17

Generalizing

Dale Hample and Yiwen Dai

1. Preliminaries

Arguments manage meanings. Normally an argument begins with something that is already known or agreed upon, and then proceeds to a conclusion that is "new" in some sense. This movement from given to new is characteristic of all social arguments. The only exception might be perfectly deductive arguments, where all the information is already understood to be contained somehow in the premises. Informal (or social) arguments involve some sort of inference, a going-beyond what is strictly expressed in the premises. By moving to a new proposition or thought, the argument can create a new meaning for someone, or reinforce an old meaning, or refute an inappropriate meaning. Arguments manage meanings.

One particularly valuable sort of meaning is a generalization. A generalization summarizes or characterizes material that is more singular than it is. If the singular things are few and straightforward, the generalization isn't especially useful (e.g., "My parents both like fruit pies" isn't much more convenient than "Mom and Dad both like fruit pies"). But the real value (and threat, but we will come to that when we discuss stereotypes) of a generalization is when it does a lot of work by summarizing too many things to think about at once (e.g., "Some thoughtful adults in nearly every human generation and culture have complained that younger people are lazy and self-centered").

Generalizations are often involved in arguments, and we can distinguish several kinds of such arguments. Some arguments move from singular premises to a more general conclusion. This involves arguing from examples or doing some approximation of scientific induction. For example, "In my survey, most people who were high on dogmatism were also high in authoritarianism; so dogmatism and authoritarianism are positively associated." We will call these *generalization-establishing* arguments. Other arguments have a generalization in the premises, but still move to a singular conclusion (you

know one famous example like this: "All men are mortal; Socrates is a man; so Socrates is mortal"). We will call these *generalization-using* arguments. Finally, some arguments contain nothing but generalizations. For example, "everyone acts from self-interest; all self-interested actions damage social cohesion; therefore everyone damages social cohesion." We will call these *generalization-saturated* arguments. We will discuss all these, but will have the most to say (by far) about generalization-establishing arguments.

As for other social arguments, we have four measures of evaluation available to us. These are criteria that one can use to evaluate the argument's strength or cogency. The first three are the ARS considerations, originated by Anthony Blair and Ralph Johnson. The argument's premises must be *acceptable* (A). They must be true, or probable, or plausible, or agreeable in some respectable way. If at least one key premise fails this test, the argument is poor. The conclusion might still be true by accident, but that argument will not have supported it. Second, the premises must be *relevant* (R) to the conclusion. Relevance is a little hard to define precisely, but the idea is that premises must bear on the conclusion, so that if the premises are true a reasonable person becomes more comfortable with the conclusion. Relevant premises assert something that is an element of the conclusion or that implies something about the likelihood of the conclusion. S stands for *sufficiency*. The idea here is that *all the premises together* are enough to prove the conclusion. One can have what is called a convergent argument that contains perhaps four individually flawed reasons for the conclusion. One at a time each reason could be set aside, but when they are combined they make a jointly suitable case for the conclusion. Sufficiency and necessity are different, by the way, and they are sometimes confused. Sufficient premises are *enough* to prove a conclusion. Necessary premises are *required* for a conclusion to be so. A premise can be necessary without being sufficient ("You have to study a lot in order to graduate"). A premise can also be sufficient without being necessary ("People lucky enough to win the lottery end up being rich"). Sufficiency is the third criterion, not necessity.

The fourth criterion is what Ralph Johnson called the dialectical tier of argument. Acceptability, relevance, and sufficiency are all internal to the argument. They are about the premises and the argument scheme. The dialectical tier is external to it. "Dialectic" describes an exchange between two ideal arguers who marshal their reasoning skills, commitment to the truth, and their *knowledge of the world* to come to a conclusion. When thinking about the dialectical level of argument evaluation, we ask ourselves what else in the world—external to these premises—might bear on the conclusion. So someone might present us with an argument for Western European military intervention into Turkey. Even if the premises about Western Europe and

Turkey seemed acceptable, we would be entitled to object to the argument if it didn't take Russia into account at all. Refutative arguments, exceptions, and counter-examples take their place on the dialectical tier.

