight, memory, and speed of responding all change over the lifespan.

For example, hearing loss is changes with cohort. OSHA rules tend to reduce hearing loss for younger generations in the workplace while the ready availability of loud music (e.g. iPods) may increase it.

C) What are the influences of these subject characteristics?

Basically, there are four patterns of results that we can find when we treat the differences between subject populations as a subject variable in a factorial design:

1. No effect
2. Additivity (effect of subject characteristic, no interaction)
3. Simple interaction
4. Cross-over interaction

The issue of generalization can be explicitly addressed by either using a subject characteristic as a variable in a factorial design or by doing the study multiple times with different populations.

D) In defense of college students and rats

Before criticizing a study simply because it was done with a particular population, we should have good reasons for expecting that the results would be different with another population.

College students are humans, and they represent most of the major cultural groups within our society (at least since the 1980s).

We need to start someplace in our research. Starting with college students may be more appropriate than jumping directly to an alternative population.

Results with rats on learning have been shown to generalize (e.g. with regard to the influence of schedules of reinforcement) and these results have proven useful in studying and modifying human behavior.

II. Effects of Culture

Culture includes shared sets of values, rearing practices with children, and educational systems.

By culture, we are NOT referring to race and ethnicity.

As an example of the influence of culture, consider the difference between collectivist cultures that emphasize the group versus individualistic cultures that emphasize the individual. In studies of self-concept and self-esteem, consistent differences are found. In Japan, self-esteem is influenced by the network of relationships with others more strongly than in the U.S. In the U.S., individual achievements have a stronger influence on self-esteem than they do in Japan.

III. Pretests, Researchers, and Laboratories

A) Pretests

The advantage of a pretest of participants is that we can look at change within the individual. This also allows us to assess whether individuals who drop out of a study are different from those who remain. (*differential mortality*)

However, in the real world, people seldom get a pretest. We do not measure peoples' attitudes before they are exposed to a political speech or commercial.

The issue here is whether the pretest reduces the generality of our results. Does it *sensitize* the individual to the nature of our measurement (producing a *practice effect*) or to the concepts being tested in the study (producing *reactivity*).

If mortality is a concern or the sensitivity of the pretest-posttest is needed, but we are also concerned about the influence of the pretest, then we can use the Solomon four group design. This allows us to assess the effect of the pretest. We could also run the study twice, with and without the pretest and compare the results.

B) Researchers

Do you get the same results with different researchers?

1. A more experienced researcher may pay closer attention to details and be less likely to produce experimenter bias. In animal studies, an experienced researcher often produces faster rates of learning (see Brogden, 1962).

2. A warm and friendly researcher can produce different results than an researcher who is cold and aloof.

3. The gender of the researcher can interact with the gender of the participant. Higher levels of performance have been found in some experiments when the participant is tested by an researcher of the opposite sex (Stevenson & Allen, 1964).

One way to ensure generality is to use two (or more) researchers. Train them consistently. Use both male and female researchers where there is any reason to expect that this will influence the results (see Rubin, 1975 for an example of this).

IV. Laboratory and Real World Settings

The laboratory setting affords greater control. However, the question of generality is whether the results of studies in the laboratory will replicate in the real world *or* have they been produced or altered by the artificial environment of the laboratory.

When we deal with the relevance of laboratory studies to the real world, we need to distinguish two different factors:

1. Mundane realism – similarity of study to events in the real world.

2. Experimental realism – do the participants take the study seriously and does the independent variable have an impact on them.

If a laboratory study has mundane realism, but lacks any experimental realism (the participants are bored and uninvolved in the study), then it is unlikely to produce results analogous to the real world. A study that lacks both mundane and experimental realism is unlikely to produce useful results.

Anderson et al. (1999) looked at 38 pairs of studies where a laboratory study and a field study were similar. These studies include work on memory, leadership style, aggression, helping, and depression. Overall, the results of the field and experimental studies were similar. The magnitude of the effects in both environments were also similar.

A program of research often includes both laboratory and field studies. This allows the researcher to establish the generality of the results and then do further research using the setting most appropriate to the question(s) being asked.

V. Reliability, Replication, and Converging Operations

A) Reliability, Replication, and Generalization

If someone conducted an opinion survey on preferences for president, opinion on tax cuts, etc., what would give you confidence in the results?

