CHAPTER 11

# Reliability and Validity

## Learning Objectives:

- What is the meaning of reliability in scientific measurement, and what are the three main approaches to assessing reliability?
- What is the meaning of validity in scientific research, and what are the four main types of research validity?
- What is the relationship of reliability to validity?
- What are some ways to increase each of the four types of validity in research?
- What are researcher-case artifacts, and what are some ways to minimize their effects?

In this chapter, we discuss two fundamental concepts—reliability and validity—that are relevant to the collection and interpretation of all types of data in all areas of scientific research, including geography.

## Reliability

**Reliability** is the "repeatability" of scores or measured values of variables. Perfect reliability occurs when you measure something the same way twice and get exactly the same value each time, assuming the thing is actually the same at the time of the two measurements. For example, reliably measuring a person's weight means that the scale reads exactly the same value on two occasions when the person actually weighs exactly the same both times. In the language of Chapter 2, high reliability occurs when the manifest variable (the variable as measured) remains nearly unchanged whenever the latent variable (the underlying construct) remains unchanged. When two or more measurements of something are quite different, even though the thing has not changed much, the measurement is of low reliability.

According to measurement theory, any observed score or data value can be partitioned into two components: (1) a **systematic component** that reflects whatever underlying construct we are consistently picking up with our measurement operations and (2) a **random component** that reflects all influences on our measurements that vary nonsystematically from measurement event to measurement event. Low reliability is caused by the random component of measurement. An example would be the way census counts of populations in particular census tracts can vary a bit depending on which day and time of day a census worker attempts to count people. It's important to recognize that only random or nonsystematic errors of measurement cause lowered reliability. Errors in the systematic component are consistent across measurement events; they cause biased or inaccurate measurement, not unreliable measurement. If I step on an old-fashioned bathroom scale to weigh myself, the fact that I read off a slightly different number on the basis of the angle at which I happen to view the dial each time would cause lower reliability. In contrast, the fact that the springs in my scale are old and stretched out would cause the dial to indicate that I weigh some 15–20 pounds lighter than I really weigh, and it would do this consistently each time I step on the scale. As long as the reading is too light by the same amount each time, it is a reliable measurement that is inaccurate. In other words, this scale would be biased low—a poor measurement tool but good for the ego.

Low reliability has a variety of detrimental effects in research. Unreliable data contain more **noise**, which is random error. Noise in data decreases the precision of estimation and the power of hypothesis tests (Chapter 9). Noise weakens ("attenuates") measured relationships between variables; a correlation between two variables will be smaller when measurement reliability is poor, even when the relationship of the underlying constructs is the same. In the extreme, completely unreliable measurement precludes finding any relationships at all between variables—random variables are uncorrelated with one another. Ultimately, low reliability limits the validity of your research conclusions, although high reliability does not by any means ensure valid research conclusions. In other words, saying that data are reliable in the technical sense does not mean that you can necessarily *rely* on their truthfulness. High reliability is necessary but not sufficient for high validity.

Given this, it is important to think about measurement reliability in research. In many types of research, it's a good thing to explicitly assess reliability. If it is too low, you can take steps to increase it. If it is satisfactorily high, you can report that, which will help convince reviewers of your research that you are measuring something in a high-quality and dependable manner. Reliability is quantified as some type of correlation or proportion of agreement between two (or more) sets of measurements of the same thing. This is a little tricky. According to measurement theory, repeated measurements of perfect reliability agree exactly only when the measurements are done in exactly the same way on the same entity at exactly the same time, which is impossible (sort of the measurement equivalent of the cryptic wisdom that you cannot step into the same stream twice). For example, I can't put two rain gauges in exactly the same place at exactly the same time. In practice, researchers

assess reliability by applying a compromise approach to measuring something more than once. They take one of three specific approaches to assessing reliability via repeated measurement, none of them ideal:

1. ***Remeasurement reliability.*** In this approach, the researcher simply measures the same variable using the same measurement procedures on the same cases, but not at exactly the same time. In the context of explicit reports with humans, it is usually called "test-retest reliability." It is not an ideal measure of reliability in every situation insofar as the repeated measurements are not done at exactly the same time. This is problematic insofar as the underlying construct may change over time so that you end up measuring something a little different on the two occasions; for example, measurements of air temperature change over the course of a day. In some situations, you may need to take steps to deal with the fact that measurement can change the case being measured. If so, use alternate ways of measuring that are considered equivalent on repeat measurements. In the extreme, measurement can even destroy the case; for example, as we mention in Chapter 14, biogeographers have sometimes killed their cases in order to study them, which would obviously preclude this approach to assessing reliability.