We will be using these four standards—acceptability, relevance, sufficiency, and dialectical quality—as we work through the various ways generalizations fit into arguments. Generalization arguments are as likely to go bad as any other kind of argument, but these also hold out the special danger of stereotyping. So as we move along we will pause to give particular attention to the possibility that arguments contain racism, sexism, religious intolerance, damaging views about other people's identities, and similar sorts of things. Stereotypes and generalizations can be hard to distinguish in a technical way. We will loosely say that stereotypes are generalizations that express unjustified hate, offense, or devaluing of some group of people, and not pursue definitions any further than that. (Actually, some stereotypes can be positive, as with a favorable view of everyone representing your political party, but we will neglect that possibility here.) As the United States Supreme Court ended up saying about pornography in the 1964 case *Jacobellis v. Ohio*, "I know it when I see it." We will organize the rest of the chapter by reference to the ideas of generalization-generating, generalization-using, and generalization-saturated arguments.

2. Generalization-establishing arguments

We begin with arguments that move from premises about specific cases to a summative conclusion that groups the cases together and offers some statement about the group. For example, "Uncle Ed is a schmuck; Aunt Raby is dumb; cousin Edward can't count yet and he is 8 years old; cousin Edie won't read a book unless you pay her; these are most of my relatives on my father's side; so my father's family is pretty hopeless." This argument moves from specifics (Uncle Ed, Aunt Raby, etc.) to a group (my father's family) and makes a remark about the group. Notice that the example also collects various descriptions (schmuck, dumb, can't count, won't read) and composes them into a summary (pretty hopeless) that is an idea that not only collects but also extends the implications of the specific descriptions. So there are actually two sorts of generalizing going on here. The first is like induction, where individual cases are examined and a probabilistic conclusion is drawn about the class of cases. The second is a kind of argument by composition, where attributes of some more general characteristic are sampled and a conclusion is drawn about the general characteristic. These two sorts of generalization don't always have to be present together, but they often are. They are quite likely to be involved in evidence offered for stereotypes.

Just by using common sense, we can immediately see opportunities to criticize this argument. We might think along these lines:

Acceptability: Are these attributes (schmuck, etc.) really true about each case (Uncle Ed, etc.)? Or are they perhaps just a series of poorly supported impressions? What's the evidence? Does everyone (or informed people, or smart people, etc.) see these people this way?

Relevance: Do these attributes (schmuck, etc.) really bear on the conclusion about the group (pretty hopeless)? Are these four people really "most of" the people in the group mentioned in the conclusion? Are these four people typical of everyone in your father's family? You are in your father's family, too, for instance.

Sufficiency: Is it possible for the attributes to be true and relevant, but the conclusion to be wrong? That is, does more need to be known before the conclusion can be fairly drawn? Are the named attributes (can't count, etc.) keys to the conclusion (pretty hopeless)? That is, do they "add up" to the conclusion?

Dialectical Tier: Do these people have other, more wonderful, attributes? Or even more awful characteristics? Is "pretty hopeless" too strong a characterization? Too weak? As a matter of fact, what could "hopeless" mean in this sort of context, and can we define and defend the core idea of "pretty hopeless?"

Perhaps you can immediately see the usefulness of our little system of four criteria for evaluating an argument. We had some trouble deciding where to put the question about what "pretty hopeless" means, by the way. The component meanings in an argument need to be clear before you can really think about any of the ARSD standards, so the question of clarity actually appears at any time you are trying to evaluate anything that bears on the term. We put it in Dialectical Tier mainly because it seemed a little philosophical. You don't need to do the ARSD thinking in any particular order anyway.

We started with what we thought might be a recognizable sort of example so that you can see where we are going with all this. But now we want to get a little bit more sterile in order to clarify some of the standards for this sort of argument. By looking at the most formally developed ways to establish a generalization, we should see the principles at work in more ordinary arguments too. We are going to abandon the argument-by-composition feature and just concentrate on induction-like sorts of generalization. You will encounter many examples of this when you read about public opinion or scientific investigation.

3. Surveys

Let's begin with public opinion polling. If you are interested in public affairs, nearly every week you will read about some poll that summarizes (i.e., makes a generalization about) public views of a politician or policy. Some of these polls are now labeled "unscientific" by the media that report them. Often the media also report that results have a "margin of error," perhaps plus or minus 5%. We will explain these things as we go. Younger readers may not appreciate that neither of these things—acknowledgement that a poll is unscientific, notice that the results are not precisely accurate—used to be done, and that these are improvements in journalism during the last half century. Even now, however, the media do not always report the exact question they are summarizing, which is really essential to evaluating their generalization.

