

Chapter 2

Measurement as Communication

In order to set the stage for understanding communimetrics as a theory of measurement it is important to set the context based on current theories of measurement, of which there are two primary conceptual models—psychometric theories and clinimetric theories. Psychometric theory has two competing approaches within its general framework—classical test theory and item response theory (IRT). The following describes the basic tenets of each of these approaches.

A Brief Review of Current Theories of Measurement

Classical Test Theory

The original psychometric theory is called classical test theory (Nunally, 1976). In this theory, one conceptualizes the universe/population of all possible questions relevant to the measurement of a single construct. Measurement involves the sampling from this population of attributes of the construct and aggregating these sampled attributes to estimate the level of the construct. Picture the population of all possible questions you could ask to measure happiness. Potential questions might involve mood state (e.g., euphoria, blissfulness, sadness) or enjoyment of activities or any number of other aspects of the construct. Classical test theory posits that if you can randomly sample from this population of all possible questions, it is possible to create a valid measure of the construct given a sufficient, representative sample of questions. In order to do a good job of measurement development according to this theory, it is first necessary to define the population of possible items and then adequately sample from it in order to achieve a representative sample. Thus, the usual first step of creating a measure from classical test theory would be to brainstorm as many possible items that might measure some important component of the construct.

Of course it is practically impossible to actually define the population of all possible questions for a construct. Similarly, it is difficult to know a priori whether a particular question actually belongs in the target population or is a better representative

of a different construct. Therefore, classical test theory goes further than just randomly sampling items. It creates a set of statistical strategies that ensure you are sampling items from roughly the same population but not ones that are so overlapping in how people respond to them that they are redundant. Measurement developers using this approach engage in a set of strategies generally referred to as *item analysis* in order to ensure a “Goldilocks” criteria of similar enough, but not too similar, items for all of the items included in a measure.

Item analysis involves the study of the intercorrelations among sets of items and correlations between individual items and total scores. The degree to which items in a set correlate with each other is used as evidence of whether the items are actually measuring the same thing. A correlation of 0.05 between two items suggests they are measuring two different constructs and therefore are not members of the same population. A correlation of 0.95 between two items suggests they are measuring exceptionally overlapping things and are essentially identical from a statistical perspective. A correlation of 0.30 to 0.60 is desirable according to classical test theory (Nunally, 1976). In other words, the items are measuring similar things, but are not too redundant. Negative correlations work the same way. A high negative correlation would be taken as evidence of information redundancy, but in the opposite direction on the construct. Factor analysis can be used to identify the underlying structure of relationships among sampled items. Factor analysis, which is the statistical cornerstone of classical test theory, takes the correlation matrix and places some formal statistical rules on the size of correlations needed to support the claim that the items share a common construct or population (Eysenck, 1971).

Factor analysis as applied to measurement development is essentially an inductive process (putting aside for the moment confirmatory factor analysis). After a set of items are generated it is used to determine statistically whether there is sufficient evidence to suggest that multiple items are measuring the same construct. Many test developers have used the results of factor analyses not only to identify items to include on a test but even to identify and name dimensions of a measure for purposes of scoring and interpretation.

Reliability and validity considerations under classical test theory come directly from the theory behind the choice of items. Although test-retest reliability and inter-rater reliability are important, classical test theory is also used to evaluate measures of transient, subjective states that are neither observable nor necessarily stable. As such, internal consistency reliability has become a commonly used and accepted indicator of reliability. Internal consistency reliability measures the degree to which items of a test correlate with each other—the higher the correlation, the higher the reliability. Generally, the more items that are on a test, the higher the internal consistency reliability will be (Nunally, 1976). Thus, classical test theory, particularly when internal reliability is the only available measure of reliability, implicitly encourages the selection of tests with more items.

Given the care used to measure one construct with multiple items, classical test theory also emphasizes measuring fewer constructs. A good measure, according

to this theory, is not multifaceted. Rather, a good measure has a stable factor structure with a discrete, probably low, number of factors. Each of those factors should have discrete validity with other measures of similar (or opposite) constructs. Classical test theory is the measurement foundation behind Eysenck's (1971) classic work on the dimensions of personality and even Leary's (1956) work on the circumplex structure of personality.