2. ***Internal consistency reliability.*** This approach can be applied when a single construct is operationalized via multiple measurements to begin with. For example, place attitudes are typically measured with a survey that asks people several questions concerning their beliefs toward a place, not just one. The effectiveness of a geographic information display would be based on several questions or several tasks involving the display, not just one. The multiple measurements are meant to be redundant, partially overlapping, measures of the same construct. If so, scores on the separate measurements should largely agree with each other. For instance, responses to half the measurement items could be correlated with responses to the other half, an approach known as "split-half reliability." The most sophisticated form of the internal consistency approach correlates the score for each item with that of all the other items combined; these are averaged to produce a single **item-total correlation**[1] ("coefficient alpha," which is unrelated to the $\alpha$ probability introduced in Chapter 9).

3. ***Inter-rater reliability.*** The final approach to assessing reliability is applicable whenever one's data are based on judgments by researchers or their assistants. Such data usually come from the coding of open-ended explicit reports or nonreactive measures such as physical traces, behavioral observations, and some archives. With such data, reliability is assessed by looking for the amount of agreement between

[1]Incidentally, item-total correlations are very useful for developing the test or survey in the first place. Items that have very low correlations with the other items, below .3, are poor items and probably should be omitted. Negative item-total correlations mean that the item has to be reverse-scored before it is combined with the other items.

two (or more) observers or coders who work on the same subset of records (that is, the same plots of land, the same videos of behaviors, and so on). This approach to reliability is sometimes called "inter-observer" or "inter-coder" reliability. We considered this approach in some detail in Chapter 5, where we discussed the coding of open-ended records.

*Perfect* reliability would be great, but it is unattainable. That's not true just in human geography, or in other social and behavioral sciences, but in any basic or applied field of science, including physical geography. But *high* reliability is certainly attainable. What constitutes "high" does vary from scientific field to scientific field; it is generally higher in the physical sciences than in the life sciences, where it is often higher than in the social and behavioral sciences. Even within specific disciplines, reliability standards vary quite a bit across topical domains, which is almost entirely because of the fact that potential reliability varies so much as a function of what and how you are measuring. The depth of the ocean within a particular area just offshore can be measured more reliably than the average concentration of a pollutant in that area. The number of people living in a household can be measured more reliably than their attitudes about using mass transit. It is nice to measure with reliability of well over 90% agreement across measurement occasions, and even in the less reliable domains of human geography, at least 80% agreement is preferable. When reliability is lower than this, steps should usually be taken to improve it. Again, we stress that in some domains, high reliability is elusive and very difficult to attain. If reliability is too low, however, there is no point to measuring and doing science in a particular domain—random numbers do not make useful data.

How can we increase measurement reliability? A basic dictum of measurement theory is that combining multiple imperfect measurements of the same thing produces a single measurement that is more reliable, as long as the measurements are not *completely* imperfect. So you can increase reliability by increasing the number of items, observations, tasks, raters, and so on. Reducing the amount of noise in each measurement is obviously a good way to increase reliability. Thus, you can increase reliability by increasing the quality of your measurement procedure—use better sensors on satellites, make sure the thermometer is placed away from intermittent influences such as occasional direct sun, clarify survey questions to get respondents to understand them more uniformly, clarify coding procedures for open-ended data, and so on. Another approach to increasing reliability is to increase and improve the training of all people who help generate data, including assistants who count, interview, code, and so on.

## Validity

**Validity** is the "truth-value" of research results and interpretation. Given that research is largely an attempt to increase the truthfulness of our understanding of the world, validity is a core concern for researchers. A widely used typology of validity in research includes four classes: internal, external, construct, and statistical conclusion.

## Internal Validity

**Internal validity** concerns the truth of conclusions about causal relationships. As we discussed in Chapter 7, causal conclusions are thrown into doubt when there are possible alternative causal variables other than the variable we conclude is causing patterns in our data. That is, "threats to internal validity" occur when researchers interpret a statistical relationship between two variables in their data as indicating a causal relationship, when in fact some other variable is the real cause. We also pointed out in Chapter 7 that increasing internal validity is the primary purpose of empirical control in research. Our discussion there showed that the likelihood of high internal validity in a particular study varies greatly with its research design; internal validity is generally maximized in experimental studies as opposed to non-experimental studies. Furthermore, particular threats to internal validity, rather than others, are more or less likely to operate in a study depending on its research design.