Two of the key preliminary steps in conducting a poll are to write the questions and select the sample. The question needs to be such that it will get the pollster an accurate answer, and that means that it must address the subject matter without any imbalance. For example, you will get different answers to these pairs of questions:

- "Do you think that women have a right to choose what happens within their own bodies? / Do you think that women have a right to abort their fetuses?"
- "If a drug would save 200 lives out of 600 very sick people, should we administer it? / If a drug would kill 400 people out of 600 very sick people, should we administer it?"
- "Do you think that everyone, regardless of race or other identity, should have an equal chance to succeed in life? / Do you think that affirmative action programs are justified?"

It's possible that a particular person might think that the first question in each pair has the same meaning as the second one. (It's also possible that someone is trying to rig a poll to mislead the public into committing the bandwagon fallacy, or is actually using the poll as a persuasive device by getting people to agree to certain things out loud.) But the questions are probably going to register differently with people, emphasizing one element of the issue more than another or using some vague but hard to question phrase ("right to live," "equal chance," etc.). When the media report the question, viewers are better positioned to evaluate the poll. But often the media will "summarize" the question for their audience, reporting either half of the first pair as "attitudes toward abortion," or either half of the third pair as "approval ratings for affirmative action programs." Professional political pollsters have quietly decided to use the same question phrasing from poll to poll and from political party to

political party, so that their clients can actually compare themselves to opponents and past results. Even if the standard question is flawed, at least the same flaw will be propagating through everyone's data, and that makes the results more useful. Data doesn't have to be perfect to be valuable: it merely has to have a known level of accuracy.

Oddly, that brings us to the second preliminary matter, selecting a sample. A perfect statement about what the citizens of Mexico think about Mexican tax rates would be to ask every single Mexican citizen, and report the answer. This would be a census, not a sample. Samples are used because they are cheaper and faster than taking a census. A sample of Mexican citizens imperfectly represents the whole population of Mexican citizens, but the degree of imperfection is known—provided that the sample is done in a particular way. This is where the "margin of error" comes from. To get a known margin of error, the pollster needs to collect a random sample of the population. A "random sample" means that every person in the population has an exactly equal chance to be in the sample. To assure this (well, to approximate it, really), polling firms might use lists of people (perhaps everyone who voted in the last Republican primary, or everyone living in an apartment building, or everyone with a driver's license in a particular state or province, depending on the population one wants to generalize to) and choose people for the sample by using a random number generator. If the random number generator's first result is 546, the 546th person on the list is contacted, and so on. The list is the population (technically, it is the "sampling frame"), and the sample result is intended to generalize to that population. Sometimes pollsters will just randomly generate telephone numbers within particular area codes so that they can cover cell phones and unlisted numbers. The size of the population is known, either from the list of voters or the number of possible phone numbers, and the size of the sample is also known. *Provided that the sample was drawn randomly*, we can calculate or just look up the margin of error.

On the next page is a standard table[1] that shows some benchmark values. The population size is the total number of people on the list, in the country, etc. It is how many people would be in a census of that sort of person. The sample size is the number of people who were actually polled. The margin of error is the plus or minus estimate of how accurate the poll's answer would be (e.g., "54% of people approved, plus or minus 5%"). The confidence in the margin of error is how sure you are that the true answer will actually be within 5% or whatever the margin of error is. (Technically the confidence statistic is not a summary of people's subjective feelings. It is a statistical calculation of how often the percentage of results would be within the margin of error if an infinite number of samples of that size were drawn from a population of that

[1] (http://research-advisors.com/tools/SampleSize.htm)

Population Size	Sample Size	Margin of Error	Confidence in Margin
100	80	5%	95%
100	89	3.5%	95%
100	87	5%	99%
100	93	3.5%	99%
1,000,000	384	5%	95%
1,000,000	783	3.5%	95%
1,000,000	663	5%	99%
1,000,000	1352	3.5%	99%
300,000,000	384	5%	95%
300,000,000	784	3.5%	95%
300,000,000	663	5%	99%
300,000,000	1354	3.5%	99%

size.) Media aren't to the point of reporting this fourth thing yet, but as more people become better educated, perhaps they will start.