Classical test theory generally views *face validity* as the least important of all forms of validity. The most important evidence of validity is captured within the broad area of information that is required to demonstrate *construct validity*. Thus, items do not necessarily have to appear consistent with what they are thought to measure so long as there is statistical evidence that these items really are measuring the construct in question. In fact, for some measures, items that might appear irrelevant can contribute to good measures. There are multiple examples of such items in classically constructed measures, such as the 338 item Minnesota Multiphasic Personality Inventory (MMPI; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). The classic example from the original version of the MMPI was an item involving whether you would sometimes cross the street to avoid running into someone you know. Most people say yes. People who are paranoid as assessed by diagnostic interview are more likely to say no.

In general, classical test theory implicitly defines as reliable and valid longer measures of single (or few) dimensions. Measures with too few items on each dimension or too many dimensions, particularly if they are not orthogonal (i.e., correlated) will be seen as less desirable within this framework. One of the most common reliability criteria in classical test theory is Cronbach's alpha, which is an indicator of the degree to which items on the scale correlate with one another (Cronbach, 1951). The equation for α is:

$$\alpha = \frac{N}{N-1} \left(1 - \frac{\sum_{i=1}^N \sigma_{y_i}^2}{\sigma_x^2} \right)$$

where N is the number of components (items or tests), σ_x^2 is the variance of the observed total test scores, and $\sigma_{y_i}^2$ is the variance of component i .

Cronbach's alpha is biased by the number of items on the scale. The fewer the number of items; the lower the magnitude of the alpha statistic. While corrections exist to this bias (e.g., Allen & Yen, 2002), it remains the case that classical test theory values multiple items to measure single constructs. The history of suspicion of single-item measures rests in classical test theory. Because of the nature of error of measurement, it is certainly true that a linear combination of items is more reliable than an individual item (Nunnally, 1976). That, however, does not imply that an individual item cannot be reliable. But you cannot perform an item analysis or factor analysis on a single-item scale, rendering the primary methods of classical test theory useless for these applications. It is in the humanity of scientists to not trust what they cannot study within the range of their methods. If you have a hammer, you tend to look for nails.

Item Response Theory

Item response theory (IRT) approaches the measurement problem in a manner that is quite different from classical test theory. IRT posits the existence of a latent continuum that is the measurable aspect of a particular construct. This continuum can be considered to extend over levels of difficulty. The goal of measurement (at least in human service enterprises) is to reliably and accurately locate a particular person (or perhaps a grouping of people, such as a family) on this continuum relative to all other possible individuals (or comparable groupings). A good measure from this perspective is one that is sensitive across the entire continuum. Therefore, the measure must have the ability to distinguish different people reliably all along the continuum.

The statistical approach to IRT can be quite varied and complex, depending on the number of parameters used to define the continuum. However, in all cases the goal is to identify a set of items that allows for the precise measurement of an individual on the latent continuum or trait. The use of a single parameter model, such as item complexity as used in Rasch scaling (Rasch, 1960/1980), is perhaps the most common approach to measure development and can serve as a constructive example of the implications of IRT for test construction.

In Rasch models, the probability of endorsing an item (if it is discrete) or the population probability of ratings at each level (if it is continuous, such as a Likert scale), is used to define where on the continuum the item is most useful to distinguish respondents (i.e., the separation reliability). The relationship of the item's pattern of difficulty to the rest of the items defines the degree to which the item lies along the latent continuum (i.e., the fit statistic). A good test from a Rasch perspective is one that has items that separate reliability, cover the range of the continuum, and lie along that continuum (Wright & Stone, 1979). Thus, Rasch modeling also consider measures with multiple items on a single dimension to be more reliable and valid. Although there are techniques within IRT that allow you to identify the fewest possible items while maintaining adequate psychometric properties, it remains a significant criterion that the included items cover the latent continuum in terms of varying difficulty (i.e., likelihood of endorsement).

IRT approaches validity from a perspective similar to classical test theory. Statistical relationships between and among items trump other methods for evaluating measures. It is possible that prediction (or statistical criterion) validity is more highly valued in IRT as compared with classical test theory; however, construct validity is again the single most important validity criterion. Face validity is nearly irrelevant as the statistical methods guide the test developer to a greater extent than the perceived experience of the respondent. When items do not fit (the item fit statistic is above 1.6 or so), cognitive testing in which respondents are interviewed while they complete the measure can be used to better understand how people are interpreting the item wording is often recommended.