There are many specific potential threats to internal validity in research studies. In Chapter 7, for instance, we discussed some specific threats that occur in developmental research designs, including history, cohort effects, and differential mortality. There is one especially interesting and subtle threat to internal validity that we want to consider in detail here, however. **Statistical regression to the mean** is the phenomenon that extreme scores, those far from the central tendency of the data set, tend to be less extreme when remeasured (often called "regression" for short, the phenomenon is the basis for the name of the general statistical technique of regression analysis mentioned in Chapter 9). For example, a year in which a city has an exceptionally high number of new housing starts will probably be followed by a year with somewhat fewer housing starts, other things being equal. In the same way, a year in which the city experiences an exceptionally low number of new housing starts will probably be followed by a year with more.

The critical point is that these changes in housing starts from one year to the next will tend to occur even without an economically or socially substantive cause, such as a change in interest rates or the city's population. The principle that extreme measurements tend to be followed by less extreme measurements—they *regress* back toward the mean—applies over and above any substantive factors that may operate to cause less extreme values. The phenomenon was identified and named by Francis Galton, the 19th-century English scientist and cousin of Darwin. He observed that sons of very tall fathers tend to be shorter (sons of short fathers tend to be taller as well). This phenomenon happens even when mothers themselves are not exceptionally short. Another standard example is the batting average of baseball players, which is the percentage of "at-bats" in which the player gets on base, over the course of the season. After a month of the season, a small handful of major league players may have averages over .400, which is a very high average (showing how tough it is to hit a major-league pitch). But as the season continues, these averages come down one by one, so that by the end of the season, no player has an average over .400; at least it has gone that way since Ted Williams finished the 1941 season with an average of .406. Why does this seasonal pattern nearly always happen? Is it because pitchers come to figure out hitters they have seen

earlier in the season? Is it because the days get hotter, which allows pitches to move to the plate faster? Do batters tire as the season wears on? Although some of these factors may operate, the major explanation is regression to the mean. The good news is that exceptionally poor batting averages will probably increase as the season goes on, at least for those players who don't get benched.

Why does regression to the mean happen? The answer is that the random component of a set of measurements expresses itself in a likely manner—moderately—after having expressed itself previously in an unlikely and extreme manner. One variable can show regression over repeated measurements whenever its reliability is less than perfect. Even more generally, a variable can show regression with respect to another variable, so that cases with extreme scores on one variable have less extreme scores on the other variable. This can happen whenever the two variables are less than perfectly related; as we learned above in this chapter, unreliability is a special case of a variable being imperfectly correlated with itself. When relationships are imperfect, one variable can only partially be used to predict the other. The two variables share some but not all of their variation. The part they do not share is random variation—noise. It's this noise that leads to regression, first by randomly acting unusually and contributing to a very high or low score, then by randomly acting more usually, neither very high nor very low.

But why is moderately valued noise more common than high or low noise? If the noise is random, aren't all of its possible values equally likely, as implied in our discussion of random sampling in Chapter 8? The answer is no, because the noise is not one thing but a *collection* of things; in the parlance of probability theory, a noise value is an "event" rather than an "elementary outcome." Take a thoroughly mixed jar of 50 red balls and 50 white balls, all balls equal in size. Given a random selection process, each individual ball has an equal chance of being selected—1% on the first pick. (Because our jar has an equal number of red and white balls, the events "red" or "white" on the first pick are also equally possible, namely, 50%.) But every possible combination of 10 balls after 10 picks, classed in terms of their color, certainly does not have an equal chance of occurring. Applying some basic probability algebra for the joint probabilities of discrete outcomes sampled without replacement (you don't put the ball back in the jar after you pick it), it turns out that picking 10 balls that are all red is very low, less than 0.1 % (one-tenth of a percent). Picking all white balls has the same tiny probability. But picking five red and five white balls is rather likely, with a probability just over 25%. Similarly, picking four reds and six whites, or vice versa, is also quite likely, at just over 21%. So very high or low noise is like a very red or very white sample of balls—unlikely. Moderately valued noise, in contrast, is like a balanced sample of nearly equal numbers of red and white balls—quite likely. (This is the same reasoning that explains why, as we pointed out in Chapter 9, a sample with a mixture of high, medium, and low scores is more likely than a sample with all high scores or all low scores.) Regression happens, therefore, because extreme scores on one variable or after one measurement are partially due to unlikely noise, which will probably be less extreme on the other variable or the second measurement by chance alone.

Given that regression is the expression of a random phenomenon, its chance of occurring in a given data set increases as the number of measurements increases.