The table's information is interesting, and runs counter to many people's intuitions. Sometimes people will hear about some poll (inevitably, one whose result is annoying), and will dismiss it, saying something like, "How can 400 answers tell us what two million people think?" The table is right and that intuition is wrong. With small populations (the example in the table is 100), you really do need to survey nearly everyone to get very close to an accurate answer. Just to get within an error margin of 5% (with 95% confidence in that 5% margin), you need to get 80 out of 100 people in your sample. If you wanted 99% confidence in a 3.5% margin of error, you would need 93 people. This is where people's intuitions come from, we suppose. But with larger populations, you need much smaller percentages. You would only need 384 people out of a million to have 95% confidence in a 5% margin of error (that's about .04% of the population). And with a population of 300 million (almost the whole population of the U.S.A.), you don't need any more people in your sample than you did for a one million population. So there are enormous economies of scale in doing this sort of sampling.

But we emphasize: the results in the table only apply to perfectly random sampling. We tend not to worry very much about tiny imperfections, like missing people who lost their cell phone that day or other little flukes of everyday life. Those are technical violations of random sampling but they probably have negligible effects on the results, because they are non-systematic. That is, we can assume for instance that an equal number of Republicans and Democrats probably lost their cell phone that day. But when the violations are systematic, that is when the media have to call the poll "unscientific." For example (this is the usual one), a TV network might put a poll up on the Internet during the newscast and invite people to respond. The only people who respond are those who are watching that network at that particular time, who have Internet access at that moment, and who are highly motivated. If the poll result is accurate that is purely by accident, and there is no way to calculate things like margin of error—because all the statistical theory requires random sampling to have been done.

So *random sampling* is the gold standard for *knowing how much inaccuracy there is*. It isn't necessarily the gold standard for getting the most accurate results or getting them most efficiently. At least two other sorts of sampling are common in scientific investigations and both should be respected provided that they are done properly. *Stratified random sampling* is done when the population is divided into strata (or subgroups), and each stratum is then sampled randomly. Suppose that your population was all the children in your community playing a youth sport. Your strata might be 6-9 year olds, 10-12 year olds, and 13-16 year olds, because that's how the sports are organized, and you know that these age groups of kids are very different in many ways. Within each age group you would sample randomly. The advantage of this is that, because you assume that the children's age will be important to their answer, you have made sure that each age group is properly represented in your sample. That makes your overall result more accurate. However, it sacrifices your ability to say that your overall result has a specific margin of error, because it is not a random sample of the *whole* population. *Cluster random sampling* is another high quality sampling procedure. This time instead of dividing your sample into some interesting strata, you divide them up, usually geographically. So you might get a map of your city, mark off and number all the city blocks, and then randomly choose a certain number of blocks. That makes it efficient to gather your data: you simply drive to one block, park your car, and go home to home (or you could randomly sample within each block), and then drive to the next block on your list. This efficiency is paid for by some possible sacrifice of accuracy, because you might miss some interesting city blocks this way (maybe many of the elderly people live in one section of the city that has retirement homes or some minor-

ity is clustered, and you under- or over-represent them). These two kinds of sampling are rarely mentioned in the media, and you are mainly likely to encounter them if you are reading original research papers. For public use, most of what is done and reported is simple random sampling, and you should pay attention for any indications that it has been done inappropriately.

Polls, as we said, are intended to produce some generalization about what the public thinks. They move from individual answers to a generalized summary. Let us consider some of the ways one might argue for such a generalized summary.

1. You could ask your best friend, who says taxes should be lowered, and then announce that the public thinks taxes should be lowered. This can actually happen if you were talking to someone you really respect, like a parent or professor, and you don't stop to think about what you're doing. It is a pretty bad argument. It might be called anecdotal evidence or overgeneralization. The problem is that the evidence in the premises isn't sufficient to support the conclusion.
2. You could interview a hundred people in the shopping mall, average their answers, and announce that 62% of all citizens think taxes should be lowered. This is "unscientific" because it isn't a random sample. You don't have any way to estimate margin of error. And you misstated your conclusion: you have no right to say anything about "all citizens" because your population was "people in the mall on Thursday night." And you didn't even randomly sample them. Your data is barely relevant to a conclusion about "all citizens," and certainly not sufficient to support any answer as specific as 62%.
3. You can do (or hire) a professional survey that starts with a good population list, samples it randomly, and obtains a large enough sample to report that 55% of the citizens think taxes should be lowered, plus or minus 4%. If the scientific procedures are followed properly, we really only need to worry about whether the question was neutrally worded.

To this point, we have been discussing public opinion polling, which is a way to generalize views (or habits, or intentions, etc.) of people. This is very common in political affairs and even has some prominence in economic news (consumer confidence, wholesale orders, etc.). It is also the source of many conclusions about correlations, for example, that extroversion is positively correlated with a satisfying social life. Polling—or to be more general, surveying—involves observation of what is naturally happening. Experiments, in contrast, intervene into what is naturally happening and make generalizations about what happens as a result.