Clinimetrics

Due to their length and the time and procedural separation between rating, scoring, and interpreting, psychometric measures were not widely accepted in medicine. Although current information technology eliminates many of these challenges, easily accessible, fast computers were not available in the decades in which clinimetrics developed as a theory of measurement. Thus, psychometric tools were seen as burdensome in medical settings. Further, the lack of concern regarding face validity in these approaches sometimes led practicing clinicians to look at the questions and be somewhat skeptical about the measurement process. In an effort to create clinically relevant measurement procedures, physicians and other health researchers have utilized a theoretical approach referred to as clinimetrics (Feinstein, 1987). The stated goal of clinimetrics is to convert “intangible clinical phenomenon into formal specified measurement” (p. 125; Apgar, 1966). Virginia Apgar is generally credited with developing the first measure from this perspective (Apgar). First introduced in 1953, the Apgar is routinely utilized as a health status measure at birth. Clinimetric tools are now quite common in medicine (e.g., Bloem, Beckley, van Hilten, & Roos, 1998; Gates 2000; Hoff, van Hilten, & Roos, 1999; Stone et al., 2001).

Perhaps more than anyone, Feinstein (1999) advocated clinimetrics as a specific theory of measurement. He enumerated six core principles to clinimetrics in comparison with psychometrics:

1. Selection of items is based on clinical rather than statistical criteria.
2. No weighting factors are needed; scoring is simple and readily interpretable.
3. Variables are selected to be heterogeneous rather than homogeneous.
4. The measure must be easy for clinicians to use.
5. Face validity is required.
6. Subjective states are not measured as they are severely limited in terms of source of observation.

Current applications of clinimetrics have some notable limitations (Marx, Bombardier, Hogg-Johnson, & Wright, 2000; Zyzanski & Perloff, 1999). Many clinimetric scales consist of a single item. Attempts to describe complex phenomena with a single item general fail to communicate complexity. For example, a Childhood Global Assessment Scale (Endicott, Spitzer, Fleiss, & Cohen, 1976), which ranges from 0 to 100, can provide a general sense of how the child is doing, but cannot capture individual dimensions of functioning that are useful to clinicians. In addition, single-item measures are not particularly sensitive to change. For these reasons, Zyzanski et al. (1999) and others (e.g., Fava & Belaise, 2005) have called for an integration of clinimetric and psychometric approaches to measurement. Marx, Bombardier, Hogg-Johnson, and Wright (1999) have demonstrated that the two theories can be complementary. Not everyone agrees. Streiner (2003) has gone so far as to argue that clinimetrics is actually a subset of psychometrics, and that for both scientific and communication reasons the word *clinimetric* should be eliminated.

Of course, the distinguishing features described by Feinstein in the process of defining clinimetrics has resulted in most applications involving single items, although in his book Feinstein does not limit clinimetric measures to single items. A single marker of disease severity is the most common type of measure using this framework. Table 2.1 provides an example of a clinimetric measure that is commonly used, the New York Heart Association rating for heart disease. Notice that it assumes the presence of heart disease even at the lowest level. Thus, the concept of normal or normative is either moot or only relevant within the population of people with heart disease.

One of the intriguing characteristics of this measurement approach is that although a key principle of the measurement theory is to keep scoring simple, with no weighting, the actual design of the anchor points creates implicit (and sometimes explicit) weighting of input criteria prior to the clinician's judgment about the rating. Thus, while scoring is simplified, ratings are more complicated. This is how the clinimetric approach differs from psychometrics in the selection of items that reflect clinical judgment. Psychometric theory would emphasize avoiding "double-barreled" items with complex, multiple meanings because they do not tend to scale as well. No such restrictions guide the creation of items in clinimetrics. In fact, if multiple constructs combine to create a continuum of severity, it is desirable to embed all relevant constructs into the anchored definitions of the rating.

The challenge of clinimetric measures is that their use is maximized at the individual patient level, but as you move to higher levels of aggregation, the utility of the measurement approach diminishes. It is hard to monitor and explain transformational processes with clinimetric measures alone. They tend to serve as excellent indicators for defining differences in patient populations but have limited value for outcomes.

Table 2.1 An example clinimetric measure

Class I	Patients with cardiac disease but without resulting limitations of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnea, or anginal pain
Class II	Patients with cardiac disease resulting in slight limitations of physical activity. They are comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea, or anginal pain
Class III	Patients with cardiac disease resulting in marked limitation of physical activity. They are comfortable at rest. Less than ordinary physical activity causes fatigue, palpitation, dyspnea, or anginal pain
Class IV	Patients with cardiac disease resulting in inability to carry on any physical Activity without discomfort. Symptoms of cardiac insufficiency or of the anginal syndrome may be present even at rest. If any physical activity is undertaken, discomfort is increased

The New York Heart Association functional classification.