It is likely with one or a few measurements but effectively certain with a large aggregate of measurements. Regression to the mean is a threat to internal validity because its attendant rise or drop in scores is often misinterpreted as being caused by something meaningful and substantive, not the whims of chance. In our housing example above, no politician worth his or her salt would miss an opportunity to take credit for a jump in a slumping market or blame his or her incumbent opponent for a decline in a booming market. Similarly, an exceptionally hot or wet year is virtually always followed by a cooler or dryer year, even if no long-term climate change is occurring as a result of human activity or any other systematic cause.[2] We hope you now know better than to accept such claims at face value.

## External Validity

We pointed out in Chapter 9 that researchers want to generalize their results and conclusions beyond the cases, measures, settings, times, and so on, that they actually use in their studies to other cases, measures, settings, times, and so on. **External validity** concerns the truth of these generalizations. In other words, external validity is the validity of inferences drawn from samples to populations (Chapter 8). External validity is clearly influenced by all aspects of how you sample, because they influence the resulting relationship of your sample to your sampling frame and the relationship of your sampling frame to your target population. And as we made clear in Chapter 8, it is not just the sampling of cases that is relevant but places, times, measures, and all other aspects of research. Thus, external validity is increased by large and representative samples of cases, places, times, measures, and so on.

However, external validity is also influenced by how the particular research settings and materials you use are specifically like the settings and materials involved in the phenomenon you study under its "natural" conditions. That is, in many research domains within geography, questions about external validity are thought to depend on how realistic the research setting was—to what degree were the conditions under which empirical observations were made similar to the normal conditions under which the phenomenon of interest exists or expresses itself. This issue of research realism or "verisimilitude" is often called **ecological validity**. It is called *ecological* because it is based on the idea that certain phenomena express themselves in a certain way because of the totality of the context in which they exist, including aspects of context that are spatial, temporal, or thematic. A geographer who studies the spread of plant communities in a laboratory might misunderstand aspects of this phenomenon because the laboratory fails to mimic the solar or atmospheric conditions that normally hold in the real world. Or a geographer who studies how analysts interpret remotely sensed images might misunderstand this phenomenon

[2]Just to be extra clear on this point, the reality of the phenomenon of regression to the mean by no means precludes the possibility that real substantive and meaningful changes over time in the phenomenon of interest are taking place. The fact that climate naturally changes over time, for instance, does not prove that humans are having no effect on climate themselves—it just proves that the phenomenon of climate change per se does not prove anthropogenic influence.

because the tasks given to the analysts seem "artificial" to them and thus do not engage their normal reasoning processes. Thus, ecological validity is primarily a concern for domains in which phenomena exist within a context that has important implications for the expression of those phenomena. In other words, it is a concern for most geographers; those who work in relatively abstract areas of physical geography close to physics or chemistry are probably safest in ignoring context, or at least in having a clear understanding of which aspects of context are relevant and need to be controlled.

A concern for ecological validity, as a way to ensure the external validity of conclusions about phenomena to the "real" world, is one reason to collect data in the natural context of phenomena, using methods that do not change the phenomena as a result of measurement (for example, field observations). It must be recognized, however, that a research setting high in ecological realism is not necessarily highly generalizable to many different contexts. The geographer who mimics the ecological conditions in one particular watershed may be able to learn about phenomena in that watershed but may be unable to generalize confidently to other watersheds. In fact, high ecological realism may interfere with generalizability by creating research conditions that are too narrow and specific. Thus, high ecological validity does not ensure high external validity, and high external validity need not require high ecological validity. That is, high ecological validity is neither necessary nor sufficient for high external validity. Furthermore, it should be recognized that generalizing to normal or natural conditions is not always the primary goal of a scientist. Testing theories about causal relationships across a range of conditions may require carrying out research in highly unusual or even hypothetical contexts. Or one may carry out research to show what is *possible* rather than what is likely or normal.[3]

Ultimately, questions about external and ecological validity lead us to one of the ubiquitous dilemmas of science. Scientists want to make statements of truth that are *general*, as much as possible. That is an important implication of the characteristic belief in simplicity shared by scientists (see Chapter 1). But all research studies that are actually carried out are finite, and most of them are in fact quite modest in scope. Scientific researchers must always use particular places, cities, rivers, measurement instruments, and other research entities in their studies, even though they want to conclude things about larger sets of these entities. That is, scientific research requires sampling, as we discussed in Chapter 8. So scientists always want to say something that goes beyond the direct evidence they actually have. Such inferences cannot be avoided even though they can always be mistaken.

## Construct Validity

**Construct validity** concerns the truth of how variables as operationalized for measuring represent the theoretical constructs they are supposed to represent; that

[3]A thought-provoking and humorous critique of the idea that research should always be designed to generalize widely or is necessarily faulty if it fails to imitate "natural" reality closely is provided by Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379–387.

is, it concerns how well manifest variables reflect latent variables. In Chapter 2, we learned that constructs or latent variables are the hypothetical entities that we attempt to measure in research. They are pieces of the idealized world that make up the subject matter of our theories. Measured or manifest variables are these entities as they are expressed by actual measurement procedures.