4. Experiments

In the past decades we have seen more and more media reports about experiments, especially about medical things. As people have become more sensitive to the sources of medical and nutritional advice, they have wanted to know something about the research basis of recommendations about whether they should use butter or margarine, whether coffee is healthy or deadly, what pills to take for arthritis, and so forth. The media have obliged by summarizing experimental research and conveying its generalizations to us. As practicing social scientists who actually conduct and publish experiments, we are often upset at what gets left out of these journalistic reports and what phrasing is used to replace the scientists' careful statements. You should learn to read the original reports for yourself if you are truly interested in the topic. Here, let us outline some basic ideas that will help you evaluate the generalizations that are drawn from experiments. These are issues that you should deal with yourself when you are reading the original reports, or suspicions that you should have when you are reading a summary of the reports that does not mention these matters.

An experiment differs from a poll in several fundamental ways because an experiment is straightforwardly designed to discover causality. In polls and surveys, everything co-occurs: causes, effects, and epiphenomena are mixed together; one cannot discern what things changed first and what things changed later; and one can only see the detritus of causal processes but not the causal processes themselves.

Experiments address these issues with several design features. (1) There is always at least one comparison group. Two or more groups with different histories (histories that are hypothesized to differ in a causal-process-relevant way) can be compared on a key outcome measure. The outcome measure is often called the dependent variable because its value depends on earlier causal dynamics. Sometimes it is called the criterion variable. (2) The timing of events is known. This is because the experimenter initiates a key change for the "experimental" group. (3) This change is the experimental manipulation. The experimenter interferes with the usual order of events by changing something artificially. In most social sciences this is called the manipulation because the researcher manipulates the value of the causal (independent or predictor) variable. In economics this would be called a shock to the system, which is a nice metaphor for what the experimenter does to the otherwise stable causal field. (4) Membership in the two (or more) groups to be compared is determined by random assignment. Random assignment is different than random sampling. In random sampling, everyone in the population has an exactly equal chance to be in the sample. In random *assignment*, everyone

in the sample has an exactly equal chance to be in group 1 or group 2, and so forth. Even if the sample was not drawn at random, a study can still have random assignment. So out of a group of volunteers (not a random sample of the population) an experimenter could randomly assign some to eat from one diet and the others to eat from another diet. Many descriptions of experimentation only emphasize two of these elements, manipulation and random assignment, but it is just as well to spell out the other items in order to appreciate how experiments generate generalizations.

For a simple design, an experimenter's argument goes like this. "I originally had two groups that were originally comparable in every important respect. I did not measure 'every important respect' because no one can think of all of them. Instead, I relied on random assignment to equalize the groups. With large enough sample sizes, random assignment will even out everything between the groups, even the things that aren't actually relevant. Then I intervened into the natural order of things by manipulating one of the groups' drug regimen (or their internet access, or the amount of sugar they ate, etc.). I then waited a scientifically justifiable length of time and accurately measured their health (or weight, or level of depression, etc.). I detected a difference between the two groups' average level of health (or weight, or depression, etc.) and applied proper statistical procedures to judge whether the difference was either dependable or so small that it might have occurred by chance. It was 'statistically significant,' meaning that it was large enough that it was unlikely to have happened by chance (this is the $p < .05$ standard). Since the difference between the two groups was statistically significant, I went on to calculate the effect size of this difference. This leads me to conclude that in this study, the experimental group (the one that got the manipulation) had X% better health. Therefore the manipulation causes better health to X degree."

In evaluating a study like this one, there are two categories of concern. These are internal and external validity. The classic treatment of these matters was written by Donald Campbell and Julian Stanley. These validity matters can be broken down into ARSD issues, and this would be an interesting exercise. But since there are well developed vocabularies for these two kinds of study validity, we will use those. Pure experiments are actually not vulnerable to internal validity problems because they were designed to avoid such critiques. But since every real-world experiment has some approximations and corner-cutting, we will cover those issues anyway.

