3

# Measurement Fundamentals for RETs

*If you cannot measure it, you cannot improve it*

- LORD KELVIN

## INTRODUCTION

Measurement is at the heart of scientific research and is central to RETs. Chapters 1 and 2 described the theoretical and conceptual foundations of RET design. In this chapter, I discuss measurement foundations for RETs for purposes of program evaluation.

Poor measurement in RETs can lead us to underestimate or overestimate the role of mediators in determining outcomes and it can lead to erroneous conclusions about the effects of a program on mediators and outcomes (Gonzalez & MacKinnon, 2020; Cole & Preacher, 2014). In this chapter, I first address the process of concept-measurement mapping. This activity ensures that the measures used capture the targeted concepts. I then discuss measurement metrics with a focus on Stevens's (1951) distinctions between nominal, ordinal, interval and ratio level properties. These distinctions are important because they impact how one analyzes RET data. Next, I discuss measurement error, including both random and systematic measurement error. This is followed by a discussion of the facets of measurement, which can affect the measurement choices you make. To set the stage for my use of SEM in future chapters, I describe latent variable representations of measurement. These methods can be used to help adjust for bias introduced by measurement error in RET data. Finally, I discuss issues related to formative measurement and the use of composites, measurement invariance, construction of study specific measures, the practice of dichotomizing measures of continuous constructs, whether to make baseline assessments of outcomes and mediators, and decision making surrounding the frequency and timing of assessments. The material in this chapter is important because without strong measurement, conclusions made from RETs are suspect. Strong measurement is necessary for strong science.

## CONCEPT-MEASUREMENT MAPPING

The measures one uses to represent constructs in an RET should, of course, be guided by the conceptual definition of each construct. For target variables in an RET, it is important that you adequately specify the universe of relevant content for each variable and then decide what to sample from that universe for measurement purposes. For example, if a conceptualization of social support specifies four types of support (tangible, emotional, informational, companionship), then these four domains need to be represented in a measure unless you are interested in only a specific type of support. The extent to which a measure adequately represents the totality of a construct is called **content validity** (Nunnally, 1978). As Cronbach, Gleser, Nanda, and Rajaratnam (1972, p. 380) state, "if one claims that observations represent a universe, one ought to define the universe clearly. Readers should agree that the conditions appearing in the study fit within the universe and that they are reasonably distributed over its whole range, not confined to some narrow sub-universe."

Inevitably. a healthy "give and take" exists between conceptual definitions and measurement, with conceptual definitions guiding the choice of measures for an RET and the choice of measures nuancing more abstract conceptual definitions. Consider the case of poverty reduction programs in which poverty is defined conceptually as the state of being very poor. How might this state be measured? A researcher might use income as an index, with income values below a cutoff standard (sometimes called the "poverty line") reflecting poverty. Another researcher might agree with this approach but further reason that people may have limited income yet own considerable assets that enable them to live quite comfortably. As such, this researcher might factor assets into a measure of poverty as well. How should he or she weight the different assets, such as home ownership, car ownership, and access to health care? In agricultural communities in developing countries, those who own horses, cows, or a truck often are said to possess key assets. Are these assets taken into account when assessing poverty in studies of people living in an agricultural community? What about urban areas? As we examine more closely what assets the researcher measures and how they are quantified and combined, we gain a more nuanced sense of the concept of poverty invoked in the study.

In the United States, the poverty line is set by the Census Bureau and is defined as the minimum amount of income needed to cover basic needs (in 2018, it was about $25,000 for a family of four). However, countries define poverty using different criteria. For example, many countries currently define poverty based on the concept of how much it costs people to have enough to eat. Having enough to eat is often quantified as 2,000 calories per day, but this value might be adjusted upward or downward depending on circumstance. Policymakers then calculate the smallest amount of money necessary to buy a food bundle with 2,000 calories. Other countries base the index on the cost of purchasing essential living necessities that include factors other than just having enough to eat. What are these necessities? These also vary from country to country. In most European countries, the poverty line is set at 60% of the median income of the country. Using yet another approach, in research in small rural towns, one might measure poverty by asking residents to identify the poorest individuals in the township. These subjective nominations constitute yet another measure of poverty.

In all of these cases, the conceptual meaning of poverty is "the state of being very poor," but the meaning of "very poor" is somewhat different and nuanced. In one case (nominations by others) it is subjective, while in other cases it is based on different criteria with different cutoffs. Measurement and conceptual definitions go back and forth, with measurement often giving greater clarity to the meaning of a concept in an RET but with the essential meaning of the concept driving the choice of measures. When designing RETs, one must pay careful attention to the mapping of concepts onto measures and the mapping of measures onto constructs.

As another example, consider depression, a psychological state that is assessed in more than 1,000 empirical studies each year. There exist over 275 scales for depression with little content overlap between them, representing more than 50 different symptoms across inventories (Fried & Flake, 2018). One third of the symptoms in the widely used Center of Epidemiological Studies Depression (CES-D) scale do not appear in the most commonly used depression scales. Given this, it is not surprising that results for clinical trials with depression often lead to different conclusions even when different measures are used in the same study. The field could benefit from better concept mapping.

## MEASUREMENT METRICS

The term metric refers to the numbers that an observed measure can take on when describing individuals' standings on the construct of interest. For the construct of self-esteem, for instance, the metric of a measure might range from the lowest possible rating of 1 to the highest possible rating of 7. The metric for grade point average in the United States in many schools is 0 to 4 (4 = A, 3 = B, 2 = C, 1 = D, 0 = E/F).

Stevens's (1951) proposed a metric classification system for measures that has been influential in the social sciences. **Nominal measurement** uses numbers merely as labels. For example, an investigator might classify a group of people according to their religion—Catholic, Protestant, Jewish, and all others—and use the numbers 1, 2, 3, and 4 for these categories. The numbers have no special quality; they are merely labels. **Ordinal measurement** assigns numbers to objects in ways that allow the objects to be ordered on an underlying continuum. For example, if on the Scholastic Aptitude Test (SAT) that measures verbal ability you are told you scored in the 50th percentile and someone else scored in the 48th percentile, you know you have more verbal ability than that person, but you have no idea by how much. When cast in a metric of percentiles, the SAT represents an ordinal measure. **Interval-level measurement** assigns numbers in ways that have ordinal properties, but with an additional mathematical property: the difference between two consecutive numbers will reflect the same amount of the underlying dimension as any other two consecutive numbers. For example, on a Fahrenheit scale of temperature, the difference between 32 and 33 degrees reflects the same amount of heat as the difference between 45 and 46 degrees. **Ratio-level measurement** has all the properties of interval-level measurement, but in addition, ratio statements are meaningful. For example, for distance in miles, 4 miles is twice as far as 2 miles and 20 miles is twice as far as 10 miles.

Stevens's system is important because the type of analytic method one uses for RET data depends on the type of metric properties it has. One decision that analysts must confront is whether a set of observations should be analyzed using methods that assume ordinal-level but not interval-level properties, or interval level properties but not ratio level properties, or ratio level properties. Most students learn that scales are either ordinal, interval, or ratio and that they are to adapt analytic approaches accordingly. Such dichotomous thinking is misleading. It is similar to saying someone has a fever because their body temperature is one tenth of one degree above the standard of 98.6°F (37°C). The person does indeed have a fever, but it is small and inconsequential. The same is true of metric properties like ordinality; a measure can approximate intervalness, but technically not be interval. In such cases, the measure might be analyzed using methods that assume interval properties, as long as the approximation is reasonably close.

Ordinal, interval, and ratio properties are characteristics of measures (observations), not just the scales used to generate those measures. A measure has as its referent not only a particular scale (e.g., the Beck depression inventory), but also an individual on whom the measure is taken, a time at which the measure is taken, and a setting in which the measure is taken. All of these factors influence the properties of a set of measures. I can illustrate this idea using height as an example. Suppose the height of five individuals is measured on two different metrics, inches (I call this measure X) and a second measure that is nothing more than the rank order of the individual for the set of individuals being studied (I call this measure Z). Most would agree that because it is a rank order metric, Z constitutes an ordinal scale, while the metric of X has interval (indeed ratio) level properties. Here are the scores of five individuals on each measure:

| Individual | Height (in inches) | Ordinal Measure |
|---|---|---|
| A | 72 | 5 |
| B | 71 | 4 |
| C | 70 | 3 |
| D | 69 | 2 |
| E | 67 | 1 |

Focus on X, the measure that uses inches. For this measure, a difference of 1 between two scores always represents the same physical difference on the underlying dimension of height. For example, the actual height difference between individuals A and B (72-71=1) corresponds to the same true underlying height difference as that between individuals C and D (70-69=1). Similarly, the metric difference between individuals D and E (69-67=2) and the difference between individuals A and C (72-70=2) are the same on the observed metric, and this also reflects the same amount of height differential on the underlying dimension. This is the essence of interval level measures.

Note, however, that these properties do not hold for the ordinal measure. The difference in scores between individuals A and B on the ordinal metric is 1 (i.e., 5-4) and the difference in scores for individuals D and E also is 1 (i.e., 2-1). These identical observed score differences correspond to differing degrees of height disparities on the underlying dimension of height. The true difference between individuals D and E is larger than the true difference between individuals A and B, as can be seen by referring to the measure using inches for these individuals. The measure has ordinal properties.

Now suppose I am studying five different individuals with the following scores:

| Individual | Height (in inches) | Ordinal Measure |
|---|---|---|
| A | 72 | 5 |
| B | 71 | 4 |
| C | 70 | 3 |
| D | 69 | 2 |
| E | 68 | 1 |

Note that *for these five individuals*, the ordinal measure now has interval-level properties. The difference in scores between individuals A and B is 1, as is the difference between individuals D and E. These differences correspond to the same amount on the underlying physical dimension. *For these individuals*, the ordinal measure is interval level.

Suppose individual E in the above example was not 68″ tall, but instead was 67.9″ tall. In this case, the "ordinal" measure Z is no longer strictly interval. However, it is very close and probably can be treated as if it is interval without adverse effects. In this sense, for these five individuals, the approximation to intervalness is so close that the measure can be treated as if it is interval without consequence. It is like having such a small fever that it basically does not matter. The measure, functionally, has interval level properties.

This example illustrates that the critical issue when estimating relationships between constructs is not whether the measures that represent those constructs are interval or ordinal. Rather, the issue is the extent to which each measure approximates interval-level properties. If the approximation is close, then the data can be effectively analyzed using statistical methods that presume interval-level properties, such as Pearson correlations or traditional regression. If the approximation is poor, alternative analytic strategies might be called for, such as ordinal modeling. For discussion of how to create measures that likely approximate interval level properties, see Jaccard and Jacoby (2020).

A problem with many ordinal analytic methods is that they make assumptions about variables that, when violated, can produce parameter estimates that are even more misleading than what one would find by treating the measures as being "interval enough" (Taylor, West & Aiken, 2006). Simulation studies suggest that measures can depart from intervalness rather substantially and still yield valid inferences about the underlying construct, depending on the question being asked and the nature of departures from intervalness (Davidson & Sharma, 1988, 1990, 1994). Finally, many ordinal methods require sample sizes that are larger than those for standard regression methods, particularly when the number of categories in the outcome variable is small or the distribution of the coarsely categorized outcome variable is skewed (Taylor, West & Aiken, 2006). Given the above, it sometimes is better to analyze ordinal data as if it were interval as long as you can build a case for a reasonable approximation to intervalness.

Parenthetically, some researchers conflate the number of categories a measure has with whether the measure has ordinal or interval properties. The assumption is that quantitative measures with few categories are, by definition, ordinal. The number of categories of a scale refers to the **precision** of that scale, with more categories being more precise. This property is not isomorphic with whether the numerical representations are equally spaced along the underlying dimension being measured. When we represent continuous variables using a measure that has only four or five values, the conceptual "spacing" of the values for that coarse measure can still approximate interval properties in the sense that they are equally spaced across the underlying continuum. Granted, the more fine-grained variations within a numerical category are glossed over because the within-category values are treated as equivalent. However, as long as the metrics are not too coarse and the categories are roughly equally spaced on the underlying continuum, we may not get into trouble using them as if they have interval level properties. Simulation studies that have addressed issues of scale coarseness with continuous constructs suggest that 5 to 7 categories are enough for many applications (Bollen & Barb, 1981; Green, Akey, Fleming, Hershberger & Marquis, 1997; Lozano, García-Cueto & Muñiz, 2008; Lubke & Muthén, 2004; Taylor, West & Aiken, 2006; Maydeu-Olivares, Fairchild & Hall, 2017). The idea that precision is isomorphic with ordinality also can be discredited using a simple counterexample: A measure of the number of children in a family usually has few values (in most research, the values are between 0 and 4), yet it has ratio level properties.