The idea of construct validity reflects the recognition that the scores in our data sets are *attempts* to assess the values that cases actually possess on our constructs of interest, attempts that are nearly always a little mistaken, maybe very mistaken. Above, we pointed out that any observed data value can be partitioned into a systematic component and a random component, and that low reliability is caused by large error in the random component. In contrast, construct validity is about the systematic component in data, specifically the *meaning* of the systematic component. Low construct validity is a mismatch between what you are trying to measure according to your conceptual framework and what your measurement procedure actually measures consistently. That is, only reliable components of variation in your data are relevant to construct validity.

Low construct validity is fundamentally construct "misrepresentation." It arises from some combination of missing aspects of your construct that are part of its essential meaning, and including aspects that are irrelevant or even contradictory to its essential meaning. In Chapter 2, we discussed the poor construct validity in John's research presented from Chapter 1. By operationalizing the construct of "dissociative institutions" by counting liquor stores in the phone book, for example, John missed much of the actual meaning of dissociative institutions at the same time that his data contained variation not relevant to the meaning of dissociative institutions. Similarly, biogeographers who operationalize the construct of "plant communities" by measuring only latitude and climate conditions will miss out on some aspects of the meaning of that construct at the same time they pick up variation in their data that is not relevant to which plant community is present in a particular place.

Of course, scientists attempt to increase construct validity by operationalizing variables (designing operations to measure constructs) in a thoughtful and carefully reasoned way that is informed by previous research and scholarship. A systematic approach to establishing and increasing construct validity is called **convergent-discriminant validation**. It is based on the fact that poor construct validity means that you are failing to capture some of your construct in your measurement, you are capturing some other constructs in your measurement, or both. To perform a convergent-discriminant validation, you attempt to measure multiple constructs with multiple measurement operations, that is, multiple manifest variables. Logically, variables that should conceptually tap into the same construct should largely agree, that is, should correlate highly across cases—this is the convergent part. By the same token, variables that should conceptually tap into different constructs should largely disagree, that is, should not correlate highly across cases—this is the discriminant part. In other words, you explore construct validity by measuring in more than one way and by measuring in ways that you do not expect to be relevant to your constructs of interest. Even if you do not apply a full-blown convergent-discriminant validation in your research, you should still avoid resting your

research conclusions on a single way of measuring your most important constructs. No measurement technique is perfect, so measuring in multiple ways is the best way to avoid the dangers of drawing conclusions from any single technique—the so-called **mono-method bias**.

## Statistical Conclusion Validity

**Statistical conclusion validity** concerns the truth of conclusions you draw from statistical analyses of data, a function of the appropriate use and interpretation of statistical tests. We discussed this in Chapter 9, but as we pointed out there, you should learn more about it in courses dedicated to statistical analysis. Issues of statistical conclusion validity include applying the correct descriptive indices to data at different levels of measurement; for example, the mean of a nominal variable is meaningless. Another issue concerns the truth of assumptions made in inferential statistics that allow the construction of the correct sampling distribution for determining probabilities about population inferences from samples. These include assumptions about the population distribution, the independence of scores, and correct and complete statistical model specifications.

There are statistical tests for the distributional assumptions. Some texts recommend applying these tests to your data before analyzing them. For instance, we can test the normality of a population distribution by examining the normality of a sample taken from that population. As we pointed out in Chapter 9, these distributional assumptions apply only to parametric statistical tests. Many researchers prefer to restrict themselves to nonparametric tests when these assumptions are shown to be dubious. Of course, these tests of assumptions about populations are based on data from samples and are therefore inferences themselves. Furthermore, several of the standard assumptions for parametric inferential tests turn out not to be very important; that is, a violation of population normality or homogeneity of variance across groups has very little distorting effect on the probabilities found as part of inferential statistical tests. In other words, the inferential tests are robust to violations of many of the standard assumptions.[4] Parametric tests can nearly always be used safely, even with samples that are fairly nonnormal or that reflect heterogeneous variances. Given this, we recommend using parametric analyses over nonparametric analyses in most situations—they are more flexible and typically more powerful.