From The Criteria Committee of the New York Heart Association, Inc. *Diseases of the Heart and Blood Vessels: Nomenclature and Criteria for Diagnosis*. 6th ed. Boston: Little, Brown, 1964.

Comparison of Communimetrics to Psychometrics and Clinimetrics

Measurement can be conceptualized as having at least two distinct phases, input and output. Each aspect requires that decisions be made regarding how that aspect is conceptualized and managed in the measurement process. The input phase involves all the operations of observation and scoring. The input aspects of measurement involve decisions about what to observe, under what conditions to observe, and using what information source for the observation. The output phase involves all the operations involved in using and sharing the measured values. The output process involves decisions about how information is scaled, combined and reported. As demonstrated in Chap. 1 (Fig. 1.2), considering measurement for its communication value shifts the focus from the input side of the measurement process to the output side of the same process. Blanton and Jaccard (2006) have argued that many psychometric measures are arbitrary because the numeric values generated have no grounding in reality. The goal in emphasizing the output applications to the measurement process is to help ensure that the measure is not arbitrary and values generated from a measurement process will be accepted for use within human service enterprises. Therefore, in order to maximize output value, decisions regarding input choices are guided by applications of the measure on the output side.

All measurement theories have to make decisions regarding how the input and output processes interact and inform decisions about each other. Communimetrics differs as a measurement theory from psychometric theories with regard to input, output, and their interaction. Communimetric theory differs from clinimetrics primarily in terms of output decisions.

Input Processes in Measurement

In designing a measure, the first decision that must be made is what aspect of the human condition is to be measured. There are a potentially infinite number of things about people that might be measured; they vary from large, rather global constructs (e.g., job skills, depression) to rather molecular behaviors (e.g., eye blinks, simple arithmetic skills).

The first decision about what construct to measure often has clear and immediate implications as to many of the significant decisions regarding the operations required to measure. For example, if you want to assess eye blink frequency it requires a process that involves prospective, external observation since self-monitoring eye blinking behavior likely influences its frequency. And nobody remembers whether they (or someone else) blinked after even just a short period of time, so recall methods of observation are not feasible. Sadness, on the other hand, is something only available through introspection on the part of the target person.

The second step of the measurement process is in regard to the procedure or operations to be used. With measuring humans, there are three basic choices: self-observation, other observation, and instrument observation. Self-report is preferred when the construct is an internal state that only the individual has access to observe. Knowledge is the best example. The only way one person knows what another person knows is by asking (or testing) them with regard to their knowledge. Other observation is generally used when either self-report is not feasible or cannot be trusted to be accurate. Instrumentation is often seen as the most scientific of all measurement approaches, but it requires a construct for which an observation instrument has been developed. Thus, applications have historically been limited to very specific constructs in which measurement has a clear value and the operation can be automated in some manner (e.g., temperature, weight). Our information culture and the microsizing of computers has created a dramatic increase in instrument measurement in stores and other venues. For example, phenomenon such as Web surfing can be measured using instrumentation (e.g., how many hits on a site). In health care, instrument measurement of humans is widespread, with examples ranging from blood pressure (which still has another observation component in many cases) to positron emission tomography.

The conditions under which a measure is applied is generally the third decision of the measurement input process. In physics and chemistry, there are often powerful assumptions regarding the conditions of measurement (e.g., standard temperature, no gravity). In the measurement of humans, many have tried to be equally rigorous, but the realities of the processes necessary to obtain information often compromise rigid rules regarding the conditions of measurement. Standardized tests in the education field are good examples of attempts to enforce routine conditions on the measurement process. People administering standardized tests have a set of rules and time frames that they must follow in order to ensure comparability in conditions across different measurements. The administration of these standard educational tests via computer has made this type of procedural control more efficient. Other examples include measurement at intake into a clinic or program followed by repeated readministration of the measure at fixed intervals (e.g., every 3 months, at discharge).

The final consideration, although not necessarily the last, in the input process of measurement is the satisfaction of whomever must complete the measure regarding its ease of use, suitability, etc. These characteristics are included what has been called “face validity,” i.e., the measure is valid on the face of it. Existing measurement theories weigh the importance of this consideration differentially.