A related problem involving the confusion of precision and interval properties is when researchers sum multiple items into a total score and then treat the total score as being interval in nature despite the fact that each individual item is rated on a single, few-category "Likert type" scale that has blatantly ordinal adverb descriptors. It is not the case that summing multiple ordinal items magically results in an interval level total score. To be sure, it might do so, but then again, it also may fail to do so. I try to use multi-item composites where a reasonable case can be made that the individual items of the measure have roughly interval level properties.

A final mistake to be careful of is equating the properties of a metric when it is used to measure one construct as having those same properties when the metric is used to measure another construct. For example, the metric of seconds has ratio level properties when used to reflect the amount of time that has passed since an event has occurred. However, the measured time it takes to respond to a question posed on a computer screen as a measure of the psychological construct of decision uncertainty does not necessarily have ratio level properties as a measure of uncertainty: Someone who takes 4 seconds to respond to the question is not necessarily twice as uncertain as someone who takes 2 seconds to respond to the question. Always ask if it is reasonable to make ratio-like statements for the underlying construct you are assessing based on its metric.

A final common error is to assume that any single-item rating scale with four to seven adverb qualifiers constitutes a "Likert" scale and that, by definition, it is ordinal and must be analyzed as such.[1] To better approximate interval level properties for rating scales, psychometricians recommend using adverbs that connote roughly equal intervals across the underlying dimension. For example, when rating the importance of each of several factors that entered into a decision, a commonly used format is:

___ Very important
___ Somewhat important
___ Not very important
___ Not at all important

Note that there seems to be unequal "psychological spacing" between these qualifiers. The difference between "not very important" and "somewhat important" seems slight compared to the difference between "somewhat important" and "very important." The choice of these adverbs creates an ordinal metric and likely should be analyzed as such. Having said that, it is possible to choose adverbs that connote roughly equal spacing and, given this, data using such adverbs might be amenable to interval level modeling.

There are large literatures in psychometrics that can guide a researcher's choice of adverbs (Beckstead, 2014; Rohrman, 2015). Using psychophysical scaling methods, Cliff (1959) found that describing something as "slightly good" is generally perceived to be about 0.50 times as "good" than the simple, unmodified "good." He scaled a large number of adverbs with the idea that the derived scale values can assist adverb selection. Rohrman (2015) presents adverb analyses in English, German, and Chinese for dimensions of frequency (e.g., never, seldom, sometimes, often, always), intensity (not at all, a-little, moderately, quite-a-bit, very much), probability (certainly-not, unlikely, about-50:50, likely, for-sure), quality (bad, inadequate, fair, good, excellent) and agreement (fully-disagree, mainly disagree, neutral, mainly-agree, fully-agree). Beckstead (2014) summarizes research on qualifying values for frequency judgments and magnitudes. I do not recommend assuming that the qualifying values in these reports necessarily apply in your research because studies show that qualifying values can vary as a function of different measurement facets (McClelland, 1975). However, the studies can serve as rough guidelines if you need to construct your own rating scales for your RET. Coupled with common sense that is sensitive to creating equal psychological

[1] Likert developed multi-item scaling models that used rating scales for the items comprising the scale. I have no doubt he would turn over in his grave if he knew that most every form of single item rating scale is attributed to him.

differences between categories as well as proper cognitive response testing (see below), reasonable adverb choices can be made to produce approximately interval-level data (see Jaccard & Jacoby, 2020). When I evaluate existing measures for possible use in my RETs, I am more confident in the interval-level properties of a measure if it uses adverb qualifiers for individual items that are equal interval appearing.

## MEASUREMENT ERROR

The measures we use in RETs often have measurement error. Such error can bias estimates of causal effects and undermine valid inference. In this section, I characterize the concepts of random and systematic measurement error and foreshadow methods I use in future chapters to adjust for the bias that measurement error creates.

### Random Measurement Error

One type of measurement error is **random measurement error**. Such errors are random events that arbitrarily push an individual's observed score up or down relative to the person's true score. For example, a person might misread an item or the item might be ambiguous leaving it open to varying interpretations. The reliability of a measure is the extent to which it is free of random error. Psychometricians quantify reliability using an index called the **reliability ratio**, which ranges from 0 to 1.00. It is the degree to which a measure is free of random error. If the reliability ratio (or more simply, the reliability) is 0.70, then 70% of the variation in the measure is systematic and 30% is random error. If the reliability ratio is 0.90, then 90% of the variation in the measure is systematic and 10% is random error.

It turns out that it is difficult to know the true reliability of many measures, but psychometricians have developed (imperfect) methods for estimating it. These estimation strategies include test-retest methods, split half methods, alternate form methods, methods based on SEM, and coefficient alpha for multi-item scales, among others. Consideration of these methods is beyond the scope of this book; interested readers should read Price (2016) and Furr (2017). Some scientists confuse reliability estimation methods with the concept of reliability per se. For example, some researchers describe reliability as the extent to which a measure at one point in time is highly correlated with or reproduces scores for that same measure at another point in time, assuming the construct remains stable. This is not reliability. It is the test-retest method for estimating reliability. Different estimation methods can be sensitive to different types of random error and make different assumptions, but ultimately, our goal when estimating reliability of a measure is to estimate the extent to which a measure is free of random error.

As noted in Chapter 1, measurement unreliability can distort the inferences we make. If we want to determine the degree of association between two variables, X and Y, and if both measures have a reliability of 0.60 (i.e., 40% of the variability in each measure is random noise), then we are likely to underestimate the true association between the constructs because the measures each contain substantial amounts of randomness. Correlating randomness with randomness results in correlations near zero. When we conduct an analysis that seeks to control for a confound by including a measure of it in a regression analysis, if the measure is contaminated by random error, then perhaps we really have not adequately controlled for it. One does not control well for SES, for example, if one's measure poorly reflects SES because 50% of its variation is random noise. Cole and Preacher (2014) offer a detailed analysis of the negative effects of measurement error in causal modeling, identifying five adverse consequences, (1) as measurement error increases, path coefficients will be under- or overestimated, (2) even small amounts of measurement error can cause valid models to appear invalid, (3) measurement error lowers statistical power, (4) differential measurement error in various parts of a model can change substantive model conclusions, and (5) all of these problems become more serious as models become more complex.

There are multiple procedural strategies one can use to reduce unreliability, including (a) creating an accommodating test environment that has optimal heating, lighting, freedom from noise, and does not have others present who might serve as a source of distraction, (b) minimizing respondents rushing through the assessment tasks by keeping assessment burdens low, (c) providing respondents with practice items so they accommodate to the task, the rating scales, and the testing environment, (d) ensuring instructions are clear (confusing directions lead to confused respondents), (e) ensuring items have no ambiguities (e.g., "Have used drugs in the past month" – the term "drugs" and "month" are both ambiguous), and (f) where possible and as appropriate, using multiple items to assess constructs so that when aggregated across items, random errors cancel. I discuss SEM based statistical methods for addressing measurement error below.

## Systematic Measurement Error

A second type of measurement error is **systematic measurement error**. Systematic measurement error is not random. As one example, constant error occurs if measured scores are biased upward (or downward) by a constant value for everyone, such as when a scale to measure weight is consistently 5 pounds too heavy. Another type of systematic error, socially desirable response tendencies, refers to an individual difference variable that reflects a propensity to want to create good impressions on others. Individuals who are high on this trait are more likely to systematically underreport such things as drug use, unprotected sex, and depressive symptoms and to over report income, life satisfaction, and accomplishments. Systematic error can bias estimates of means, correlations, and other statistics when we conduct theory tests. For example, social desirability response tendencies might impact self-reports of both drug and alcohol use, thereby inflating the estimated correlation between the two variables.

A common strategy for dealing with systematic error variance is to obtain a measure of it and then include the measure as a covariate in one's statistical model. For example, measures of social desirability response tendencies (Stoeber, 2001; NieBen et al., 2019) can be obtained in a study and then used as a covariate in the prediction equation. Baumgartner and Steenkamp (2001) measured five response styles that represent systematic error: (1) acquiescence (the tendency to endorse items/questions, independent of their content), (2) disacquiescence (the tendency to disagree with items/questions independent of their content), (3) extreme (the tendency to use the extremes of a rating scale independent of item/question content), (4) midpoint (the tendency to use the midpoint of a rating scale independent of item/question content), and (5) noncontingent (the tendency to respond to items carelessly, randomly, or nonpurposefully). These researchers found that the response styles accounted for an average of about 25% of the variation in 14 different consumer behavior constructs. To be sure, there is debate about the extent to which response styles are problematic for social science research (Lance, Dawson, Birkelbach, & Hoffman, 2010; Podsakoff, MacKenzie, & Podsakoff, 2012; Spector & Brannick, 2009; Van Vaerenbergh & Thomas, 2013). However, I think it is good practice to consider the possibility that the response styles operate and then introduce correctives, should evidence for response set bias be present. I discuss this point below when I address conducting measurement-oriented pilot studies.

Sometimes systematic error can be addressed using procedural rather than statistical controls. For social desirability, the following practices help minimize its impact: (1) use a self-administered as opposed to a face-to-face reporting format so that respondents do not have to report sensitive behaviors directly to an interviewer; (2) use anonymous/confidential conditions and provide respondents assurances that identifying information will not be associated with their data; (3) stress that results in research reports will only be reported for groups of people, not for individuals; (4) provide motivational instructions at the outset that encourage honest reporting (e.g., how the quality and ability of the research to "make a difference" depends on people being honest); and (5) instruct respondents not to answer a question if they are not going to be truthful (and then use specialized analytic methods to handle the missing data).

A measure is said to be valid or have **validity** to the extent that it is absent of both random error and systematic error, that is, the only thing causing variation in it is the true variance of the construct the measure is thought to reflect. We build a case for validity by empirically demonstrating that a measure is correlated with constructs it should be correlated with and not correlated with constructs it should not be correlated with. The latter property has been extended to a concept known as **discriminant validity**, which focuses on whether a measure of a construct shows correlations with measures of other distinct constructs that are low enough that the target measure can be regarded as indeed measuring a distinct construct (see Rönkkö & Cho, 2022, for a discussion of how to establish discriminant validity).

I routinely build into all of my RETs empirical tests that affirm the validity of my measures. For example, I might use two measures of depression to show that they are highly correlated with one another because if they are valid, they should be (a form of validity known as **convergent validity**). There are well established differences in depression between males and females. If my measure does not show such differences, it raises questions about its validity. I always look for opportunities to build a case for measure validity in my RET be it through discriminant validity, convergent validity, predictive validity, concurrent validity or some other psychometric standard .

## THE FACETS OF MEASUREMENT

Social scientists often speak of the reliability and validity of a scale. Like the ordinality or intervalness of metrics, this is not a technically correct frame of reference. As noted, when we administer a scale, we do so in a specific setting, for a specified population of individuals, and at a given point in time. The reliability and validity of a set of observations is a function of each of these facets of measurement, scale inclusive. A scale can be reliable and/or valid for one population but not another. It can be reliable and/or valid when administered in one type of setting but not another. A scale can have good psychometric properties at one point in time but not another point in time. In this sense, we should reference the reliability and validity of a set of observations rather than a scale.

There are many examples of the impact of measurement facets on the reliability and validity of measures. For example, there is controversy about whether intelligence tests are valid indicators of general intelligence for different ethnic groups (Onwuegbuzie & Daley, 2001). Studies also have explored whether web-based administration of a scale alters the reliability and validity of the observations produced by that scale relative to traditional face-to-face administrations (Meade et al., 2007). Research has found that the reliability and validity of a scale can change with the age of a child as children mature and acquire different levels of cognitive and emotional ability (Borgers et al., 2000). Research also has documented differences in the validity of self-reports of physical activity by adults as a function of their educational levels (Annemarie et al., 2015).