1. A larger sample. More likely to generalize to the target population.

2. Similar results from a second sample (repeatability).

Repeating an experiment, survey, etc. and getting the same or similar results increases our confidence that the results are reliable.

Repeating a study is usually done with a different sample of subjects. If we get the same or similar results, we have demonstrated *experimental reliability*.

Generally, we prefer experimental reliability to statistical reliability. Statistical reliability, which is based on the odds that our results are due to chance, can still be wrong. If we replicate an experiment, it is unlikely that we will have all of the other, extraneous factors exactly the same, so if the earlier results were due to chance, then the results the second time are likely to be different.

Thus, experimental reliability, in which the pattern of results from one study recurs in additional studies, is preferred to statistical reliability.

There are three types of replication of a study:

1. Direct (repeat the same study)
2. Systematic (vary other factors)
3. Conceptual (use new operational definitions)

1. Direct replication.

Repeat a previous study with new subjects. Keep the changes between studies to a minimum.

2. Systematic replication.

How strong (or fragile) is the effect? Does it hold up for different populations of individuals and different specific details of how the study is done?

Systematic replication is an attempt to establish generality by varying factors that are *not* thought to make a difference.

3. Conceptual Replication

Here, the goal is to establish the generality of the concepts by changing operational definitions.

Conceptual replication is part of the process of establishing external validity. Here, we are changing factors that could make a difference to explore the generality of the concepts we are testing.

In summary, direct and systematic replication are all part of establishing reliability. Conceptual replication (and to some extent, systematic replication) are part of the process of establishing validity and the generality of our understanding.

B) Converging Operations

In converging operations, we want to explore the generality of a concept. In addition, we want to rule out alternative explanations.

We do this by using multiple experiments (or multiple independent variables) or multiple studies to provide a pattern of data that can rule out some of the alternative explanations.

In converging operations, we deliberately change the operational definitions that relate our independent and/or dependent variables to the concepts we are exploring. If we find the expected pattern of results, then we have established validity for our concepts. That is, the concept has meaning and generality beyond a particular set of definitions.

C) Example: The Stroop Effect.

1. Background

Construct three lists of items. The first is a series of color words (red, green, blue, yellow, etc.) written in black ink. The second is a series of color patches, in different ink colors (red, blue, ...). The third is color words written in conflicting ink colors (e.g. the word "red" written in blue ink).

Two groups of subjects participate. One group is given the list of color words in black ink and asked to read the words. They are also give the list of color words in conflicting ink colors and asked to read the words. This is the reading group.

The second group is given the list of color patches and asked to name the ink colors. They are also given the list of color words in conflicting ink colors and asked to name the ink colors. This is the naming group.

All subjects do each of their lists as fast as possible. List order within each group is counterbalanced.

The results show that the speed of reading the color words is unaffected by the ink color. Reading times for the two lists are the same. The speed of color naming is much slower for the color words written in conflicting ink colors than for the color patches. This result is termed Stroop interference. It has been widely replicated, both directly and systematically.

How general is this interference between reading and other “tasks”.

2. Conceptual Replication

Is this interference is not just about colors and color words? Can it occur for other materials? For example, suppose we ask our participants to count the number of items in each row of a display and say the count aloud.

In the display in the next page (List A), they would go down the row saying “one”, “four”, … In the display following that (List B), they would also go down the list, giving the count of the number of characters in each row, saying “one”, “four”, … However, this second list is much harder because the digit in each row conflicts with the count.

When asked to read the digit in each row, List B is not harder than a control list.

Clearly, Stroop Interference is a general phenomena due to conflict between reading and other use of the information.

List A

X	response “one”
XXXX	response “four”
XXX	
XX	
XXXX	
XX	
XXX	
X	
X	
XX	
XXX	
XXXX	
XX	
XXXX	
X	
XXX	
XXXX	
XXX	
X	
XX	
XX	
XXX	

List B

2	response “one”
2222	response “four”
11	
444	
3	
111	
44	
3333	
1111	
4	
222	
33	
11	
2222	
4	
444	
33	
111	
3333	
2	
44	
222	

3. Converging operations and different theories.

What causes the Stroop effect?

a) Interference during the input or perceptual processing of the information. Reading is faster than color naming in perception, so it is possible that reading color words inhibits the perception of the ink color. Call this the Perceptual Inhibition hypothesis.

b) Interference in the process of choosing the correct response to make. After the subject has perceived both the word and the ink color, there is competition between the two color names. The word name is available first or is more strongly associated with the response (from practice reading), forcing the subject to respond slowly with the ink color name. Call this the Response Competition hypothesis.

