Measurement

In every study, researchers measure behavior. The behavior could be *overt* (obvious, easily visible) such as bumping into someone or saying something, or it could be more covert, such as an increased heart rate or looking at a particular location on a computer screen. In some studies (such as experiments), researchers may first manipulate respondents' environments before measuring their behavior, but in every study, researchers measure behavior. In this chapter, we focus on some of the central issues you should consider in measurement.

Scales of Measurement

A choice you will need to make early on is the "scale" you will use for your variable. This decision will influence the sensitivity of your measure and the kinds of conclusions you can make, so it's important to choose a scale that will enable you to answer your question[1]. The **scale** of a variable describes the degree to which the variable exists along a continuum, such as from *low* to *high*. Temperature exists on an underlying continuum from low (cold) to high (hot). In contrast, there is no underlying continuum for nationality. Although nations can be rated on several variables that exist along a continuum (e.g., per capita income), the nations themselves do not exist along a continuum. I begin below with a scale that carries the most information about the underlying continuum and end with a scale that has no underlying continuum.

1. *Ratio scale.* For a variable to be on a ratio scale, *equal ratios must have equal meaning.* A ratio is a proportion, such as "half" or "twice." For example, money is on a ratio scale because the ratio of $50 to $100 (1/2) has the same meaning as the ratio of $100 to $200 (1/2). In addition, it makes sense to talk about one item costing *half* as much as another item or one job paying *twice* as much as another. In contrast, it doesn't make sense to talk about one person being "twice as happy" as another person, so happiness is not on a ratio scale. For a variable to be on a ratio scale, it must be possible for there to be a 0 value. It is possible to have $0 and, in theory, it is possible for there to be an absolute zero temperature (the complete absence of molecular motion), but it is unclear whether it is possible to have zero happiness.

2. *Interval scale.* For variables on an interval scale, we assume that *equal intervals have equal meaning.* For example, the difference between an I.Q. score of 120 and a score of 110 (a 10-point interval) represents the same performance difference as the difference between I.Q. scores of 80 and 70. Because there can be no 0 value for I.Q. and because it makes little sense to describe one person as "twice as intelligent" as another person, I.Q. is not on a ratio scale.

3. *Ordinal scale.* Variables on an ordinal scale contain information about sequence (order) such as first, second, third, etc. If you are observing a race and you record only who finished first, second, and third (and not the exact finish time), you are using an ordinal scale. If you ask participants to rank-order (1st, 2nd, 3rd, etc.) a set of films according to how much they enjoyed them, the data are ordinal scale.

4. *Nominal scale*. A nominal scale variable carries values that do not exist on an underlying continuum, such as religious affiliation (Catholic, agnostic, atheist) or position on capital punishment (for, against). Note that for the latter variable, the researcher has the option of representing opinion either as a two-category nominal scale variable or as an interval scale from "for" to "against." While nominal scale *independent* variables are common and easy to analyze, nominal scale *dependent* variables can be more challenging. For student projects, I would recommend selecting dependent variables that are on an interval or ratio scale.

1 Although many researchers believe that the scale of measurement dictates which statistical test is most appropriate, this is not true. The appropriateness of a particular test is dependent only on the mathematical assumptions of the test (e.g., that data are normally distributed), and not on scale (see Gaito, 1980).

Reliability

Reliability is the degree to which a measure is free from measurement error. **Measurement error** refers to the "noise" or random distortions in your observations that are caused by factors such as imperfections in your measurement instrument, errors in procedure, lack of participant motivation, or fluctuating environmental conditions.

Distinguishing reliability from construct validity. As discussed in a previous reading, **construct validity** is your confidence that the operational definition of a variable accurately reflects the underlying construct of interest and not some other construct. An example of an operational definition with high construct validity is the use of a thermometer to measure temperature. Using a thermometer correctly, you can be reasonably confident that the numbers generated by your instrument accurately reflect temperature. Whereas construct validity tells you how well you have isolated your variable from other variables, **reliability** tells you how well your measure is free from measurement error. An analogy I have found useful is with tuning a radio: Reliability is the degree of static in the signal you are receiving, while construct validity is the degree to which you have tuned to the station you want.

According to **Classical Test Theory**, every time you measure a variable (e.g., take someone's temperature), you obtain a combination of "true score" and measurement error. Unlike validity, which is not expressed as a number, reliability is a statistic. It is the proportion of a measure that is error-free. It varies from zero (no true score, all measurement error) to one (all true score, no measurement error), although in practice reliability is always less than one. Methods for estimating reliability were developed by asking, "If a measure were reliable – that is, if it had very little measurement error – what would we expect it to do?" Generally, the answer to this question is that we would expect a reliable measure to be *consistent*. It should be consistent across time (repeated observations should produce similar values), consistent across instruments (multiple measurement devices should produce similar values), and consistent across judges (independent judges of the same behavior should report similar values). Below, I discuss methods for estimating reliability that draw on these three assumptions.