These examples underscore the importance of keeping in mind each of the facets of measurement if you choose an existing measure to represent a construct in your RET. Researchers often choose measures with the idea that their validity has already been established and therefore need not be addressed. For example, there are well developed measures of depression, financial literacy, brand loyalty, political ideology, alcohol use, and health related expectancies, all of which have documented favorable reliability and validity data. When deciding whether you can reasonably use one of these measures, you need to think about whether the facets of measurement used in the prior reliability and validity studies map onto the measurement facets of your study and whether any disparities in that mapping matter. If reliability and validity were established using college students, but your study uses economically disadvantaged, inner city youth of color in middle school, does the prior documented reliability and validity of the measure apply to your research? If not, you will need to ensure the reliability and validity of the measures you use independent of this past psychometric research.

## CRONBACH'S ALPHA AND ISSUES OF DIMENSIONALITY

A common index researchers use to assert reliability and/or unidimensionality of a multi-item measure is Cronbach's coefficient alpha. It turns out alpha is limited for this. Psychometricians have argued for years that coefficient alpha is outdated and that more modern indices should be used (Sijtsma, 2009; McNeish, 2018; but for counterarguments and broader perspectives, see Savalei & Reise, 2019 and Sijtsma, Elli & Borsboom, 2024). To illustrate, for purposes of documenting dimensionality, consider the following correlations for two six item scales, each with a coefficient alpha of 0.86:

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | - | | | | | | - | | | | | |
| 2 | .8 | - | | | | | .5 | - | | | | |
| 3 | .8 | .8 | - | | | | .5 | .5 | - | | | |
| 4 | .3 | .3 | .3 | - | | | .5 | .5 | .5 | - | | |
| 5 | .3 | .3 | .3 | .8 | - | | .5 | .5 | .5 | .5 | - | |
| 6 | .3 | .3 | .3 | .8 | .8 | - | .5 | .5 | .5 | .5 | .5 | - |

If one uses the alpha of 0.86 to assert unidimensionality, the assertion is wrong for the scale on the left. Alpha does not test for unidimensionality; rather it *assumes* it.

Respecting the dimensionality of measures can be important because of the possibility of **aggregation bias** when distinct dimensions are summed or averaged. For example, some theorists distinguish four dimensions of depression, (1) a cognitive component, (2) an affective component, (3) a somatic component, and (4) a behavioral component, all of which are correlated about 0.35 with each other. Suppose that an outcome Y is correlated 0.30 with the cognitive component but is zero correlated with the other three components. If I sum the scores on the components into an overall index, the meaningful correlation of Y with the cognitive component will be masked because I have contaminated it with the other components that are functionally operating as random noise relative to the prediction of Y. Even if a significant correlation between Y and the total score of depression manages to emerge, it is misleading because it implies changing any of the four components will impact Y when, in fact, it is only changes in the cognitive component that will do so. If the four constructs are highly correlated and form a unidimensional scale, then a composite of them is unlikely to result in aggregation bias.

Psychometricians distinguish between internal consistency and homogeneity. **Internal consistency** refers to the degree of interrelatedness of items whereas homogeneity refers to **unidimensionality** or the extent to which the covariance structure among items can be accounted for by a single latent factor. These constructs are distinct. If 10 items are all intercorrelated 0.20, the correlational pattern among them can be accounted for by a single factor, i.e., the items are unidimensional. However, their internal consistency is low that the items are only correlated 0.20. Although coefficient alpha is impacted by the internal consistency of items, it is a weak index of it, in part, because its magnitude is impacted by factors other than internal consistency, such as the number of items on the scale. For a 10-item scale with correlations between items of 0.20, the coefficient alpha is 0.71, yet any pair of items share only 4% common variance.

Tests of dimensionality are best pursued using confirmatory factor analysis. On the resources tab of my website, I present a primer that shows how to use Mplus to do so. In place of coefficient alpha, Kelley and Pomprasertmani (2016) recommend a reliability index called **omega hierarchical** for interval-level item response metrics and **omega categorical** for binary or ordinal item response metrics. These indices estimate the classic reliability ratio of a composite measure but allow for minor deviations from unidimensionality in the form of correlated errors. On the programs tab of my webpage, I provide a program for calculating these indices (the program is called *composite reliability*).[2] If you obtain a low index of composite reliability, then this is a signal that you may need to revisit your composite and consider the possibility of splitting it into

[2] A popular index of composite reliability called omega does not accommodate correlated errors. The term *omega hierarchical* has been used for indices other than that of Kelley and Pomprasertmani (2016), so be aware of this.

different measures because it likely is measuring different constructs. This also is true when tests of unidimensionality of the composite blatantly fail or when you lack reasonable levels of internal consistency. I return to this issue in more depth below.

Faced with psychometric results of one form or another that question the aggregation of items of a scale in your study, it is *not* necessarily the case that you should drop offending items to improve the psychometric properties of the aggregate. If the items were generated by scale developers to represent the content universe of the construct in question, then dropping items can undermine this property. If dropping an item undermines content coverage, then you might retain the offending item(s) but treat it as a separate construct from the aggregate. If you ultimately decide to drop items, then you should recognize that you now are focusing on only a portion of the construct you originally sought to study. Of course, it is possible the remaining items still adequately represent construct content, in which case dropping an item may not be consequential.

## ADDITIONAL CRITERIA FOR CHOOSING A MEASURE

Reliability and validity are not the only criteria used to evaluate measures. When discussing the characteristics of a good measure of economic well-being at a country level, Sumner (2004) states the measure should “be policy-relevant, a direct and unambiguous measure of progress, specific to the phenomena, valid, reliable, consistent, measurable, user friendly, not easily manipulated, cost effective and up-to-date.” Summer’s list includes reliability and validity but it also includes practical criteria.

Blanton and Jaccard (2006) argue that in addition to reliability and validity, good measures have non-arbitrary metrics, i.e., the values of the measure have intuitive meaning tied to meaningful benchmarks. Metric arbitrariness can exist independent of reliability and validity. Consider height of individuals as expressed in meters. For most residents of the United States, if told that a person is 1.83 meters tall, people will have no sense of how tall such an individual is. Yet, expressing height in meters uses a scale that is completely valid and reliable. If told that another individual is 1.52 meters tall, one knows the latter individual is shorter than the former individual; but by how much? Most U.S. citizens would have no idea. This is because, for them, the metric of meters is arbitrary. Many self-report measures have this arbitrary quality. Given a choice between a measure with a non-arbitrary metric (e.g., on how many days in the past 30 days have you smoked cigarettes) or an arbitrary metric (e.g., on a rating scale, indicate “how often have smoked cigarettes in the past 30 days,” with response options, “not at all,” “a few times,” “a moderate amount of the time,” “quite a bit”), my preference is for the former. For a discussion of making arbitrary metrics non-arbitrary, see Jaccard and Bo (2019).

Finally, another criterion for choosing a measure is its level of measurement, namely whether the measure has ordinal, interval or ratio level properties. One should prefer measures higher up on the Steven's measurement hierarchy, all else being equal.

## MEASUREMENT AND LATENT VARIABLES

There are many strategies one can use to correct for measurement error but one of the most popular approaches is a strategy based on latent variables coupled with structural equation modeling (SEM). In this framework, a latent variable is the true construct that you are interested in making statements about—for example, depression. The observed measure of that latent variable is the observable response (e.g., responses to a depression inventory) that you use to infer a person's standing on the latent construct. Figure 3.1 depicts a measurement model per SEM. The latent variable of depression is in a circle and the observed measure thought to reflect depression is in a rectangle. A causal path is drawn from the latent variable to the observed measure under the assumption that how depressed a person is influences his or her responses to the questions on the inventory. There also is an error term, ε, that reflects measurement error; that is, factors other than depression that influence a person's responses on the inventory.

The relationship between the construct and the indicator is usually assumed to be linear in accord with the equation

$$X = \alpha + \lambda \, LX + \varepsilon \qquad [3.1]$$

where X is the observed variable, LX is the latent variable, $\varepsilon$ is measurement error, $\alpha$ is the measurement intercept and $\lambda$ is the path coefficient linking LX to X. However, the function also can be nonlinear. Interval measures are characterized by a linear function whereas ordinal measures have monotonic, non-linear functions.

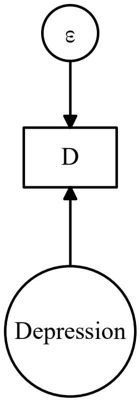


**FIGURE 3.1.** Measurement model

Sometimes we obtain multiple, interchangeable indicators of a construct. For example, a researcher might administer three measures of depression, the Center for Epidemiologic Studies Depression scale (CES-D), the Beck Depression Inventory (BDI), and the PHQ-9 to study participants. A measurement model for this case is in Figure 3.2. The latent variable of depression influences each of the observed measures, and each measure has some measurement error. The indicators are said to be **interchangeable** in the sense they all are thought to measure the same construct. With interchangeable indicators, the $\varepsilon$ often can be conceptualized as random error, hence they reflect unreliability, but they also can contain systematic error, as I discuss shortly.[3]

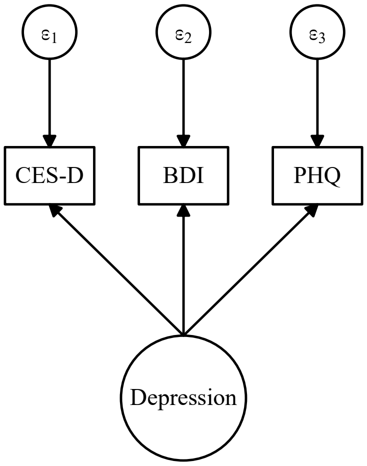


**FIGURE 3.2.** Multiple indicator measurement model

A powerful feature of SEM is that it provides estimates of parameters after adjusting for modeled measurement error. Figure 3.3 presents an example where we seek to estimate the correlation between depression (a mediator) and alcohol use (an outcome) and each construct has three interchangeable indicators. The "true" correlation is represented by parameter *a*. We have nine pieces of information to estimate this correlation, namely, the observed correlations between A1 and D1, A1 and D2, A1 and D3, A2 and D1, A2 and D2, A2 and D3, A3 and D1, A3 and D2, and A3 and D3. This puts us in a far stronger position than if we had only a single indicator of each construct. SEM allows us to use all of this information to estimate parameter *a* as well as to obtain estimates of measurement error ($\varepsilon_1$ through $\varepsilon_6$). It also provides perspectives on the

[3] The meaning of a latent variable depends on the broader theory in which it is embedded. I treat it here as the hypothesized "true" construct of interest. However, its technical meaning is more nuanced (Bollen & Hoyle, 2012).

strength of the relationships between the constructs and their indicators ($\lambda_1$ through $\lambda_6$).

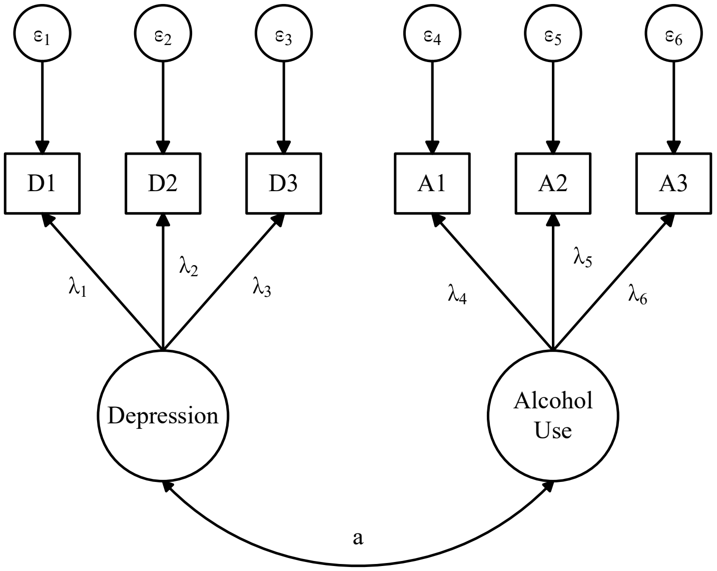


**FIGURE 3.3.** Estimating correlations with a multiple indicator model

If a source of systematic error is unmeasured, there are still ways of taking it into account in SEM. Figure 3.4 presents the case where we seek to estimate the correlation between alcohol use and marijuana use. Each construct has two interchangeable indicators, a self-report (SRA for alcohol use and SRM for marijuana use) and a biological indicator (BA for alcohol use and BM for marijuana use). The two self-reports could be influenced by social desirability tendencies, thereby inflating the correlation between the two measures. If social desirability resides in $\varepsilon_1$ and also in $\varepsilon_3$, we would expect these error terms to be correlated, per Figure 3.4. In this case, SEM estimates the "true" correlation between alcohol use and marijuana use taking into account the correlation between SRA and SRM, SRA and BM, SRB and SBM, and BA and BM. It also adjusts the estimate for unreliability of the measures and for the correlated error.