For internal validity, the basic question is, "Are we sure that we understand the causal dynamics within the experiment?" In other words, if we observed a difference between the two groups, are we sure that the difference was due only to the experimental manipulation? Setting aside obvious impurities such as we see in movies (e.g., an evil doctor messing with the medicines, an

unscrupulous Senator getting his daughter into the treatment group, etc.), here are the standard internal validity issues: history, maturation, testing, instrumentation, statistical regression, selection biases, experimental mortality, and selection-maturation interaction. History refers to outside events that might have affected results. For instance, during World War II sugar was not ordinarily available in England and that would have affected a diabetes experiment that began in 1935 and ended in 1945. Maturation refers to the natural growth and development of study participants due to time passage and time passage alone. For instance, in a long-term study of adolescent drug usage the participants would get older and perhaps wiser. Maturation could also happen in relatively short-term studies. For example, participants might become tired or hungry during a two-hour experiment. Testing refers to the possible effect of taking a test once on taking it again. For instance, if students are given the same math test at the end of every week they might eventually learn to answer those exact questions. Instrumentation has to do with inadvertent changes in the measurements. For instance, a new set of counselors might be brought in partway through a study to evaluate depression levels, and they could use different standards. Statistical regression happens when groups with extreme characteristics move toward more ordinary values. For instance, if a group of extremely obese people were in a weight-loss study, some of them might naturally lose some weight over time because there is not much room to gain more weight. Selection biases have to do with who gets assigned to the groups. In other words, participants in the experimental group may differ from those in the control group to begin with. For instance, if study administrators make sure their friends get into the treatment group, this would make the two groups incomparable. Experimental mortality refers to differing levels of dropout in the two groups. For instance, if everyone stays in the control group but half the people in the experimental group resign from the study, the groups are no longer comparable. And finally, selection-maturation interaction is the possibility that people selected for the two groups have different maturation experiences. For instance, if the control group had normally sized people but the experimental group were composed of people with dietary issues, the second group might get hungrier and react differently in the experiment. What all of these internal validity threats have in common is that each one of them points to a different causal explanation of why the two groups might have had different results in the study. Random assignment, perfectly executed, should protect against all of these threats. That is why the “gold standard” for medical research designs requires random assignment (and also that both participants and experimenters are blind to each person’s assignment to one or the other group).

Assuming that these internal validity threats can be argumentatively nullified (or measured and accounted for), then we move to the question of external validity: "Given that we understand what happened within the study, can we be confident that it will also happen out in the real world?" In other words, external validity is concerned with the extent to which a study result can be generalized to a different population, a different context, and a different time. The issues here are reactive testing, interaction of the manipulation and selection, reactivity of the experimental circumstances, and multiple-treatment interference. Reactive testing refers to the possibility that the actual testing makes the participants unrepresentative of the general population. For instance, if people in the study have to weigh themselves every morning, that might make them more focused on their weight than the general population. Interaction of the manipulation and selection is the possibility that the people selected for participation in the study react differently to the manipulation than the population would. For instance, if the study consisted of people who are allergic to the ordinary medicine for some condition, their reaction to the new drug might not generalize to people who aren't allergic to the standard treatment. Reactive effects of the experimental circumstances is the chance that people act differently in the experiment than they would in ordinary life. For instance, if participants in a depression study were brought to a lab full of scientific equipment and were attended by considerate professional personnel, they might be so taken with their general treatment that they became more optimistic than they would ordinarily have been, and therefore reacted better to the drug or counseling than people in a non-experimental setting will. Finally, multiple-treatment interference happens when multiple treatments are used in the study. For instance, if people are first counseled and then given a drug, we could not be sure that the drug alone would have had the same effects. These external validity issues point to arguments that might justify setting aside the experiment's results when trying to understand how the overall population would react to the manipulation.

These issues summarize what might well be most of a course in study design, but what little we have said here should illustrate the kind of critical thinking that can be done about a study announcing some new generalization about diet, medicines, counseling, exercise, sleep patterns, or the many other things that journalists and bloggers are now reporting. These issues are likely to have been explicitly addressed in the original research report if the experimenters themselves were worried about the issue. Most experimenters are happy to do this because it justifies someone funding them for the next study. Corporations, however, might like their drug studies or other products never to be questioned so that they can persuade people to request the product. Therefore it is well to pay attention to who did the study, where it was pub-

lished, whether there were conflicts of interest, whether the results can be replicated, and perhaps to wonder how that study out of all the studies on the topic was put in a journalist's hands to report on.

We do not mean to be encouraging cynicism about scientific experiments. Cynicism is when things are rejected out of hand ("All politicians lie all the time, so ignore anything Senator Smith says"). Skepticism, which we do encourage, is withholding judgment until the evidence and arguments have been carefully examined.