1. *Consistency over time: Test-retest reliability*. Test-retest reliability is an estimate of the consistency of a measure across time. To use test-retest reliability, you must be able to assume that your variable is relatively stable over time. **Test-retest reliability** is the correlation (*r*) between the same measure collected at two time points. For example, the test-retest reliability of a measure of extroversion would be computed by giving a group of people the measure of extroversion, waiting for some period of time (e.g., two weeks, six months), giving the same group the same test again, and computing the correlation between those two tests. As a researcher, you would hope for the correlation to be as high as possible, generally above +0.7. One potential problem with using the same test twice is that people may remember their earlier responses and answer the same way to appear consistent. To avoid this problem, you could develop two slightly different versions of your measure and use the different versions at the two time points.

2. *Consistency over measures: Internal consistency reliability.* **Internal consistency reliability** is an estimate of the consistency across multiple measures of the same construct. Imagine a group of researchers conducting an experiment in which the precise measurement of temperature is essential. Rather than rely on a single thermometer, which could be flawed, they are likely to use several thermometers. If the thermometers all report the same temperature, the researchers can be more confident in their measurement than if the thermometer readings differed from one another. The most common usage of internal consistency reliability in psychology is for questionnaires, where multiple items are used to assess a single construct (e.g., optimism). For example, a researcher might develop 10 statements related to optimism and ask participants how much they agree or disagree with each one. The researcher hopes that all 10 statements are related to optimism, but it is possible that some items do not fit with the others. For example, the item "I see the glass as half full" would only be useful if the respondent were familiar with the idea of optimists seeing a glass as half full and pessimists seeing it as half empty. The most common way of measuring the consistency across multiple items is to correlate all possible pairs of items. If two measures are conceptually related, then they should be correlated with one another. For example, the two items "I tend to think that things are getting better and better" and "The future

looks bright to me" should be positively correlated. If participants strongly agree with one, they are likely to strongly agree with the other. If they strongly disagree with one, they are likely to strongly disagree with the other.

a. The statistic used to measure internal consistency reliability is **Cronbach's alpha**, often expressed as the Greek letter α, and is the average correlation among all pairs of items, adjusted for the number of items. A large number of items will increase alpha. Like the other measures of reliability, it varies from 0 to 1. As a researcher, you want α to be as high as possible, preferably above 0.7. An α below 0.7 suggests that at least some of the items in your measure do not "fit" with the others.

b. *Reversed items.* The two optimism items described above are phrased so that agreement with one will tend to co-occur with agreement with the other. It is a good idea to phrase some items in a questionnaire so that agreement with one will co-occur with *disagreement* with the other. For example, in measuring optimism, it would be a good idea to include an item endorsing pessimism ("I typically expect that things will not work out."). An item phrased so that responses should be the opposite of other items is called a **reversed item** and is useful for detecting the **acquiescence bias** (also known as the **yea-saying bias**): the tendency for some respondents to unthinkingly agree to items because of fatigue, poor motivation, or ambiguities in the item phrasing. If a respondent strongly agrees both to the items phrased to endorse optimism and also the items phrased to endorse pessimism, you should question the usefulness of their responses. If many respondents show this pattern, you should question the construct validity of your measure.

c. *Multiple measures*. In general, multiple measures of a behavior are preferred to single measures. This is because any single measure could be biased, poorly constructed, or administered improperly. Using many measures helps reduce the influence of any one flaw. In addition, some concepts may be too broad to be captured by a single measure. All things being equal, internal consistency reliability increases as the number of measures increases.

3. *Consistency over judges: Inter-rater reliability*. Inter-rater reliability is a measure of the consistency, or agreement, among multiple judges who are evaluating the same behavior. Some behaviors, such as playground aggression, can be ambiguous, leading different judges to record different values of aggression. If the group of judges all report similar levels of aggression for a particular child, then the measure of aggression has less measurement error than if the group of judges reported very different levels of aggression. Inter-rater reliability is measured in a variety of ways depending on the scale of measurement:

a. *Percent agreement*. If you have two judges who are using a nominal-scale dependent variable (e.g., helpful, neutral, hurtful), an estimate of inter-rater reliability can be obtained by counting the number of times that two judges agree (record the same behavior) and dividing it by the number of possible agreements. A problem with percent agreement is that it can be biased upward (seem better than it really is) if there are a very large number of possible agreements or only a few levels of the nominal-scale variable. Cohen's Kappa was developed to address this problem.

b. *Cohen's Kappa*. Like percent agreement, this is a measure of inter-rater reliability for two judges who are using a nominal-scale dependent variable. However, Cohen's Kappa is a more conservative measure of reliability because it corrects for the number of agreements that would be expected by chance.

c. *Correlation*. If judges are providing ratings (e.g., on a 1-6 scale), a measure of their agreement is the correlation between their ratings of the same stimuli. For example, if two judges provided ratings of the same 50 ice skaters, you could compute the correlation between those ratings as a measure of their reliability.

d. *Cronbach's alpha*. Correlation is effective when there are only two judges providing ratings, but for ratings from three or more judges, Cronbach's alpha can be used to estimate the average correlation between all possible pairs of judges, adjusting for the number of observations.