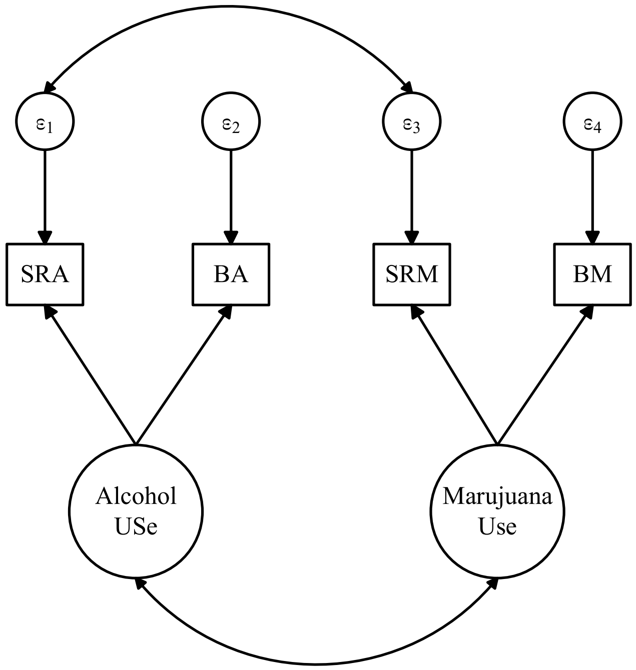


**FIGURE 3.4.** Systematic measurement error as correlated error

In sum, in RETs, we need to be concerned about the biasing effects of measurement error. Fallible measures can distort estimates of correlations and covariances between constructs, leading, in turn, to bias in estimates of causal coefficients. The best way to address measurement error is to use measures that are highly reliable and relatively free of systematic error. However, if this is not possible, we can use SEM to help correct for bias due to measurement error. This latter fact should not be seen as an invitation to use poor measurement in SEM-based research; SEM works best with reliable and valid measures and it too can be undermined by sloppy measurement.

In SEM, there is a phenomenon known as the naming fallacy (Kline, 2015). The fallacy refers to the belief that naming a factor or latent variable means it truly represents the name applied to it. Like exploratory factor analysis, the name assigned to a latent construct does not define that construct; rather, the nature and quality of the indicators of the construct help to define its meaning, per my earlier discussion of concept-measurement mapping.

I generally recommend that for RETs, when choosing multiple indicators of latent constructs one should seek to use interchangeable indicators. This is consistent with treating the indicator error terms as measurement error. However, I recognize that scenarios can occur where alternatives are required. For example, some researchers define a latent variable in terms of the common variance underlying distinct constructs, such as when a generalized locus of control latent variable (the extent to which someone feels they have control over what happens to them in life) is thought to underlie specific types of locus of control, such as lack of control due to luck, lack of control due to powerful others, and lack of control due to fate (Wallston, Strudler-Wallston & Devellis, 1978). Measures of the latter three constructs might be used as "indicators" of generalized locus of control in the sense that generalized locus of control is thought to influence, more or less, all three of them and thus represents the source of their common variance. The focus of the RET might be on changing this generalized locus of control with the idea that doing so will permeate through to the specific types of locus of control.

You need to be careful about treating two or more variables as reflective of a single latent variable just because they seem vaguely related or because they both fall within some general theoretical category. For example, a researcher might treat the behaviors of (a) voted in the last election, (b) contributed money to a political party, (c) worked in a political campaign and (d) watched election debates on television as indicators of the latent variable "political participation." One needs to ask in such cases if there truly is a well-defined common cause to these indicators that might be labeled "political participation" and if you want to turn these rather straightforward variables into what some might argue is a vague construct like "political participation" that reflects the common variance underlying the indicators. I personally think the power of multiple indicator strategies is maximized when they are invoked in the spirit of interchangeable indicators to address random and systematic measurement error, but I also recognize the potential utility of using distinct indicators to study meaningful higher order constructs.

## COMPOSITES, RELFECTIVE MEASUREMENT, AND FORMATIVE MEASUREMENT

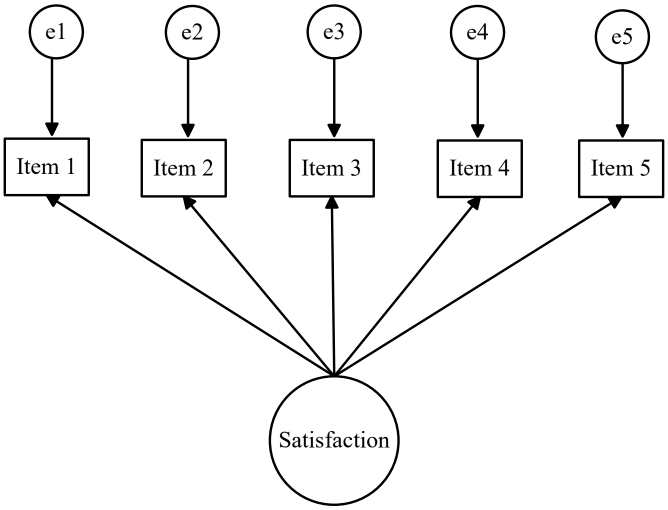
One of the most common approaches to measurement in social science research is the use of item composites. Although definitions of composites vary, I use the term here to refer to the summing or averagaging a set of items usually from a multi-item scale. Composites typically are represented as a weighted linear combination of item/variable responses:

$$\text{TS} = w_1\,\text{V1} + w_2\,\text{V2} + w_3\,\text{V3} + \dots + w_k\,\text{Vk} \qquad [3.2]$$

where TS is a total score for a person, V is the person's response to a given item on a multi-item scale, *w* is a weight attached to the variable/item response and *k* is the number of items. When we sum responses across items, the weights are uniform and all equal 1.0; when we average responses, the weights are uniform and all equal 1/*k*.[4]

Social scientists rely heavily on composites because working with individual items of scales in RET-based SEM often is impractical. If one assesses four constructs each measured by 15 items at two points in time, then if you analyze your SEM model using item level data, you must contend with a 4X15X2 = 120 by 120 covariance matrix in your statistical modeling, which can be unwieldy and sample size demanding. Items on scales often have a diffuse factor structure with large measurement errors, which also creates analytic challenges for latent variable modeling. With a composite approach, you would collapse the items for the four constructs into 8 "total scores" (4 constructs at two time points each) yielding an 8X8 covariance matrix for purposes of analysis. In applied RET program evaluations, it is not unreasonable to work with composites as long as one ensures that one's measurement house is in order in terms of reliability and validity and the formation of the composite fits with one underlying scaling theory.

Composites are formed using different measurement models. One approach adopts what is known as a **reflective measurement model**. Responses to items are assumed to be impacted by an underlying latent variable that represents the construct of interest, i.e., the items "reflect" the underlying latent construct. This measurement model is captured by an influence diagram of the following form, shown here for a measure of peoples' relationship satisfaction with their romantic partner for a five item scale:



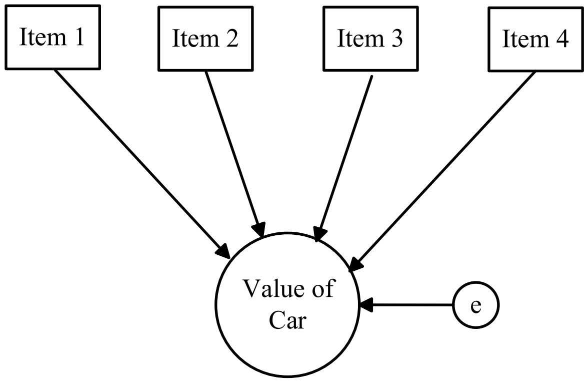
The scale might assess the extent to which people agree or disagree with items like "*I am very satisfied with my relationship with my partner*" or "*My partner and I understand*

[4] I assume in my discussion that negatively scored items for a composite are reverse scored.

*each other perfectly*" or "*Overall, our relationship is a success*." In such models, one expects the items to be highly correlated because they all share and should be dominated by a common cause, namely the underlying latent construct of relationship satisfaction.

A typical implementation of this approach is to use a scale in which the items have been factor analyzed. Those items that "load" highly on the same factor (and only that factor) are said to be unidimensional. If the loading magnitudes are fairly similar to one another for items that load on the same factor, then it is not unreasonable to sum or average them (i.e., to assume uniform weights), which is what researchers typically do. Summing or averaging scores presumes that any item differences between the magnitude of loadings that load on the same factor are not substantively or practically meaningful and/or possibly are sample specific due to factor analytic overfitting of the sample data. McNeish and Wolf (2020) caution researchers about making uniform weighting assumptions, arguing for the need to check for loading uniformity before summing or averaging items. They argue that loading uniformity is a specialized measurement model within the class of reflective models and that its properties need to be empirically affirmed.

A different measurement model is when the focal or latent construct is thought to be determined by the item properties, i.e., the focal or latent variable is "formed by" the observed variables in a causal sense. This approach is often referred to as **formative measurement** because the item properties cause the latent construct rather than vice versa. Consider the case where I assess the value of people's cars and express that value as a function of four attributes represented by four "items," the car's age (Item 1), its condition (Item 2), its size (Item 3), and its make (Item 4) per the following figure (note: curved arrow correlations between the items are not shown to reduce clutter):



In this case, the value of a car is determined by the item properties (car age, condition, size and make) with unspecified other determinants of value (as well as measurement error) lumped into the error term, *e*. The attributes captured by the items may or may not be correlated with one another; e.g., the make of a car is not necessarily correlated with its age. If a car's value changes over time, this does not necessarily mean the other defining causes of the car value change; the car does not turn from being a Mercedes into being a Ford as the car's value changes. Instead, being a Ford or a Mercedes is a determinant of being more or less valuable. To determine the value of a car that a person owns, we might standardize or place each item attribute onto a common metric of worth and then sum, average, or combine the scores in some way to form a composite of the overall value of the car. Despite our use of a composite as is done in traditional measurement discussed thus far, the underlying measurement theory is distinct from that of reflective measurement. If you believe your items can legitimately be uncorrelated or modestly correlated and you use a composite of the scale values of the separate items, then this signals that you may be working from the perspective of a formative measurement model with your composite index defined accordingly. By contrast, if you believe your items should all be highly correlated because they are being influenced by the same underlying construct, then this signals that you probably are working with a reflective measurement model.

Consider some other examples of formative measurement. In the literature on country economic development and world systems theory, a construct that has received a great deal of attention is the extent to which different countries have favorable conditions for development. This construct of "conditions for economic development" is often indexed by a composite known as the human development index (HDI). The HDI composite score includes country-level indices of health, education, and income. Scores on these "items" *define* the favorable conditions for development and a change in any one of the items literally changes the value of the focal construct (potential for development). This is formative measurement.

As another example, suppose the underlying construct is the frequency of relationship violence on the part of one's partner that a woman has been exposed to in the past 12 months. I might obtain a frequency judgment from women in each of four domains of relationship violence of the extent to which their partner has engaged in the category, namely (1) psychological aggression, (2) physical aggression, (3) sexual coercion, and (4) injury. I might assess multiple events within each category, such as exposure to different types of physical aggression, exposure to different types of sexual coercion, and so on. The items comprising the scale are not expected to be highly correlated with one another. An abusive partner who physically hits a woman (one item) may or may not also kick her (another item) and may or may not sexually coerce her (a third item). However, when summed across all items, the composite represents a total exposure index of relationship violence in a formative measurement sense. In this case, if the various items on the scale are not that highly correlated, this fact does not bother me as a psychometrician because I do not expect someone who does one violent behavior to, by definition, do all the other violent behaviors. However, summing across the items meaningfully gives me a sense of the total exposure to relationship violence that the person has experienced. This is a form of formative measurement.