As with surveys, people can do little approximate experiments themselves. You can set out a piece of fruit and a piece of pie to see which one your nephew will take first. You can make hints for your boyfriend to see how he will react. You can organize a group of fellow students to laugh at an instructor's jokes only when she or he is looking at the ceiling. In each of these cases, you will have done a manipulation but you will be implicitly comparing your results to what you would have expected based on your experience, you will have a tiny sample size, you will not have a randomly assigned control group, and you will have taken many other huge shortcuts that would not be allowed in a real experiment. This doesn't mean that your results are useless, but you should have great reserve about trusting or generalizing them too far.

5. Meta-analyses

In the early 1970s, one of us was taught in graduate school that "One study never proves anything." By itself, this is an over-generalization. But as a guiding principle for reading individual studies it is a good impulse because it encourages reserve and proper skepticism. When evidence cumulates over many studies, generalizations from that data become increasingly reliable.

A meta-analysis is an "analysis of analyses." A standard survey or experiment gets its data from individual respondents. A meta-analysis gathers its data from the results of previous surveys or experiments. The outcomes of those studies are averaged and analyzed to produce conclusions that are based on many studies and large groups of respondents.

Meta-analyses became increasingly common in the last quarter of the 20th century. At first they were mainly offered as summaries of large research traditions and their merit was that they damped out the effects of occasional statistical flukes. But in the 21st century additional analytic methods have been developed to critique whole lines of research. We have recently become aware of "replication crises" in medical and psychological research. The problem is that certain results don't replicate—that is, that they can't be reproduced by studies that had a good chance to generate the original findings. The new techniques can detect if failed research has been withheld from

the public record, so that only the glittery results-revealing studies are published. The likelihood of missing studies can now be detected by examining the patterns of studies that do make it into print. If—and this is sadly not rare—the meta-analysis discovers that the big effects are reported with small-sample studies and that large-sample studies show lesser (or no) effects, the meta-analysis will now show this clearly. In other words, statistical scrutiny of the research record can now predict pretty well whether results will replicate or not. The techniques for doing this are complex and the scientific community is still in the process of choosing which ones deserve to be standard.

Meta-analyses are not yet reported in the media as commonly as dramatic single studies, but we can hope that this will change. A good meta-analysis is more trustworthy than a good survey or experiment. You can even do informal meta-analyses yourself. Perhaps you notice that most of your friends believe in some generalization, and then take on that generalization yourself. Although this can be an instance of the bandwagon fallacy, it may also be more reliable than trusting only your own impressions, especially if you have some reason to respect the views of the people who are influencing you.

6. Summary about Generalization-establishing arguments

And that brings us to a key point about generalizations generated by arguments, whether the argument takes the form of a survey, an experiment, a meta-analysis, or some less disciplined real-life activity (such as noticing a couple of examples of something or feeding your dog weird food for a day). Unless your reasoning was just irredeemably ridiculous, the generalization you came to should have *some* argumentative value. The question is how much. Generalizations, like other conclusions, need to be qualified so that they fit the strength of their argumentative support. A generalization can be "maybe true," "occasionally true," "possible, at least for people with histories of psychiatric upset," "often true for men in Western cultures," or something along those lines. The worst reasoning about generalizations happens when people let the generalization get loose from the argument that produced it. The best thing you can do when thinking in general terms is to keep in mind the nature of the generalization's support and the inevitable flaws in those underlying arguments. Weigh the flaws properly and carry that weighing forward to the generalization.

So what about stereotypes? The nasty ones—remember, we are only claiming "to know them when we see them"—are based on terrible arguments. Often the evidence is biased. Perhaps you notice and remember every flaw in a postmodernist philosopher's life but ignore or forget the very same flaws in a logician's life. Or maybe there was no evidence at all. Perhaps your mother warned you against postmodernists before you set off for college, and

you completely absorbed her warning that they are all pernicious, immoral youth-corrupters, without exception. Or perhaps you swallowed an unqualified generalization from your news feed without inquiring about the evidence that generated it. No one can investigate the basis for every opinion they hold—life is too short for that—but we are each responsible for what we believe and say. If you expose your generalizations to a good sample of the rest of the world, the world will let you know which of your opinions might actually be stereotypes, and those are the ones you need to investigate further. That is the nice thing about living in a civilly argumentative environment, rather than being completely self-enclosed or restricted to only like-minded familiars.