There are statistical tests for the independence assumption too. In contrast to distributional assumptions, independence is something you should take quite

[4]Evidence for the robustness of statistical tests to violations of certain distributional assumptions comes from statistical Monte Carlo "experiments" like those discussed in Chapter 7 in which simulated sampling distributions are created by drawing a large number of random samples from hypothetical populations with particular properties. For example, the "central limit theorem" states that a population must be quite nonnormal and samples must be small before the sampling distribution of the sample mean will deviate much from normalcy itself and produce distorted probability values.

seriously. Nonindependence means that you actually have much less independent information than you might think. This can lead to radically mistaken estimates of inferential probabilities concerning what sample results indicate about population values. It is widely recognized that nonindependence in geographic research is often based on spatial relations, as we discussed in Chapter 9, but it is also frequently based on temporal relations or relations of identity, as we discussed in Chapter 7; two cities measured once provide more independent information than one city measured twice.

There are two other common threats to the validity of conclusions about statistical significance when using the inferential technique of hypothesis testing (Chapter 9). The first is **inadequate power**. It is possible whenever you make a decision to retain both the null and alternative hypotheses—that is, when you do *not* obtain statistically significant results. If you fail to obtain significance, it is always possible that you have made a mistake (a Type II error in Figure 9.5). You are likely to make this mistake when beta ($\beta$) is too high, or equivalently, when power is too low. So whenever you fail to obtain significance, there is always the question of whether your test had adequate statistical power to find an effect that is really there in the population. The statistical technique of power analysis introduced in Chapter 8 is the formal way to estimate how much power you have in a given hypothesis test, and how you can increase it to a particular level.

A second common threat to the validity of statistical conclusions about significance is, in a sense, the converse of inadequate power: the possibility of **alpha ($\alpha$) inflation** whenever you conduct multiple significance tests on the same set of data. Alpha inflation is a potential threat to validity when you make a decision on some of your tests to reject the null and accept the alternative hypotheses—that is, when you *do* obtain statistically significant results. If you obtain significance, it is always possible that you have made a mistake (a Type I error in Figure 9.5). You are likely to make this mistake after multiple tests because the $\alpha$-level per test may be .05, for instance, but the effective **studywise $\alpha$** will be higher, much higher if a large number of tests are conducted. That is, the chance of making at least one Type I error in a set of tests is higher than the chance of making one on just a single test. To appreciate this intuitively, imagine that you are drawing a single card from a standard deck, and you are concerned about getting one of the two black Jacks (you harbor some superstitions). On one draw, the chance of getting a black Jack is only 2/52—less than .04. However, if you are forced to draw five cards all at once, your chance of getting hexed by at least one of the black Jacks is much higher—around .20. So whenever you conduct several significance tests on the same data, there is always the question of whether one or more of the significant ones reflect an inflated studywise $\alpha$. This problem is much more serious when one conducts a large number of comparisons, even all possible ones, on a single data set without theoretical expectations to guide which tests to conduct and how to interpret their meaning. Such an approach is charmingly known as "going fishing." Unfortunately, such an atheoretical approach to empirical work is all too common in geography and likely has produced a great deal of apparent evidence for phenomena that are not true, or at least that are not very strong. The antidote to $\alpha$ inflation is using one of several possible techniques to reduce the $\alpha$ per test and

therefore the overall studywise $\alpha$ (see references in Bibliography). Better yet, try to perform statistical tests that are theoretically guided—leave the fishing for seafood suppliers and recreational anglers. At the very least, phenomena discovered on a fishing trip must be replicated in new data sets.

## Researcher-Case Artifacts

We finish with a brief discussion of certain special classes of validity threats that arise because research is an activity that can influence ongoing reality, and researchers are humans that can be influenced by their understanding of a situation as being research. In human geography, researchers' cases are often humans that can also be influenced by a situation as research. These **researcher-case artifacts** can produce biased data or data interpretations that reflect various expectancies or beliefs people, whether researchers or human research subjects, have about research situations, about outcomes, or about classes of people being measured or doing the measurement. They also derive from various aspects of the activity of the research situation. That is, researchers can act in ways that alter their phenomenon of interest, or they interpret that phenomenon in biased ways. These artifacts potentially threaten any of the four types of validities.

Researcher-case artifacts can be grouped into two classes, interactional and noninteractional. **Interactional artifacts** result from the interaction of the research situation (including the researcher himself or herself) with the case, providing an opportunity for the research situation to directly influence the cases and their values on the constructs being measured as data. That is, interactional artifacts actually change the value of the case on the construct being measured. In human geography, one usually considers the danger of interactional artifacts when the cases are sentient creatures, such as individual human subjects who are directly contacted by a researcher. Nonhuman animals studied by biogeographers could also show these influences, but interactional effects in general are undoubtedly rare in physical geography. An example could be a geographer who is trying to show that soil interflow depends on ground slope.[5] He or she might excavate a soil pit and insert measuring devices at various levels to intercept the water flow. The faces of the pit will act as "macropores," however, that alter the pressure relations experienced by the flow and possibly cause water that would have flowed downslope not to flow. This will lead to an altered indication of flux in the water.