As a final example, the number of calories burned in a day (the underlying construct) is impacted by how much a person has walked that day, whether a person has engaged in each of a set of exercise activities that day (e.g., playing basketball, using a treadmill), and a host of other such variables. The engagement in each of these separate activities are assessed as "items" using a metric of the typical number of calories burned for that activity and then combined into a composite to determine the total number of calories burned. In such an approach, it is not the case that the items need be highly correlated. A person who performs one activity may or may not perform one or more of the other activities; there are many ways one can burn calories. Nor are the "items" assumed to be unidimensional. This is an example of formative measurement.

There is a substantial literature on psychometric representations of formative measurement models and formative constructs in the social sciences (Bollen & Lennox, 1991; Bollen, & Bauldry, 2011). Controversy exists about the best way to construct and model them in SEM settings (Diamantopoulos et al., 2008; Bollen & Diamantopoulos, 2015; MacCoun, 2013; see also the special issue on formative measurement edited by Diamantopoulos, 2018). In my opinion, the literature surrounding reflective measurement, formative measurement, and composites is fraught with inconsistent definitions, divergent psychometric characterizations of the concepts, and non-comparable measurement practices. I do not want to get sidetracked into that literature here. The main points I emphasize is that (a) for practical reasons, social science researchers rely heavily on single indicator composites when working with measures of constructs, and (b) that the composites typically are conceptualized from the perspective of either reflective or formative-like measurement models. As you design your RET, you need to decide if the use of a composite is reasonable and then how you want to conceptualize the items designed to assess the presumed underlying construct, either as reflective or formative.[5] This, in turn, impacts how you evaluate the psychometric properties of those measures. For reflective measurement, you typically will test for composite reliability, unidimensionality of items, and loading uniformity, but this will not be the case for formative measures. For formative measures, items can be correlated or uncorrelated and they may or may not be unidimensional. However, the composite forms

[5] Hayduk et al. (2007) note a third measurement model called reactive measurement. See their article for details.

a meaningful whole for the underlying concept in ways that make conceptual and logical sense. I often encounter studies that are clearly working with formative-inspired composites but that then report alpha coefficients for them, which is psychometrically incoherent. For both reflective and formative measures, issues of reliability and validity need to be front and center but the ways we determine reliability and validity for the different types of measures can be distinct. For example, classic composite reliability statistical indices are appropriate for reflective measurement but not for formative measurement. See my website for a document called *Testing for Unidimensionality* that describes how to test unidimensionality, loading uniformity, and evaluate the composite reliability of reflective measures.

## Additional Qualities of Composites

Because multi-item composite measures are so widely used, I briefly consider in this section five decisions or cautions with respect to them that you should keep in mind, (1) whether to use differential weighting of items (2) whether to standardize items, (3) whether to sum or average scores across items, (4) the presence of minor factors, and (5) aggregating scores across subscales.

### *Uniform Weighting versus Differential Weighting*

As noted, summing or averaging item responses imposes uniform item weighting. By contrast, some researchers choose instead to use differential weights for items. When differential weights are used, the weights often come from factor analysis, with the total score taking the form of **factor scores** for the latent variable in question. The weights typically are assigned to best account for the covariances between the items, but other criteria can come into play as well. Principal components analysis generates scores based on weights to yield what are known as **component scores**. In contrast to factor scores, the weights seek to maximally capture the total item variance in the observed variables comprising the composites. The goals of factor scores and component scores are distinct and can result in different composite scores because of the use of different item weights.[6] Yet another approach to defining weights for a multi-item scale is to use **partial least squares** (PLS). This approach works with items for two constructs X and Y where X is a predictor of Y. The weights are assigned to each composite in a way that maximizes the squared correlation between the two composites (Rönkkö et al., 2016).

[6] A criticism of factor scores is the problem of factor score indeterminacy (Schönemann & Steiger, 1978), which refers to the fact that there are many sets of factors scores that satisfy the requirements of a common factor model. The choice of which set to use is said to be arbitrary, implying that factor scores also are arbitrary. This problem does not plague component scores nor is it problematic in confirmatory factor analytic models in SEM that posit factor structures in which indicators "load" on only one factor with zero loadings on other factors.

Proponents of differential item weighting for composites argue that doing so is more realistic because it recognizes that some items are more strongly related to the underlying latent construct than others. The more relevant items, the argument goes, should be given more weight. Critics argue that when using empirically defined weights, the weights can vary across different samples thereby potentially changing the meaning of the composite from one sample to the next (see my discussion below of loading invariance). They also argue that weight differences can be unreliable due to sample-specific overfitting. Finally, some critics argue and provide empirical evidence for the fact that differential weights when applied to reasonably constructed multi-item scales using reflective measurement rarely lead to different conclusions than uniform weighting so that the use of differential weighting really does not matter all that much (Rönkkö et al., 2016). Arguments by both proponents and critics of differential item weighting when forming composites have merit and I can't unequivocally recommend one approach over the other across all RET scenarios. You should consider differential weighting if you have solid theoretical or substantive reasons for justifying weight differences or if item analyses speak forcefully to using differential weights. I recommend Rönkkö et al. (2016), Henseler (2020), Bobko et al. (2007), Rhemtulla et al. (2019), and Bollen and Diamantopoulos (2015) for informative discussions of these issues.

### *To Standardize or Not Standardize*

With uniform weighting, it is not necessarily the case that each item contributes equally to variability in the total score. This is because under the scenario of equal weighting, items with more variability contribute more to total score variability than items with less variability. A simple example makes the point obvious: If my composite consists of two items and one has a standard deviation (SD) of 3 and the other an SD of .05, when I sum the two items, the variability in the total score will be dominated by the first item.

Sometimes letting items contribute to total score variability as a function of naturally occurring variability makes conceptual sense. Other times, the differential variability is artifactual due to item wording. An item that states "Last week, I was very sad" likely will have different variability than if the item reads "Last week, I was sad." This is because more extreme statements tend to push people away from one or the other metric extremes, thereby affecting variability. If differential item variability is a product of arbitrary wording choices, one might want to remove those variance differences. A strategy some researchers use is to standardize items before summing or averaging them. This process ensures a standard deviation of 1.0 for each item. Care must be taken in such cases because you implicitly use additional parameters to form the aggregate, namely item variances. This can limit generalizability of your results to other populations.

Several researchers (e.g., Cohen et al., 2003; Widaman et al. 2011; Little 2013) suggest the possible use of a scoring method known as **percent of maximum possible** (POMP) scoring. Each item is rescaled to occur between 0 and 1, and then averaged into a total score. For example, if an item is originally scored on a 1 to 10 scale, you can subtract 1 from each person's score so scores now are between 0 and 9 and then divide by 9. POMP scoring is advantageous in that all item scores fall on a scale with the same metric (0 to 1) but the items can differ in terms of their variance, thus avoiding the standardization problem of forced equal variances. A score of 0.50 on an item means the person scored halfway between the lowest possible score and the highest possible score; a score of 0.80 means the person scored 80% up from the lowest possible score; and so on.

### *To Sum or to Average*

Suppose you have a 10-item composite where each item is responded to on a 1 to 5 rating scale of 1 = strongly disagree, 2 = moderately disagree, 3 = neither agree nor disagree, 4 = moderately disagree, and 5 = strongly disagree. If you sum scores across the 10 items, the total score ranges from 10 to 50. If you average scores, the total score ranges from 1 to 5. Whether you sum or average the item scores when forming a composite will not affect the correlation between the total score and other variables nor will it affect their significance tests in traditional regression modeling. In my opinion, however, the averaged scores often are easier to interpret and can make more intuitive sense when used to generate unstandardized regression coefficients. By averaging, we can readily relate the total score to the verbal anchors of the response metric of the individual items. In the above example, a total score near 1.0 means the person tended to strongly disagree with all items; a total score near 5.0 means the person tended to strongly agree with all items; a total score near 4.0 means the person tended to moderately agree with items; and so on. For regression coefficients, the coefficient is the predicted mean change in the outcome given a one unit change in the metric of the predictor. With the averaging strategy as applied to a predictor, a one unit change in the predictor has associated with it intuitive verbal anchors, such as from "strongly disagree" to "moderately disagree," or from "moderately disagree" to "neither agree nor disagree." By contrast, in summation scoring, a one unit change for scores of, say, 30 to 31 is difficult to grasp. To me, a one unit increase in an averaged total score usually is more interpretable than a summed score.

### *Unidimensionality and Minor Factors*

Sometimes a strict unidimensional model will not fit a set of items well because in addition to the single major factor underlying item responses there also exist minor factors that are localized and that selectively influence a few of the items in minor ways.

The presence of these minor “factors” do not change the fact that a single major factor dominates covariation between the items, but the minor factors can disrupt model fit at the item level and potentially obscure the dominance of a single factor model.

Consider the one factor model in Figure 3.5 in which a generalized communication quality factor between parent and child is thought to influence adolescent agreement with statements about parent-adolescent communication quality in six topical domains, X1 to X6. The estimated standardized factor loadings shown in the Figure are large and it turns out the correlations between the indicators are large. I can see this because in a single factor model, the correlation between two variables is the product of their factor loadings. Suppose that X1 and X2 are both in topical domains that pertain to school but this is not true of the other domains. The predicted correlation between X1 and X2 is (0.85)(0.85) = 0.72. Suppose that the observed correlation between the two items is 0.86 and, further, that all of the other observed correlations between X1 through X5 are well reproduced by the product of their respective loadings; it is just this one particular pair of variables that the correlation is under-predicted by a single factor model. This suggests there is a “minor factor” that needs to be taken into account, i.e., a factor that serves as an additional, weaker common cause of just X1 and X2. This factor can be modeled in different ways, but one approach adds correlated errors to X1 and X2 because each error likely contains the same localized common cause that impacts communication quality for just items X1 and X2. Perhaps this unmeasured common cause is conflict between parents and adolescents about schoolwork independent of general communication quality. This minor common cause influences X1 and X2 but not the other domains represented by X3 through X6. Figure 3.6 presents the model.

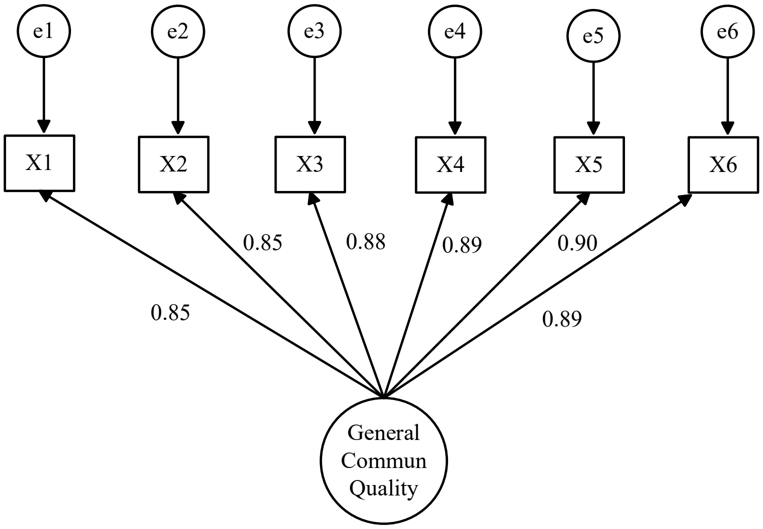


**FIGURE 3.5.** Unidimensional Model of Communication Quality

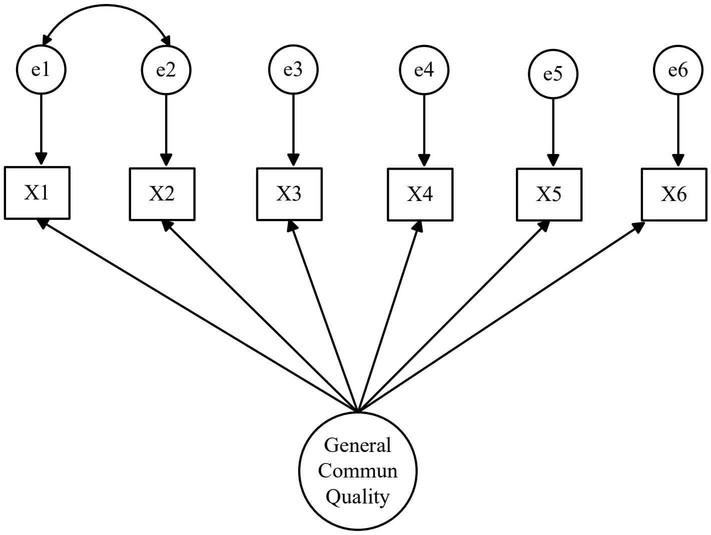


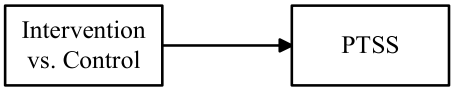
**FIGURE 3.6.** One Factor Model with Correlated Residuals

With unequal factor loadings and the presence of minor correlated residuals, it turns out Cronbach's alpha often is not a good estimate of the reliability of the composite. The indices of omega-hierarchical or omega-categorical tend to do better.