7. Generalization-using and Generalization-saturated arguments

To this point, we have been mainly considering arguments that conclude in a generalization. However, generalizations can appear elsewhere in an argument. An argument can *use* a generalization in its premises: Politicians say false self-serving things all the time; Prime Minister Cobbler says he is innocent of taking bribes; so Prime Minister Cobbler is lying. An argument can also be completely *saturated* with generalizations: Professors are all brilliant; brilliant people are all excellent friends; so professors all make excellent friends.

As far as we can tell, generalization-using and generalization-saturated arguments can take just about any form. They can be deductive, or causal, or practical, or many other things. Each of these other things is an argumentation scheme of its own. Every known argumentation scheme has its own form and its own associated critical questions. Douglas Walton, Chris Reed, and Fabrizio Macagno have written a book that details dozens of these schemes and lists their individual critical questions. All those critical questions, along with those you might generate if you were looking at an argument that didn't quite fit their typology, come down to the ARSD standards, one way or the other.

Our message about such arguments has to do with A, the acceptability standard. Every premise that expresses a generalization ought to have been established by some prior argument. Perhaps it is based on scientific methods such as those we have discussed. More probably, it is based on some informal approximation of those scientific forms of reasoning: intuition, impressions, life experience, something you tried once with success, and so forth. Scientific method and statistics are basically just codifications of the best and most reliable kinds of informal reasoning. After investigation of a generalization, you should be able to connect its basis to what the best possible standards for the argument that established that generalization ought to have been.

A flaw in an argument's premises interrupts the argument's movement toward its conclusion. However, "flaw" is a matter of degree. A premise would have to be astonishingly awful to have no value at all. Just as we have tried to show you how to weigh and qualify the generalized conclusions to arguments, you should be able to apply that advice to the premises of other arguments. If you encounter a generalization-using or saturated argument that contains generalizations that are unqualified, all or nothing, black or white, you should try to figure out what the qualification should have been. That qualification should propagate throughout the argument, so that a qualification attached to a premise should reappear somehow in the conclusion. An exaggeration in the premises should be edited so that the conclusion isn't exaggerated as well.

8. Conclusions

Generalizations are an inevitable and useful element of human thought and argument. It simply isn't possible to store every individual bit of data in our heads or list them all in an argument. Even if it were possible it would be boring and overwhelming. We need generalizations. By summing things up, generalizations allow us to estimate what is likely to happen, to appreciate what might be going on now, and to understand what occurred before.

Like other kinds of thought and expression, generalizations can be imperfect. We can have over-generalizations that lack proper qualifying, we can have under-generalizations that don't take proper account of important evidence, and we can have generalizations that are so poorly founded that they might as well be fictional. The essential way to evaluate and correct a generalization is to know where it came from, and to be able to evaluate its origins. Knowing the evidence for or against a generalization and the means that were used to generate that evidence, are essential to improving our own thought and our public discourse. The standards of acceptability, relevance, sufficiency, and dialectical quality can all be complex and challenging to apply, but they are essential in one form or another if we are to make and receive arguments as best we can.

Suggested Readings

Blair, J.A. and Johnson, R.H. (Eds.) (1980). *Informal Logic: The First International Symposium*. Inverness CA: EdgePress.

Campbell, D.T. and Stanley, J.C. (1966). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Cooper, H., Hedges, L.V. and Valentine, J.C. (Eds.) (2009). *The Handbook of Research Synthesis and Meta-analysis* (2d. ed.). New York: Sage.

Johnson, R.H. (2000). *Manifest Rationality: A Pragmatic Theory of Argument*. Mahwah NJ: Erlbaum.
Walton, D., Reed, C. and Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press.

About the authors:

Dale Hample has been publishing research on argumentation since the 1970s. He has written two advanced books on argumentation, *Arguing: Exchanging Reasons Face to Face* (2005) and *Interpersonal Arguing* 2018), and with William and Pamela Benoit has co-edited a book of essays, *Readings in Argumentation* (1992). He has published about 150 journal articles and book chapters. His discipline is Communication, which has led him to integrate quantitative social science with the humanistic traditions of rhetorical theory. He has given keynote addresses at argumentation conferences in the U.S., Canada, Chile, and the Netherlands. He is an Emeritus Professor of Communication at at the University of Maryland.

Yiwen Dai received her Ph.D. in Communication from the University of Maryland, College Park. Her research interests include argumentation and interactions in romantic relationships. She has published in *Argumentation and Advocacy* and *Western Journal of Communication* and presented her work at national and international communication conferences.