Error vs. Bias

Error is random noise that contaminates all measurements. Every measure contains a degree of error, and reliability is the main way of estimating how much error exists in a measure. In contrast, **bias** is a *consistent* raising or lowering of scores away from a person's "true score". For example, Steele and Aronson (1995, Experiment 4) found that Black students who reported their race *before* they answered questions from the GRE scored significantly lower than Black students who reported their race *after* they answered questions. For White students, reporting race had no significant effect. This suggests that a testing procedure that has students report their race prior to taking the test is *biased* against Black students.

In general, it is more common for differences among your subjects or variations in how you conduct your study to contribute *error* to your measurements than to contribute *bias*. If you use the word *bias*, you should explain how scores are being consistently raised or lowered.

Reactivity

Reactivity is the degree to which observations are influenced by the measurement itself. It is not possible to measure something without influencing it in some way, thus there is no such thing as a purely objective measure. However, measures vary in the degree to which they influence the object of measurement. A **reactive** measure is one that strongly influences its object, whereas a **non-reactive** measure provides a more objective assessment. For example, a measure of racism that explicitly asks respondents whether they hate members of a particular race may be reactive because respondents may become self-conscious about their racial attitudes and attempt to conceal negative aspects of themselves. The reactivity of a measure is influenced by the following factors:

1. **Social desirability**. If participants suspect that some responses are more socially desirable (likely to meet with praise or admiration rather than criticism) than others, they may respond in a way that casts them in a good light. This may be intentional or unintentional.

2. **Demand characteristics** are details of the study that betray the hypothesis (Martin Orne, 1959). If participants learn the researchers' expectations for the study, they may change their behavior either to conform to those expectations or to frustrate them. For example, if participants in Stanley Milgram's experiment on obedience discovered that the study was not on the effects of punishment on learning (which is what they were told) but rather on the effects of authority on obedience, they may have shown much greater resistance to the authority figure's commands.

3. **Self-awareness** is the degree to which participants are thinking about how they appear to others. Generally, increasing a person's self-awareness causes them to behave in a more socially desirable way. Self-awareness increases when people believe they are being *recorded*, for example on videotape or with an electroencephalogram (EEG).

4. **The Hawthorne effect**. The Hawthorne effect is the effect of awareness of being in a study, especially due to novel or favorable treatment. It is named after a series of studies conducted at the Hawthorne works of the Western Electric Company plant in Chicago in the 1920s and 30s. In these studies, workers were selected to perform under varying conditions (e.g., different levels of illumination). Over the years, the story of these studies has been simplified to the finding that regardless of the manipulation (increasing illumination, decreasing illumination), worker productivity steadily increased, with the usual explanation that the workers in the experimental conditions felt special and worked harder than the workers on the assembly line. However, Olson, Hogan, and Santos (2006) point out that the Hawthorne studies actually had a large number of complex findings and it is possible that no "Hawthorne effect" actually occurred during the Hawthorne studies. Regardless, the "Hawthorne effect"

has come to symbolize a type of reactivity effect due to awareness of being in a study. To reduce Hawthorne effects, researchers should consider how participants' knowledge that they are in a study is influencing their behavior. In particular, researchers should ensure that *both* experimental *and* control groups feel that they are in a study.

5. **Observer effects**. In a study using naturalistic observation, participants are typically unaware that they are being observed. **Observer effects** occur if the observer is discovered and participants change their behavior because of increased self-awareness or because of social desirability concerns.

6. **Direct effect of instrument on dependent variable**. In some cases, the instrument itself may directly influence the dependent variable it is designed to measure. Consider the spinal tap procedure, in which a needle is inserted between a person's vertebrae and spinal fluid is extracted. Even if spinal fluid carried information about anxiety, researchers would be reluctant to use a spinal tap measure of anxiety because the procedure itself would likely cause a dramatic increase in anxiety levels.