*Aggregating Across Subscales*

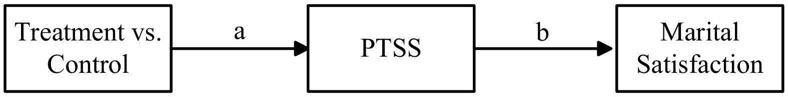
A frequent empirical practice in the social sciences is to factor analyze items on a scale with the goal of identifying subscales or subcomponents of the target construct. It is not uncommon to identify 3 or 4 subscales that are modestly correlated with one another but that are conceptually distinct. For example, factor analytic studies have identified four subscales for social support (1) emotional support, (2) tangible support, (3) informational support, and (4) companionship. Despite the widespread practice of identifying subscales, researchers often aggregate across the subscales to form a total score for the construct and then use that total score in program evaluations. This practice can be justified using a formative measurement logic model relative to the subscales, but it then makes no sense to report a coefficient alpha or composite reliability that assumes unidimensionality of the composite. Also, you need to be careful of aggregation biases for the total score.

The issue of aggregation bias with subscales is fundamental for RETs and it merits elaboration here. To illustrate my points, I use as an example the classic four factor/subscale representation of post-traumatic stress symptoms (PTSS) that distinguishes four clusters of symptoms that tend to co-occur within a symptom cluster but not necessarily across clusters, (1) **re-experiencing**, which includes symptoms like having distressing and unwanted memories of the traumatic events, having vivid nightmares about the events, and experiencing flashbacks, (2) **event avoidance**, which includes avoiding activities, places, and people that might bring back memories of the trauma, (3) **emotional numbing**, which are symptoms associated with losing interest in positive activities, feeling distant from others, and experiencing difficulties having positive feelings, and (4) **hyperarousal**, which are symptoms associated with difficulty sleeping, feeling irritable, quickly losing one's temper, having concentration difficulties, and feeling constantly on-guard. Suppose I form a composite of items for each subscale that assess the extent to which combat-experienced military veterans experience each symptom type and I then aggregate these subscales into a total score representing the overall construct of PTSS. I might test an intervention to reduce PTSS on the total score via the following influence diagram:



Suppose the intervention produces a statistically significant difference between the intervention and control groups on the mean total PTSS scores. I might conclude that I have indeed developed an intervention that impacts PTSS. But suppose the only component of PTSS that the intervention truly affected was that of event avoidance. I would not know this if I failed to analyze program effects on the separate PTSS subscales. My overall conclusion might be correct in that I have indeed lowered PTSS, but the message I *should* take away is that my intervention was only partially successful because it failed to impact re-experiencing, emotional numbing, and hyperarousal. If I had conducted subscale analyses, I would learn that aggregation bias has misled me.

Taking aggregation dynamics a step further, suppose I treat PTSS as a mediator to a distal outcome, marital satisfaction, with the idea that higher levels of PTSS lead to lower levels of marital satisfaction. Here is the influence diagram I might posit:



Suppose path *a* in my evaluation study turns out to be statistically significant as does path *b*. This pattern of results might lead me to think I will obtain differences in marital satisfaction as a function of my intervention. However, suppose the underlying subscale dynamics are such that (1) the effect of the intervention on PTSS is entirely due to its effect on event avoidance, and (2) the effect of PTSS on marital satisfaction is entirely due to the effect of emotional numbing on marital satisfaction. The net result is that the intervention will not affect marital satisfaction despite the presence of significant *a* and *b* paths. Only an analysis of the PTSS subscales would reveal why this occurs; it is because the intervention fails to affect the "active ingredient" of emotional numbing for marital satisfaction; it instead influences event avoidance which is irrelevant to satisfaction.

There are interesting measurement models that can be explored when using subscales. Figure 3.7 shows one example that uses higher order factor analysis. Boxes S1 through S12 are measured indicators of different PTS symptoms. The four latent factors in the middle of the diagram represent the different symptom types that underlie different manifest symptoms. Each of these latent factors are thought to be influenced by a higher order factor (which I call *generalized post-traumatic stress*), which results in the four factors being correlated by virtue of common cause dynamics. However, each specific PTS factor also has unique variance that is distinct from generalized PTS reactions (d2 through d5). The intervention is thought to influence global PTS and, to the extent it does, the effects on it will filter through to each of the separate symptom specific factors. The dashed arrows reflect the possibility that the intervention has independent effects on the four specific factors over and above its effect on global PTS. Note that I can represent a simplified version of this model using composites, as shown in Figure 3.8.

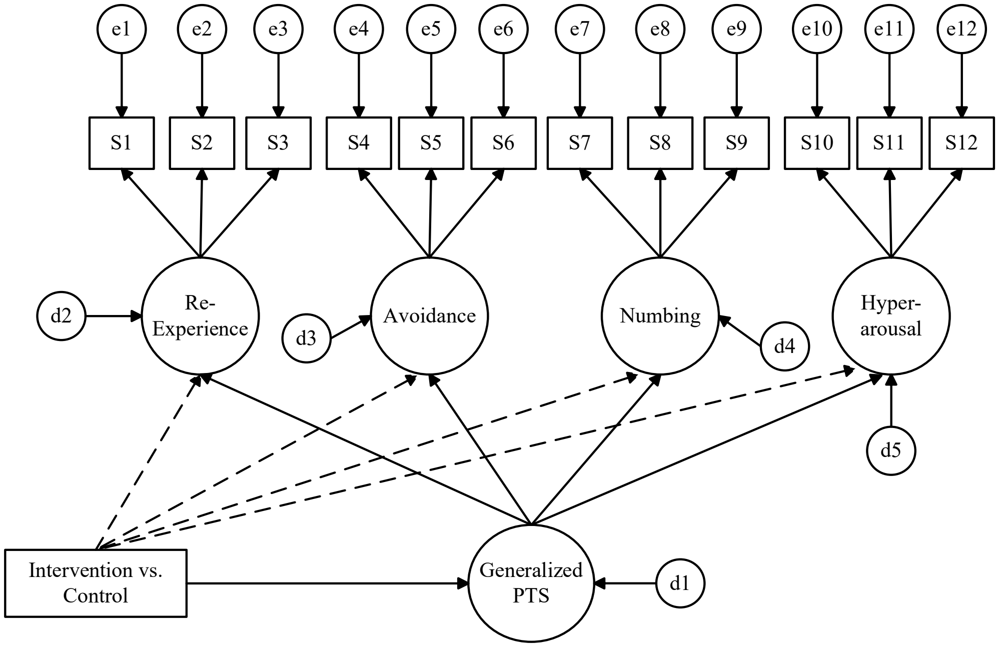


**FIGURE 3.7.** Higher Order Factor Model for Treatment Effect

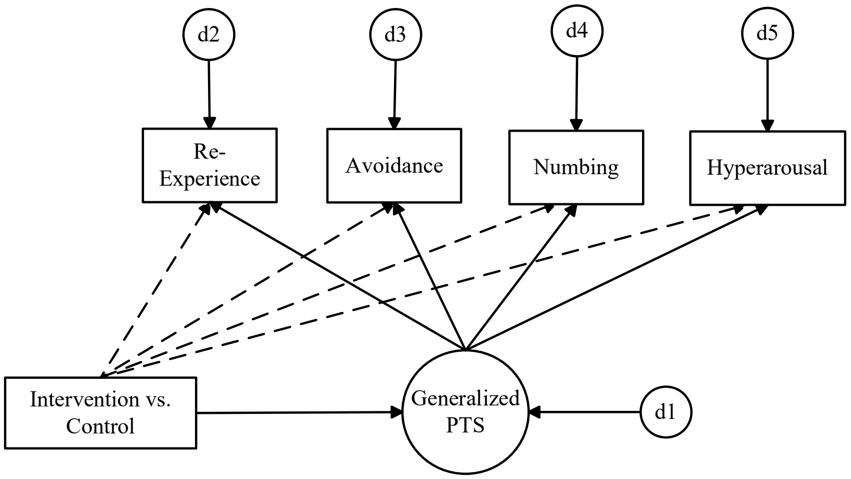


**FIGURE 3.8.** Higher Order Model with Composites

*Concluding Comments on Composites*

In sum, when working with composites, you need to make decisions about whether to use differential weighting of items, whether to standardize items, whether to sum or average scores across items, whether to treat a measure as functionally unidimensional despite the presence of minor factors, and whether to aggregate scores across subscales, which probably is reasonable to do if all the subscales are highly correlated. In models that explore unidimensionality, a one factor model can be fit to the data that includes the presence of this minor "factor" and the model fit should be improved accordingly. Technically, the model is no longer a one factor model because we have acknowledged the presence of a minor factor that influences X1 and X2. But it also might be evident based on large standardized loadings for each item for the generalized communication quality factor that the dominant source of the correlations between the various X is generalized communication quality. In this case, we can still form a meaningful composite for the items, with the caveat that there is minor contamination between X1 and X2 with respect to the total score.

There is considerable controversy and conflicting advice associated with the use of composites, especially composites based on unit weighting, such as sums or averages (e.g., Lai & Tse, 2024; Liu & Pek, 2024; McNeish, 2018, 2023; McNeish & Wolf, 2020a,b; Sijtsma, Ellis & Borsboom, 2024); Widaman & Revelle, 2022, 2023). I tend to gravitate to the arguments of Sijtsma, Ellis and Borsboom (2024) in their article titled *Recognize the Value of the Sum Score: Psychometrics' Greatest Accomplishment*, but I fully recognize the ultimate need to map one's measurement strategy onto the underlying psychometric model and scaling theory that one adopts in a given research study.

## BREADTH VERSUS DEPTH OF CONSTRUCT COVERAGE

A difficult decision you will face when designing an RET is whether to collect multiple interchangeable indicators of the same construct or to collect many single indicators of different constructs. Collecting multiple interchangeable indicators usually puts you in a stronger position to statistically adjust for measurement error and to use multiple pieces of information to estimate relationships. However, given time and resource constraints that typically apply, you then may be forced to assess fewer constructs of interest. You need to find a balance between assessing all the constructs that are essential to the evaluation and optimizing measurement practices for strong scientific inference.

Some methodologists disparage single item measures of psychological constructs but the fact is there is considerable evidence that shows that single item measures often perform admirably (Allen, Iliescu & Greiff, 2022; Jaccard, Weber & Lundmark, 1975). Strategies exist for adjusting for measurement error in SEM for single indicator cases (Bollen, 1989; Hayduk, & Littvay, 2012; Oberski & Satorra, 2013; Cole & Preacher, 2014). To the extent that these methods are viable, then one can increase the number of constructs assessed because the need for multiple measures of the same construct is lessened. The single indicator measurement model in Figure 3.1 generally cannot be estimated in SEM because it is statistically under-identified; there are too many unknowns relative to knowns to obtain a unique solution of both the error variance and the path coefficient linking the latent variable to the observed indicator. The most common approach to resolving this problem is to make an educated guess about the unreliability of the measure and then fix the error variance to that value in your modeling. SEM will then make adjustments to the parameter estimates to take this into account. I show how to implement this strategy in a primer for the current chapter on the resources tab of my webpage. I also address in that primer alternative strategies to address measurement error, such as the use of item parceling for multi-item scales.

The ability to address measurement error using SEM is a decided strength of the technique. However, as noted, this should not be taken as an excuse for using substandard measures. The typical advice to those using SEM is to get one's measurement house in order before pursuing model analysis. The fact is that SEM brings with it many of its own (tenuous) assumptions, as I outline in future chapters. The challenge of creating correctly specified models and ruling out alternative models that might equally well represent the data while also having to model multiple sources of systematic measurement error and complex measurement error structures is not a good situation to be in. Using psychometrically sound measures can greatly simplify the modeling process.