Experiment by Egeth et al. (1969)

In this experiment, a new variation of the Stroop task was tried with three conditions. The first was a control condition, the second was designed to test the idea of Perceptual Inhibition and the third was designed to test Response Competition. One group of subjects was run through all three conditions.

a) *Neutral control*. On each trial, the subject sees two color patches. If they are the same ink color, subject says "same", if they are different, subject says "different".

b) *Perceptual Inhibition*. On each trial, there are two color words written in various ink colors. If the ink colors match, say “same”. If the ink colors do not match, say “different”. If color words inhibit the perception of ink color, then this condition should be slower than the control. There should be no response competition here since the subject is not responding “red”, “blue”, etc.

c) *Response Competition*. On each trial, the words SAME and DIFF (different) appear written with either the same or different ink colors. If both ink colors are the same, say "same". If the ink colors are different, say “different”. Here, the response can compete with the printed words and if response competition is occurring, the speed for this condition should be slower than the control. However, there is no reason for perceptual interference between the words SAME and DIFF and the ink colors.

It is possible that all three conditions will produce the same speed of response. This would indicate no Stroop effect in this study and rule out both explanations.

Finally, it is possible that both the perceptual inhibition and the response competition conditions will be slower than the control, indicating either that some third factor is causing Stroop interference or that Stroop interference has both input and output components.

Condition	Response	Stimulus	
Neutral	same	XXXX	XXXX
	different	XXXX	XXXX
Perceptual	same	BLUE	BLUE
	different	BLUE	BLUE
Response	same	SAME	SAME
	different	SAME	SAME
	same	DIFF	DIFF
	different	DIFF	DIFF

Results:

The RTs in the control and Perceptual Inhibition conditions were the same. Subjects were slower in the Response Competition condition.

If Response Competition is correct, then if this experiment were run again with a simple change in the instructions about what response to make, we should be able to make the effect of response competition disappear. We tell subjects that whenever the ink colors on a trial match, say "yes" and whenever they don't match, say "no". Now, the response should not compete with the printed materials, so no response competition should occur. Also, since our materials are the same as before, we should get no difference between control and perceptual inhibition conditions.

This simple change in instructions eliminated all interference. The response times for the three conditions were the same.

Condition	Response	Stimulus	
Neutral	yes	XXXX	XXXX
	no	XXXX	XXXX
Perceptual	yes	BLUE	BLUE
	no	BLUE	BLUE
Response	yes	SAME	SAME
	no	SAME	SAME
	yes	DIFF	DIFF
	no	DIFF	DIFF

Conclusion:

Through the use of converging operations, it has been shown that the Stroop Effect is largely the result of competition between alternative responses to a situation or set of materials. These results show that the effect is reliable and robust. Furthermore, the different presentation conditions and instructions given to the subjects across the experiments allows us to rule out one of the potential explanations for Stroop Interference.

Chapter 14 Sample Questions

1) It may be problematic to generalize results from a group of volunteers to the general population in an experiment on the effects of various types of persuasion because: a) volunteers generally have a higher need for social approval b) volunteers generally have a lower level of education c) volunteers never yield the same results as non-volunteers d) all of the above

2) Experimental reliability can be established by: a) using inferential statistics b) changing the operational definitions and re-running the experiment c) replication of an experiment with a new sample of subjects d) a & b above

3) Confidence in the generality of an experimental result increases when: a) the experiment is repeated and we obtain the same results b) the experiment is based on a large, random sample from the population c) the experiment is systematically and conceptually replicated d) b & c above

Answers for Chapter 14 Sample Questions

1) – a; 2) – c 3) – d;

Note here that b) is not correct for question 2. Once the operational definitions are changed, the experiment is now concerned with validity and generalization. To establish experimental reliability, we repeat the original study.