The classic type of interactional artifact in human research occurs when research subjects change their behavior or their responses to explicit reports because of the way they interpret the meaning of the research situation, including its purpose, its risk, its relevance to their reputation or class grade, its political bias, and so on. In other words, the concept of reactance that we defined in Chapter 4 is an interactional artifact. This expression of interactional artifacts reflects a definition of research as

[5]This example comes from Reid (2003), referenced in Chapter 4.

seen from the perspective of a research subject[6]: Research takes place in "a context of explicit agreement to participate in a special form of social interaction known as 'taking part in a study.'" In this situation, artifacts can be brought about by the way researchers ask questions, the identity of the agency or organization sponsoring the research, or just the name or stated purpose of the study. Even the appearance of the researcher could affect subjects' responses if data collection occurs during an episode of direct contact between the researcher and the subject. Examples include a researcher's apparent sex, age, ethnicity, or physical attractiveness.

**Noninteractional artifacts** result from the researcher himself or herself, not from the interaction of the research situation with the case. They change the recorded values of data or the way they are interpreted but do not actually change the value of the case on the construct being measured. Instead, they change the values or interpretation of data as a result of bias or limitations in the way researchers perceptually and cognitively process information. To some extent, researchers see what they expect or hope to see (we discussed the "nonobjective" nature of human cognition in Chapter 6). Sometimes researchers might even intentionally bias the way they record or analyze data because of human foibles such as greed or hubris, which, we noted in Chapter 1, scientists suffer from as much as anyone else. Such intentional bias obviously violates ethics in research, which we return to in Chapter 14. A good example of an unethical noninteractional artifact would be making a decision that an observation is an outlier and deserves to be removed from the rest of your data only *after* you have seen that its inclusion weakens support for your hypothesis. Noninteractional artifacts potentially occur in much the same way and to the same degree in any area of geographic research. In general, they would be more common whenever data creation and analysis have less strict and predetermined rules for their conduct—that is, whenever the researcher's personal judgment plays a greater role determining data treatment. In some contexts, such as the coding of unstructured interviews, that would be more common in human geography. This is one of the important reasons for preferring systematicity in data treatment.

How can we reduce the occurrence and impact of researcher-case artifacts? With respect to interactional artifacts, it is important to recognize that they do not necessarily occur and do not occur equally to everyone. Just because a person is aware he or she is being questioned for a study, for example, does not guarantee that the person will deceive the researcher about the truth, as he or she understands it. Likewise, most of us change our behavior somewhat when we know are being watched, but skilled and experienced public speakers or performers often demonstrate the ability to avoid this change.

Having made that point, interactional artifacts such as reactance are a real possibility. Avoiding them is a major motivation for using so-called nonreactive measures in human geographic research (as we discussed in Chapter 4). Research subjects cannot change their behavior, thoughts, or feelings in response to the researcher or

[6]From Rosenthal & Rosnow (1991).

research situation if they are not aware they are in a research study, or if the data are not in fact produced as part of a research study. If nonreactive measures are not possible, such as when your research question deals with constructs like explicit attitudes, then you need to spend a great deal of effort during pilot testing and research design to produce unbiased instruments. In some situations, it would be worthwhile to show that the same data result when items are designed with alternative wordings. Sometimes it is appropriate to keep assistants "blind" to the specific hypothesis of the research or to the specific condition to which a subject has been assigned. This helps with noninteractional artifacts too, as when open-ended data coders are kept blind during coding. You should also consider the characteristics of assistants or interviewers, such as their sex or ethnicity, that could influence subjects' responses. Sometimes you should counterbalance these characteristics by making sure to use both male and female assistants, for example. In other situations, perhaps you should use only assistants with particular characteristics.

In the case of noninteractional artifacts, it is clear that new researchers need to be trained, through both explicit instruction and exposure to role models, in the ethical scientific values of neutrality, objectivity, honesty, and so on. We are not so naive as to think that such training, even when very high in quality and quantity, guarantees that researchers or their assistants will always treat data properly and interpret it objectively. Neither are we so jaded to believe that personal integrity in research is mostly a lost cause. We try to conduct our own research according to these values and believe that we almost always succeed at it. Ultimately, however, the social nature of science that we discussed in Chapter 1 comes to the rescue here. The fact that scientific results and conclusions are necessarily and appropriately subjected to critical scrutiny by other researchers at other places goes a long way toward muting the threat of researcher artifacts. Findings should be skeptically doubted (with justification) and should be independently replicated by other researchers with neutral or even opposing views.