Another strategy researchers can use to reduce measurement burden is to create short versions of multi-item existing measures by reducing the number of items to assess targeted constructs (Widaman, Little, Preacher & Sawalani, 2011). Measure reliability can be affected by the number of items comprising the scale, in part, because the greater the number of items, the more likely it is the positive and negative random noise will cancel each other out when the items are averaged to form a composite. Given this, shorter measures generally will be less reliable, everything else being equal. The need for more items, however, can be offset by ensuring insisting on high internal consistency among the items (i.e., high shared variance among the items) and large degrees of true score variability for the items. In general, when working with unidimensional constructs, one should select subsets of items that (a) remain faithful to the underlying construct, (b) are highly correlated with one another, and (c) have more rather than less true score variability. A 5-item unidimensional scale with internal consistency of 0.50 (i.e., all items are intercorrelated about 0.50) will have a composite reliability near 0.83; a 3-item scale will have a reliability of 0.75. For internal consistencies of 0.40, the corresponding composite reliabilities for a 5 and 3 item scale are approximately 0.77 and 0.67.

## MEASUREMENT INVARIANCE

The concept of **measurement invariance** focuses on the problem of group differences in the way people orient to items on a scale thereby creating artifactual differences between groups on means or on correlations between variables. The technical definition of measurement invariance is best illustrated using an example outside the context of an RET. Suppose I want to compare two groups, males and females, on their mean levels of depression. Using Equation 3.1, I can specify a latent variable model for depression scores in each group. For males, the model is

$$D_M = \alpha_M + \lambda_M \; LD_M + \varepsilon_M$$

where D is the observed depression measure for a given male, LD is the latent variable of depression (i.e., the "true" depression), $\alpha$ is the measurement intercept, $\lambda$ is the measurement slope (also called a loading or a factor loading), and $\varepsilon$ is an error term reflecting measurement error. This equation is just a linear equation with an intercept ($\alpha$), a regression coefficient ($\lambda$) and an error term ($\varepsilon$); D is said to be a linear function of LD. The corresponding model for females is

$$D_F = \alpha_F + \lambda_F \; LD_F + \varepsilon_F$$

Suppose I calculate a mean D score for males and a mean D score for females and, in doing so, also calculate the mean of the right-hand side of the two equations. I obtain the following:

$$\mu_M = \alpha_M + \lambda_M \ \mu_{LD\text{-}M} \qquad [3.3]$$

$$\mu_F = \alpha_F + \lambda_F \ \mu_{LD\text{-}F} \qquad [3.4]$$

where $\mu$ refers to a mean, using population notation. Note that the error scores drop out of these equations because the mean of the error scores is always zero; the random positive errors cancel the random negative errors when the errors are averaged. I now develop several implications of Equations 3.3 and 3.4.

In practice, my primary interest is in making statements about the true mean depression differences for males and females, but all I have to work with are the two observed means. If the group latent depression means are equal, I would hope that the observed group means also are equal. It turns out that even if the latent depression means are equal, I can mistakenly find a non-zero observed mean depression difference if the two measurement intercepts are unequal, i.e., if $\alpha_M$ does not equal $\alpha_F$. Stated another way, to make a valid comparison between the groups with respect to their mean levels of depression, I must assume the measurement intercepts are **invariant** across the groups.

What could cause the intercepts to be different? Suppose an item on the depression inventory asks respondents to report how many days in the past two weeks they have cried. It turns out that women tend to cry more often than men independent of depression (Vingerhoets, 2013). If we identify a group of men and women who have the same levels of average depression (i.e., their mean LDs are the same), I would find that the women would report more instances of crying than men simply because women tend to cry more in general. This represents a case where the two measurement intercepts are different for this item. If I examine the observed responses to this "depression" item, I would conclude that women, on average, are more depressed than men when, in fact, this conclusion is wrong. It is just that women cry more in general. This item is not a generalizable indicator of depression and should not be used on a depression scale.

The comparison of observed mean depression scores for two or more groups to make inferences about group differences in true depression also requires another measurement assumption. Inspection of Equations 3.3 and 3.4 reveals that in addition to equal measurement intercepts, we also must assume the two measurement slopes, $\lambda_M$ and $\lambda_F$, are equal. If they are not, then we again might mistakenly infer the groups differ on true latent depression means when, in fact, they only differ on the measurement slopes. Different measurement slopes imply that the observed score D is calibrated differently to changes in the underlying latent depression for males and females. If $\lambda_M = 0.50$ and $\lambda_F =$

0.80, this means that for every one unit true depression changes, D will change, on average, only 0.50 units for males but it will change 0.80 units for females. These calibration differences can undermine mean comparisons for groups and they can create false interactions in regression analyses. It is like comparing income for two groups or performing a regression analysis in two groups using income as a predictor or outcome, but in one group income is measured in dollars but in the other group it is measured in pesos. The calibration of dollars and pesos differs as indicators of income.

Measurement non-invariance as a source of artifactual results in RETs is something we must be careful about. I develop this idea more in Chapter X. Issues of measurement invariance also are important for longitudinal analyses where measurement intercepts and measurement slopes can change over time. For example, as children age, they show changes in cognitive and emotional development as a function of maturation and these changes can alter measurement intercepts and/or measurement slopes. A strength of using SEM to analyze RET data is that it can provide perspectives on measurement invariance across groups or across time. On the resources tab of my webpage, I provide a primer for how to test for measurement non-invariance and how to deal with it should it occur.

## MEASUREMENT-INTERVENTION CORRESPONDENCE: CREATING STUDY SPECIFIC MEASURES

A great deal of scientific research relies on measures of constructs that have been subjected to rigorous psychometric evaluation in prior research. However, when evaluating programs, it is likely you will need to develop some measures specific to your goals and the context in which the evaluation takes place. A heuristic I use for screening measures of my mediators and outcomes is to examine each item on the measure and then envision if I can see the program altering responses to each item. If this is not the case for most items, I may not use the measure (or I ask myself why the intervention seems irrelevant to most items of the measure). For example, when working with measures of the construct of hope, I sometimes find the items on widely used measures refer primarily to hopefulness in the distant past (which no program can change) or they are so general that they likely will not be responsive to the specific hope constructs my program targets. It is important to have correspondence between what the program targets and the measures used to reflect changes in those targets.

When constructing my own self-report measures, I find it useful to take into account three core processes. First, people must understand the question posed to them i.e., there must be **comprehension**. Second, people must form or retrieve from memory judgments and opinions in response to the posed question to form an answer in their minds. This **judgment** process typically involves cognitive and affective mechanisms playing themselves out in a person's working memory. Third, once the judgment/opinion is formed, people must communicate that judgment/opinion to me. Sometimes people provide open-ended responses and other times they make ratings on a scale. The act of communicating one's answer to the investigator is called **response translation**. I have found that formulating specialized measures greatly benefits from thinking about factors affecting comprehension, judgment, and response translation for the population I am studying. For a detailed description of how to use this framework for measure construction, see Jaccard and Jacoby (2020).

A useful tool for measure refinement is **cognitive response testing** (Beatty, 2004; Lapka, Jupka, Wray & Jacobsen 2008; DeMaio & Rothgeb, 1996; Willis, 2004). Cognitive response testing is a form of qualitative research that allows one to identify assessment problems by asking a small number of people who are representative of the study population to complete the measures but then to paraphrase the items, discuss thoughts or emotions that came to mind when answering questions, and offering suggestions for measure improvement. Even if this feedback is obtained from only a few individuals prior to conducting an RET, the benefits can be considerable.

One commonly used approach asks questions in seven different categories after a person has answered each question. Consider the target question "how many times did you go to a doctor in the past 12 months?" The categories are (1) comprehension and interpretation (e.g., "what does the term 'doctor' mean to you?"), (2) paraphrasing ("can you repeat the question I just asked in your own words?"), (3) confidence ("how sure are you of your answer? Why?"), (4) recall probe ("how did you remember that you went to the doctor five times in the past 12 months?"), (5) specific probes ("I noticed that you hesitated at one point - tell me what you were thinking when you hesitated") (6) general probes ("how did you arrive at that answer?  Was the question easy or hard to answer?") and (7) improvement probes ("can you think of ways we can improve the question?").

Cognitive response testing helps identify sources of random error and systematic error in measures. It can be applied to new measures or to existing measures that have been developed in populations that are non-trivially different from the population in your RET. For a useful summary of the approach, see Jaccard and Jacoby (2020).

## CONDUCTING MEASUREMENT ORIENTED PILOT TESTS

I often find it useful to conduct small scale psychometric studies prior to embarking on a formal RET to ensure my measures will have acceptable reliability and validity for the RET. How extensive the study is depends on time and budget constraints. When I conduct my own federally funded scientific research to evaluate a program I have developed, the psychometric study can be more elaborate. When hired by an agency to evaluate their program using an RET, I may only be able use small scale cognitive response testing. Most of my research is with low income, inner city adolescent populations and I always bring a healthy skepticism about whether the psychometric history of well-developed measures in the scientific literature generalize to my populations and settings. A typical psychometric study I conduct begins with cognitive response testing for half a dozen or so respondents. After measure refinement based on these data, I administer my measures to about 80 or so individuals (more, if possible; less, if constraints dictate otherwise) in a test-retest design with about a one-week interval between assessments. Longer time frames run the risk of the constructs changing over time, thereby confounding reliability with stability. Shorter time frames run the risk of people trying to recall their earlier responses, which I encourage them not to do in instructional sets for the second assessment. I then evaluate the test-retest reliability of every item to identify reliability issues. I resolve them, as needed, through additional cognitive response testing or by eliminating or refining items.

I include in the psychometric study a social desirability measure (for a brief measure, see NieBen et al., 2019 on my website) and, where possible, I explore other response sets similar to those described by Baumgartner and Steenkamp (2001). I flag items or composite measures that are artifactually correlated with social desirability and make revisions or adopt better instructional sets to minimize such bias. I also include measures of variables that allow me to assert validity by showing that my measures predict other variables they are supposed to predict and are uncorrelated with measures they should be uncorrelated with, based on prior research or logic. My focus is on empirically building a case for the concurrent or convergent validity of my measures. In their article “Measurement Matters”, Fried and Flake (2018) provide recommendations for improving measurement. One of their recommendations is the following: “Stop using Cronbach’s alpha as a sole source of validity evidence. Alpha’s considerable limitations have been acknowledged and clearly described many times (e.g., Sijtsma, 2009). Alpha cannot stand alone in describing a scale’s validity.” My psychometric pilot studies are intended to embrace this recommendation. The general idea is to proceed to the RET with measures I am confident in and that can withstand scientific skepticism. I often role play a scenario of hostile reviewers determined to undermine conclusions I make from my RET on measurement grounds. I try to make sure I can effectively counterargue any objections they might offer.

## FALSE DICHOTOMIZATION OF MEASURES

In some disciplines, researchers frequently dichotomize or trichotomize continuous measures to make them easier to work with or to facilitate interpretation of regression coefficients associated with them. The noted biostatistician Stephen Senn (2012) refers to the practice as **dichotomania**. Unless there is strong theoretical justification for doing so, the practice is questionable (MacCallum, Zhang, Preacher & Rucker, 2002). Consider the case of dichotomizing a measure of IQ that ranges from 60 to 140 with a median of 100 by using a median split. In doing so, you essentially reduce a relatively precise scale to a crude, two-point scale ("low" IQ versus "high" IQ). You treat someone with an IQ score of 60 as being the same as someone with an IQ score of 99 (because both are in the "low" group) and you also treat someone with an IQ score of 101 as being the same as someone with an IQ score of 140 (because both are in the "high" group). At the same time, you treat a person with an IQ score of 99 (who is in the "low" group) as being different from someone with an IQ score of 101 (who is in the "high" group). This is not good practice. If you seek to control for annual income in a regression analysis and use a covariate that is a median split ("low" versus "high" income), have you really adequately controlled for annual income in light of the above dynamics? Probably not.

Studies have shown that false dichotomization or trichotomization can reduce statistical power, bias estimates of effect size, and create spurious main effects and interaction effects (MacCallum et al., 2002). Ironically, some researchers use these properties as misplaced justification for dichotomization. The argument is that given decreased power, dichotomization yields a more conservative test, so a statistically significant result is that much more impressive. This argument places too much emphasis on statistical significance and ignores the fact that dichotomization yields biased effect size estimates. Conservative tests are not an ideal; accurate inference is. If conservativeness is the ideal, then we also should prefer small sample sizes and unreliable measures in our research, but we do not.