The Output Process in Measurement

There are also a number of important decisions that the developer of a measure must confront regarding the output of the measurement process. Scaling is the first decision of the output process. What is the unit of measure? How many levels

exist in each item? What is the proposed relationship among those levels? Historically, scaling has been categorized as categorical, ordinal, interval, and ratio. Categorical scales describe things in discrete groups that have no hierarchy. Ordinal scales have a hierarchy, but the differences among levels in this hierarchy are not comparable. Interval scales order levels in a fashion that allows for an assumption of equal differences among levels. Ratio scales have an absolute zero. Ratio scales are thought to be rare for constructs of interest within human service enterprises.

Within these standard categorizations of scales, there is further differentiation. For example, Gutman scales are a form of an ordinal scale in which each new response, when endorsed, requires that all previous responses have been endorsed as well. The classic example of a Gutman scale measures racial discrimination using something like the following questions:

1. Would you be OK with a person of [insert race] living in your town or city?
2. Would you be OK with a person of [insert race] living in your neighborhood?
3. Would you be OK with a person of [insert race] living on your street?
4. Would you be OK with a person of [insert race] living next door to you?
5. Would you be OK with a person of [insert race] living in your house?

If you said yes to question 4, then you would obviously have said yes to questions 1 through 3, but may not necessarily had said yes to item 5.

There are other types of ordinal scales: frequencies, class rank, power rankings of sports teams, etc. are all ordinal in their scaling properties. These scales are easy to use and understand, but are limited in statistical applications as they are less easily combined, and often you can't use parametric statistics with them. Frequency scales (i.e., raw counts) are often mistakenly thought of as interval scales.

Within interval scales, the most common type is the Likert Scale (Anastasi, 1968; Nunally, 1976). Generally, Likert scales assess either agreement or intensity, which is sometimes used to convert frequency measurement into an interval scale.

Agree completely	Never
Agree somewhat	Rarely
Neither agree or disagree	Sometimes
Disagree somewhat	Often
Disagree completely	Always

There are other types of ratings besides Likert Scales, such as visual analogs and 0 to 100 ratings, which are thought to normally function as interval scales. By and large, it is accepted while these types of scales can be assumed to function as interval scales; however, it is a good idea to test any scale as it functions to ensure this important characteristic.

Once scaling decisions have been made, the next decision about measurement output is how to combine items. It is on this decision that the various measurement theories diverge most dramatically, so this topic is discussed in greater detail within each of the major theories. However, overall, decisions have to be made about

which items can be added together and in what fashion in order to create scores that are used as the primary outputs of the measurement process. With psychometric theories, single items are thought to not make reliable measures; therefore, some combination is always required. With clinimetric and communimetric measures, single items can make reliable measures, so the decision making in this regard is different. Generally, with clinimetrics only single-item scales are used, so decisions about combinations are moot.

As discussed, Blanton and Jaccard (2006) have described the problem of arbitrary metrics in psychological measures. These authors define arbitrary as “when it is not known where a given score locates an individual on the underlying psychological dimension or how a one-unit change on the observed score reflects the magnitude of change on the underlying dimension” (p. 28). In other words, scores on many measures do not have independent relationships with the degree (e.g., severity, difficulty, intensity) of the construct purported to be measures: A 15 on the Beck Depression Inventory is not tied directly to a degree of depression itself. These authors press for tying levels of measures to real-world, meaningful events as necessary to making measures reliable, valid, and not arbitrary.

The third decision about output processes in measurement is how the scores are presented, displayed, or otherwise communicated. Some measures use normative transformations, such as T scores (mean of 50 with standard deviation of 10). Other measures use total scores or profiles of scores. Some strategy is necessary to ensure that the scores have meaning to those who intend to use them or that individuals who utilize the scores can be educated to interpret them appropriately. Psychometric measures must develop some strategy to ensure meaningfulness. Both clinimetric and communimetric measures are designed for immediate meaning, at least at the level of a single item.

The final characteristic of the output process is whether the use of the measure has any impact on the people who receive the information. That is, does the information taken from the measure within the human services setting actually result in a change of behavior or performance. Although a validity consideration for all approaches to measurement, this measure utility or impact is not a primary consideration in the design of measures developed out of psychometric theories. This is a central output consideration for communimetric measures.