# Review Questions

## Reliability

- What is reliability, and what causes high and low reliability of measurement?
- What are three specific approaches to assessing reliability, and why does each fall short of ideal for assessing reliability?
- What are some ways to increase measurement reliability?

## Validity

- What is validity, and what are the names and meanings of the four major types of validity?
- What is regression to the mean, why does it happen, and how does it influence internal validity?
- What is ecological validity, and how does it relate to internal and external validity?

- How is poor construct validity different than low reliability?
- What are three assumptions generally involved in valid statistical inference (introduced in Chapter 9)? How important is it for researchers to attend to threats to statistical conclusion validity stemming from violations of these assumptions?

## Researcher-Case Artifacts

- What are researcher-case artifacts, and how might they influence each of the four types of validity?
- What is the distinction between interactional and noninteractional artifacts, and what are some examples of each in geographic research?

# Key Terms

**alpha ($\alpha$) inflation:** threat to statistical conclusion validity when multiple hypothesis tests are conducted on a single data set, and some of them are significant; results from the overall increased studywise chance of a Type I error when multiple tests are conducted

**construct validity:** the truth of how measured (manifest) variables capture the meaning of the theoretical constructs (latent variables) they are supposed to represent

**convergent-discriminant validation:** systematic approach to establishing and increasing construct validity by measuring constructs with multiple variables, some of which are intended to redundantly measure the construct and others of which are meant to measure other constructs

**ecological validity:** the truth of research generalizations as a function of how well the context within which a phenomenon exists under natural conditions is present in a particular study; an aspect of external validity

**external validity:** the truth of research generalizations from the samples of cases, measures, settings, times, and so on, actually used in a study to wider populations of cases, measures, settings, times, and so on

**inadequate power:** threat to statistical conclusion validity when nonsignificant hypothesis tests may be due to power that is too low

**interactional artifact:** researcher-case artifact that changes a case's value on the construct being measured because of the interaction of the research situation (including the researcher) with the case

**internal consistency reliability:** approach to assessing the reliability of a complex measured construct that is operationalized as the combination of two or more separate but partially redundant variables (as in surveys); reliable measurement requires the separate variables to correlate with each other

**internal validity:** the truth of research conclusions about causal relationships

**inter-rater reliability:** approach to assessing the reliability of a measured or coded variable in which the data produced redundantly by separate observers or coders are compared for similarity

**item-total correlation:** way to quantify internal consistency reliability by averaging the correlation of each separate variable with all of the others in the set meant to measure a single construct

**mono-method bias:** shortcoming of research that relies on drawing major conclusions from a single way of measuring constructs

**noise:** another term for random error of measurement

**noninteractional artifact:** researcher-case artifact that changes data values or their interpretation, not the actual values of the cases on the constructs of interest, independently of any interaction of the research situation (including the researcher) with the case

**random component:** the component of a measured score value that is inconsistent across measurement events, reflecting nonsystematic influences on measurement operations; this random component causes lowered reliability and can be called noise or random error

**reliability:** the "repeatability" of measured values of variables; perfect reliability occurs when remeasuring a construct that is exactly the same on both occasions results in exactly the same measured value

**remeasurement reliability:** approach to assessing the reliability of a measured variable in which cases are measured twice or more at different times, and the resulting measurements are compared for similarity; sometimes called "test-retest reliability"

**researcher-case artifact:** threat to any of the four types of validity that arises because human researchers and/or research cases have various expectancies or beliefs about the research situation, outcome, or about classes of people being measured or doing the measurement

**statistical conclusion validity:** the truth of conclusions we draw from statistical analyses of data, a function of the appropriate use and interpretation of statistical tests

**statistical regression to the mean:** the phenomenon that extreme measured scores tend to be less extreme when remeasured, due to chance variation acting in a more likely manner after it acted in an unlikely manner; it is an especially intriguing threat to internal validity

**studywise $\alpha$:** the $\alpha$ for making at least one Type I error among multiple tests conducted on the same data set; $\alpha$ inflation is due to the fact that the studywise $\alpha$ is generally larger than the $\alpha$ per individual test

**systematic component:** the component of a measured score value that is consistent across measurement events, reflecting whatever construct we are picking up with our measurement operations; when this systematic component is inaccurate relative to the true value of the construct, it can be called "systematic error"

**validity:** the "truth-value" of research results and interpretation

## Bibliography

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Hand, D. J. (2004). *Measurement theory and practice: The world through quantification*. London: Hodder Arnold.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.