Other investigators report that the estimated correlation between variables sometimes increases after dichotomization, thereby questioning the criticism that dichotomization reduces statistical power. MacCallum et al. (2002) show that this can occasionally happen for the case of small correlations with small sample sizes, primarily because of sampling error. However, it does not change the fact that effect size estimation with false dichotomization is biased, which, again, is a property we want to avoid. MacCallum et al. (2002) rebut other faulty justifications, such as the argument that dichotomization results in more reliable measures or that dichotomization simplifies matters. They build a strong case that the practice should be avoided unless strong theory or substantive considerations dictate otherwise. If a continuous variable is to be categorized, research suggests it is best to create a minimum of five to seven categories (e.g., Bollen & Barb, 1981; Green, Akey, Fleming, Hershberger & Marquis, 1997; Lozano, García-Cueto & Muñiz, 2008; Lubke & Muthén, 2004), sample size permitting.

One common use of dichotomization in RCTs in clinical psychology is to convert a continuous outcome into a dichotomous index of whether a patient has shown “clinically significant change” (Jacobson & Truax, 1991). Specifically, a cutoff value on the outcome measure, such as a depression scale, is defined so that a value below the cutoff is deemed as “acceptable” whereas a score above the cutoff is seen as “actionable,” such as requiring treatment or requiring continued treatment. In the vast majority of clinical interventions, cutoffs are defined using the number of standard deviations (SDs) a score is above the mean of the study population or some other referent population. For example, a cutoff of 1.5 SDs above the mean is often used. A problem with this approach is that standard deviations themselves sometimes have an arbitrary character; they are influenced by the wording of items, the adverb qualifiers used for individual items, and instructional sets. As well, the mean value of different depression scales can map onto different locations of the true underlying dimension of depression. For example, the mean of the Beck inventory might reflect more severe depression for patients than the mean of the PHQ-9 inventory for the same patients because of item wording and different rating scales associated with the items on the respective inventories. This implies a different standard is used to define a cutoff depending on the scale one uses because each scale has a different reference point on the underlying depression dimension. Finally, treating someone who is 1.49 SDs from the mean as a treatment “success” and someone who is 1.51 SDs from the mean as a treatment “failure” makes little scientific or clinical sense.

I fully understand the practical need of defining points of “action” in applied clinic settings. As such, a program evaluator might want to evaluate how a program affects such points of action. Having said that, dichotomization does not lend itself well to building a solid scientific knowledge base for documenting evidence-based treatment effectiveness.

## BASELINE ASSESSMENTS: DO WE COLLECT THEM OR NOT?

There is some controversy about whether one should obtain baseline assessments in randomized trials (Assman, Pocock, Enos, & Kasten, 2000; Bolzern, Mitchell, & Torgerson, 2019). Some RCT designs exclude baseline assessments yet, given random assignment, are still viewed as valid experimental strategies for program evaluation (see Chapter 4). By omitting a baseline assessment, one does not need to worry about the biasing effects that completing a survey might have on treatment effectiveness. For example, a program to reduce anxiety may only be effective if it is preceded by assessments that make salient to people their current anxieties and coping strategies.

It often is argued that an advantage of including baseline assessments is that it allows one to assess individual change and to explore correlates of individual change. However, as discussed in Chapter 4, change scores do not only reflect response to a treatment. They also reflect testing effects, history effects, instrumentation change, regression to the mean, maturation, demand characteristics, experimental mortality, and host of other time-varying artifacts, suggesting that analyzing individual change scores might not be as fruitful as many people think.

As discussed in Chapters 2 and 4, baseline assessments can yield more effective control of confounds and they also permit adjustments for sample imbalance. As well, baseline assessments allow us to test for baseline moderators of the effect of the treatment on mediators and outcomes. For example, a treatment for depression might be more effective for people who are moderately depressed at program entry as compared to those who are very depressed. This theoretical proposition can only be tested using a baseline measure of depression. Finally, using the baseline measure of the outcome as a covariate often increases statistical power (Raush et al., 2003).

In sum, there are both pluses and minuses to obtaining baseline assessments. My own opinion is that unless pretest by treatment interaction is a serious threat (see Chapter 4), I agree with the CONSORT recommendation to include baseline measures in randomized trials (Moher et al. 2010). To be sure, there are scenarios where doing so can cause problems (Glymour et al., 2005), but pursued judiciously and with methodological care, it usually is advantageous to include baseline assessments.

## FREQUENCY AND TIMING OF ASSESSMENTS

As noted in Chapter 1, causal theory assumes that a cause temporally precedes an effect. The time it takes for a cause to translate into an effect can vary. Sometimes the effect is evident instantaneously (within milliseconds), while other times it takes considerable time for an effect to reveal itself. For RETs, researchers need to think about how long it takes for the program to affect each mediator and how long it takes for changes in the mediators to translate into changes in outcomes because these factors impact choices about the frequency and timing of assessments. In addition, the temporal function relating the cause to the effect might be of interest. The effect might get stronger as time initially passes but then weakens at later time points, indicating that changes in the effect are non-linearly related to time. In the ensuing discussion of temporal dynamics, I focus first on treatment effects on mediators, ignoring outcomes. After establishing key concepts, I then consider how long it takes for mediator changes to translate into outcome changes.

The amount of change in a mediator, M, as a function of a program likely varies as treatment progresses. Suppose a treatment lasts 10 weeks and the mediator is measured on a 0 to 100 metric, with the program seeking to increase M. One way of documenting the underlying temporal change dynamics is to calculate the treatment versus control group mean difference on M at frequent intervals during treatment. Suppose I make weekly assessments of M during treatment. Table 3.1 presents 4 different effect patterns I might observe. The column labeled “M diff” is the mediator mean difference between the treatment and control conditions at a given point in time. In Profile 1, the program affects the mediator immediately and changes in the mediator increase linearly as a function of time in treatment. In Profile 2, the program does not begin to show effects until week 5 and then there is an increase with each passing week. This function has elements of an exponential distribution. In Profile 3, the program does not affect M until week 5, but when it does, the change immediately reaches its asymptote. In Profile 4, the same threshold function is evident but it happens in the first week. Note that all four profiles have the same effect on M by the end of treatment, but each profile presents a different characterization mid-treatment; at week 4, Profile 4 shows considerable change whereas Profiles 2 and 3 show no change. In week 5, Profiles 3 and 4 show equivalent change.

**Table 3.1: Profiles of Emergent Treatment Effects on a Mediator**

|  | Profile 1 | Profile 2 | Profile 3 | Profile 4 |
|---|---|---|---|---|
| Week | M Diff | M Diff | M Diff | M Diff |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 4 | 0 | 0 | 40 |
| 2 | 8 | 0 | 0 | 40 |
| 3 | 12 | 0 | 0 | 40 |
| 4 | 16 | 0 | 0 | 40 |
| 5 | 20 | 5 | 40 | 40 |
| 6 | 24 | 10 | 40 | 40 |
| 7 | 28 | 15 | 40 | 40 |
| 8 | 32 | 20 | 40 | 40 |
| 9 | 36 | 30 | 40 | 40 |

Researchers also often are interested in decay curves after treatment has finished. The same concepts apply, but now the initial reference point is the mean condition difference at program completion, namely at week 10. Table 3.2 presents different decay profiles for the mediator across a 3-month follow-up, again with treatment-control group differences assessed on a weekly basis. Profile 1 shows linear decay, so a single follow-up at any point in the 3 months will allow you to compute the slope in conjunction with the immediate posttest (week10). For example, if you obtain assessments at weeks 10 and 14, the average change per week is (40 - 32)/4 = 2.0 or 2 units per week. By contrast, Profile 2 shows exponential decay. Profiles 3 and 4 show something yet different, namely threshold decay but with different threshold points. Note that across the four program profiles, at the one-month follow-up (week 14), the program effects are quite different for the four programs. However at the 3-month follow-up (week 22), the decay is identical. Again, the more complex the function, the more assessments needed to discern it. As with treatment emergence, you may not be interested in the decay curves per se but only if the treatment effect is diminished at selected time points, such as at 6- or 12-months post treatment. In such cases, your assessment strategy would focus on these time points. If your focus is on isolating decay functions, you need to conduct "shortitudinal" research with frequent assessments at short intervals (Dormann & Griffin, 2015).

**Table 3.2: Profiles of Treatment Decay for a Mediator**

| | Profile 1 | Profile 2 | Profile 3 | Profile 4 |
|---|---|---|---|---|
| Week | M Diff | M Diff | M Diff | M Diff |
| 10 | 40 | 40 | 40 | 40 |
| 11 | 38 | 40 | 40 | 16 |
| 12 | 36 | 40 | 40 | 16 |
| 13 | 34 | 40 | 40 | 16 |
| 14 | 32 | 40 | 40 | 16 |
| 15 | 30 | 40 | 16 | 16 |
| 16 | 28 | 40 | 16 | 16 |
| 17 | 26 | 36 | 16 | 16 |
| 18 | 24 | 32 | 16 | 16 |
| 19 | 22 | 28 | 16 | 16 |
| 20 | 20 | 24 | 16 | 16 |
| 21 | 18 | 20 | 16 | 16 |
| 22 | 16 | 16 | 16 | 16 |

The situation becomes more complicated when we seek to map program effects onto both the mediator and the outcome simultaneously. Table 3.3 presents two of the previously presented four profiles but now focused on the emergence of treatment effects for both M and Y (Y also is measured on a 0 to 100 metric). I extend the table by two additional weeks, for reasons that will be apparent shortly. For Profile 1, the outcome does not appear to respond to changes in the mediator until two weeks after the mediator change. For Profile 2, the outcome seems to respond immediately to mediator changes. At the post-test (week 10), Profile 1 results in a different conclusion about program effectiveness than Profile 2. The Y treatment-control mean difference for Profile 1 at week 10 is 32 whereas it is 40 for Profile 2. However, if we waited two more weeks to make our posttest assessments, the full potential of the program with a Profile 1 pattern becomes evident. In ways, one should assess program effects on an outcome after we are reasonably certain the program has completed its influence on the mediators and those mediators have had time to exert their full influence on the outcome. Otherwise, we might mischaracterize the program effect. I revisit this matter in Chapter XX in depth, but the above gives you a sense of issues one needs to consider.

**Table 3.3: Profiles of Emergent Treatment Effects on a Mediator and Outcome**

| | Profile 1 | | Profile 2 | |
|---|---|---|---|---|
| Week | M Diff | Y Diff | M Diff | Y Diff |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 4 | 0 | 0 | 0 |
| 2 | 8 | 0 | 0 | 0 |
| 3 | 12 | 4 | 0 | 0 |
| 4 | 16 | 8 | 0 | 0 |
| 5 | 20 | 12 | 5 | 5 |
| 6 | 24 | 16 | 10 | 10 |
| 7 | 28 | 20 | 15 | 15 |
| 8 | 32 | 24 | 20 | 20 |
| 9 | 36 | 28 | 30 | 30 |
| 10 | 40 | 32 | 40 | 40 |
| 11 | 40 | 36 | 40 | 40 |
| 12 | 40 | 40 | 40 | 40 |

The main point of this discussion is that structuring the timing of assessments matters a great deal in RETs. The timing should be structured differently depending on (a) how long you think it takes for a change in a mediator to occur as a function of the treatment, (b) how long it takes for a change in the mediator to produce a change in the outcome, (c) the expected duration of the change in the mediator, (d) the expected duration of the change in outcome, (e) whether you think there are reciprocal causal dynamics between M and Y across time (i.e., M at time 1 influences Y at time 2 which influences M at time 3, which influences Y at time 4….) and (e) your research questions and goals, e.g., are you interested in documenting "curves of change" or do you only care whether change persists at substantively compelling time points. To effectively understand causal relationships, we need to think as much about the X-to-Y interval as the X-to-Y causal structure.

## CONCLUDING COMMENTS

As you approach RET design, there are decisions you need to make that focus on measurement. These include the choice of existing measures to use, the construction of measures unique to your particular RET, the development of strategies to minimize random measurement error and systematic measurement error, whether to collect baseline assessments of mediators and outcomes, and the frequency and timing of assessments more generally. The present chapter makes salient some of the issues you need to think about when making these decisions.