Ways to reduce reactivity. One way to reduce (but perhaps not eliminate) participants' concerns about social desirability is to make their responses **anonymous** – to ensure that their responses are never associated with their identities and to inform participants that this is the case. In some cases, such as videotaped interviews, participant responses are necessarily connected to identity. To reduce social desirability concerns in those cases, researchers should maximize **confidentiality** – the restrictions placed on the circulation of participant responses. If participants trust the researcher not to disclose their answers to others, they may be more forthcoming and honest. Perhaps the best way to reduce reactivity is to employ **nonreactive** or **"unobtrusive"** measures: behavioral measures collected in a way that does not alert participants. For example, Levine and Norenzayan (1999) used the accuracy of public clocks as a measure of the importance of time in a culture.

Sensitivity

The **sensitivity** of a measure is the degree to which the measure is capable of detecting subtle differences in a behavior. For example, thermometers vary in the range of temperature they are capable of detecting. Some thermometers might be sensitive to the typical range of outdoor temperature, whereas others might be designed for use in cooking, where the temperatures are much higher. An outdoor temperature thermometer would be incapable of telling you the difference between 350 and 400 degrees Fahrenheit – it would be *insensitive* to this difference.

Sensitivity influences two aspects of design in psychological research. First, you should design your measure so that it targets the average level of behavior you will observe. For example, one group of student researchers was interested in the influence of siblings on sexual behavior and designed a questionnaire to measure sexual behavior in college students. They asked participants to indicate whether they had ever engaged in kissing, under-the-clothes fondling, oral sex, or intercourse. The problem with this measure is that 100% of their sample responded yes to kissing, 94% responded yes to heavy petting, 80-90% responded yes to oral sex, and 70-80% responded yes to intercourse. In the case of the kissing measure, it would be impossible for siblings to have any influence because every participant reported kissing. The other measures suffer from similar problems: the possible effect of siblings is limited by the high base rate. A more sensitive measure of kissing would have been the *number of different people* the participant has kissed, or the average *frequency* of kissing, rather than whether or not the participant has kissed anyone.

Failing to accurately predict the mean response can lead to ceiling or floor effects. A **ceiling effect** occurs when scores pile up at the high end – when a large proportion of participants receive the maximum score. The high base rates in the sexual behavior study are examples of ceiling effects. A **floor effect** is the opposite – when scores pile up on the low end. For example, if the researchers had asked participants whether they had ever engaged in sado-masochism, they would likely have received very few positive responses. Both cases indicate a **restriction of range** in your dependent variable: variability that is limited to a narrow range of scores. Restriction of range reduces your ability to detect an effect (your "statistical power"). You would need a very large number of participants to find an effect when the average response is very high or very low.

The second implication of sensitivity is that you should seek the opposite of restriction of range; you should create your measure to generate the largest possible **variance** in responses (different responses from different people) while still preserving some reliability. For example, you could measure political conservatism using a nominal-scale measure, asking people to identify themselves as either liberal or conservative. This would produce very little variance because there are only two possible responses. In contrast, a 10-point scale from Liberal to Conservative would generate greater variance. All other things being equal, a measure with greater variance is more sensitive. However, it would be a mistake to ask people to report their conservatism using a million-point scale. This would lack reliability because participants could not meaningfully distinguish between a score of 980,212 and 980,213 and they would be unlikely to give the same answer twice (producing low test-retest reliability).

Finding a Published Measure

Before designing your own questionnaire, it is worth consulting the scientific literature or your instructor to see if someone has gone before you. Questionnaires designed by other researchers often have the advantage of known reliability and have sometimes been evaluated with regard to their criterion, convergent, or discriminant validities. Researchers who develop new questionnaires in their research sometimes publish the questionnaire items in their articles, either as tables or appendices. If you are unable to find their questionnaire in their articles, you may contact the author and request the use of their measure using the author contact information given in one of their recent articles. Because researchers sometimes move from one institution to another, it's a good idea to verify the contact information by checking it against the contact information provided on their institution's website.

A final source of information about published questionnaires is a reference work that catalogs psychological measures. The *Mental Measurements Yearbook*, published every 18 to 24 months by the Buros Institute, gives reliability and validity information for a wide range of available measures. Another resource is *Test Critiques*, published by Pro-Ed and containing a review of over 800 measures. Both are typically carried by college libraries.

References

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564-567.

Levine, R., & Norenzayan, A. (1999). The pace of life in 31 countries. *Journal of Cross-Cultural Psychology*, *30*(2), 178-205. doi:10.1177/0022022199030002003

Olson, R., Hogan, L., & Santos, L. (2006). Illuminating the history of psychology: Tips for teaching students about the Hawthorne studies. *Psychology Learning & Teaching*, *5*(2),110-118.

Orne, M. (1959, May). The nature of hypnosis: Artifact and essence. *The Journal of Abnormal and Social Psychology, 58*(3), 277-299.