The Relationship of Input and Output Process

In the measurement development process, depending on the theory, complex relationships exist between input and output processes. In particular, in psychometric theories results of statistical analysis of item performance from the output processes have direct implications for the design of input processes. Both classical test theory and IRT have specific, well-defined characteristics for a well functioning item. Those characteristics involve the statistical performance of the item relative to other items, and sometimes, an external criterion. An item that does not perform statistically

in a manner consistent with the theory, then it should be removed from the input process. Thus, measurement development is generally defined by findings from the input side of the measurement process. For example, in classical test theory there is an optimal correlation among items and between an item and the total scale. Two items that have a high correlation are considered redundant from an information perspective and one is generally removed. An item with too low of a correlation to the total score is thought to be measuring a different construct and therefore, it is eliminated. If a subset of items can predict the total score of a larger set of items, then a shorter version of the measure (i.e., the subset) is recommended.

Item-response theory uses item fit statistics to determine whether the item is performing as expected. In other words, is the probability of endorsing different levels of an item (i.e., termed item difficulty in Rasch modeling) consistent with that item residing on the underlying continuum shared by other items in the sample. In addition, item-response theory looks for items that spread across likelihood of endorsement (i.e., item difficulty) to ensure that items are included that are sensitive at different levels of the construct. Failure on these input analyses leads to changes in the input process. Thus, if too many items are “easy” (i.e., frequently endorsed), it will result in an insensitive measure across the latent trait and the measure will have a ceiling effect. More difficult items must be identified. Similarly, if there are too many “difficult” items (i.e., rarely endorsed), then the measure has a floor effect and easier items must be added. A shorter version from an IRT is a scale that has a uniform distribution of items across levels of difficulty while maintaining good item fit statistics on the continuum.

Input and Output Processes in Human Service Enterprises

Human service enterprise settings have very different priorities than research settings. Accommodating these technical and contextual requirements requires a broad scope for models of measurement. The measurement model must include guidelines for utility in operations as well as reliability and validity. It is not necessarily true (as psychometric measurement theory assumes) that if you develop a good measure from an item analysis, it will result in a useful measure within a human service enterprise. Measures intended for the assessment of transformational offerings should be easy to use and brief. Their output should be clear, unambiguous, relevant, easy to translate into intervention planning recommendations, and accessible to providers, consumers, and policy makers. Classical test theory, IRT, and clinimetrics are not able to fully inform the development of measures meeting these requirements in human service enterprise applications.

As discussed in Chap. 1, in entities that provide help for people, the primary role of measurement is to communicate. That communication is first between the consumer and the provider (e.g., what do we need to work on together?), but the communication can be far more complex than that. Often, human services are paid for by the government or other entities. Thus, third parties (the consumer is the first

and the provider is the second party in the transaction) are involved in payment for these interventions. Communication between providers and payor also is important. In addition, fourth parties are involved, including accreditation and other entities that monitor human service enterprises. In some situations, even fifth, sixth, and seventh parties are involved because the nature of the intervention requires the participation of multiple system partners. For example, in the child-serving system, it is not unheard of that child welfare, juvenile justice, mental health, and educational representatives are involved with the same youth. Communication with each of these parties is important to the work. That communication should focus on the nature of the work—the *human* in the human service enterprise. Measurement as communication is different in some important ways than other forms of measurement.

Principles of Communimetrics

Considering the communication value of a measure from the beginning changes some core principles of measurement design. This is particularly true when a constitutive view of communication is taken in which communication is viewed as the creation of a shared meaning. There are six key principles of measurement as communication—communimetrics:

1. Each item has implications for differential action.
2. Levels of each item are immediately translatable into action.
3. Measurement must remove the context, including:
 - a. Services already in place
 - b. Culture
 - c. Development
4. Measurement is descriptive and minimizes cause–effect assumptions.
5. Observation windows can be trumped by the action levels.
6. Information integration

Each Item Has Implications for Differential Action

Like clinimetric measures, communimetric tools are designed so that they can operate at the item level. As described, clinimetric measures have proved false the psychometric theory position that only multiple item scales are reliable by demonstrating the feasibility and utility of single-item scales in medical settings. Communimetrics also emphasizes the use of single items, but also encourages multiple item approaches to allow comprehensive assessments of multiple constructs to facilitate decision making and outcome monitoring.

Given the action orientation of communimetric tools, items are included in a measure if they have a potentially meaningful relationship to what happens next

in the human service enterprise. In other words, the assessment is a planning process for any interventions that follows; items exist to inform choices among possible interventions or approaches. An item that is irrelevant to the planning process should not be included.

Levels of Items Translate Immediately to Action

A unique requirement of a communimetric measure is that the levels of measurement on each item should translate into action. In other words, the individual items are selected to guide decision making. The levels of these items should further guide decision making by indicating what level of service effort is required. A standard four-point communimetric scale might look like the following:

- 0 No evidence, no need for action
- 1 Watching waiting/prevention or keeping an eye on something
- 2 Action is needed
- 3 Immediate or intensive action is needed

Thus the design of the levels of an item on which ratings are made should immediately communicate the meaning of the item from a planning perspective. Here would be an example of a communimetric scale for a strategic planning process:

- 0 Not relevant
- 1 Parking lot
- 2 Issue to be addressed
- 3 Priority issue

An issue that is classified as not relevant can be dropped from the discussion. A “parking lot” issue is something that isn’t immediately important, but should be returned to at a more appropriate time. Items rated a 2 or 3 should be addressed in the strategic plan with those being rated a 3 taking priority.

Strength-based planning has increasingly become a best practice in child serving systems (Healy, 2005). The following is an example of a communimetric scale from the Child and Adolescent Needs and Strengths for strengths measurement:

- 0 Centerpiece strength
- 1 Useful strengths
- 2 Identified strength
- 3 No strengths identified

In this model, a centerpiece strength can be used as the focus of a strength-based plan. For example, if a child is removed from his or her parents due to abuse or neglect, but grandparents are available who are willing to take the child into their home, that is a centerpiece family strength. A useful strength is something that can be included in a strength-based plan but cannot serve as a centerpiece (e.g., knitting when stressed, enjoying singing in a choir, youth soccer for an 8 year old). An identified

strength gives you a window into where a strength could be built (e.g., a particular vocational interest in the absence of any knowledge or skills), and if no strength is identified that would preclude its inclusion in a strength-based plan. Thus, using this communimetric strength scale, strengths rated a 0 or 1 could be included in strength-based planning and those rated 2 or 3 might become the focus of strength identification and building efforts.

The action orientation of a communimetric tool is one of its greatest strengths. It eliminates the arbitrariness of a Likert scale as there is a clear link between the level of the measure and the external world. It makes the link between assessment and intervention planning transparent in support of supervision and other forms of accountability. It facilitates a full understanding of when interventions are no longer necessary, although meaningful applications for outcomes management. The levels of the items communicate between assessor and various parties who might be involved in providing transformational experiences based on the assessment findings. People who are assessed often report that this is the aspect that they most appreciate because it provides them with a framework for the work they have ahead of them if they wish to change aspects of their lives. However, the action orientation is not without controversy.

By establishing a clear, visible link between assessment processes and interventions opportunities for accountability are dramatically enhanced. I was doing training in Florida on a mental health version of the CANS and I presented the basic action levels: no evidence, watchful waiting/prevention, action, and immediate/intensive action. Someone came up to me at the break and said, "Well John, you realize this means we have to do something." They were quite distressed at thinking that once the tool had been applied it became clear to youth and families that something had to be done. I was struck by the irony of this concern. I answered, "That's exactly what it means. If you rate an item 2 or 3, then something has to be done." Isn't that the point of assessment after all—to figure out what needs to be addressed?

In New Jersey, I completed training and e-mailed people who had passed and not passed the certification test demonstrating reliability. One particular person was not reliable, and I e-mailed her with the news and feedback on what she missed. Essentially she had consistently underestimated needs of the youth in the test vignette. Reacting to this feedback, she replied that underestimating needs was just how they worked at her office. She stated that they really didn't have any options available to serve children and youth and had just found that they were better off pretending that treatment needs just didn't exist. Of course, this is missing the point of the framework for these types of tools. It isn't about pretending everything is OK. The process should be about identifying needs and if you can't meet them, then you have succeeded in identifying an unmet need. Documenting unmet (or unmeetable) needs becomes important information for improving the human service enterprise in the future.

This concern about action continues to be a sticking point for some people as they attempt to implement communimetric tools in human service enterprises. But just because you identify an item that requires action, it does not mean that a specific

action should automatically follow. Because of the fifth principle of communimetric measurement, “It is about the what, not the why,” there is no need to define precisely what must be done. In fact, creating hypotheses about the why (i.e., cause-and-effect relationship leading to the identified need), is the nature of transformational interventions. It is one thing to identify whether an entrepreneur has human resource management skills; it is a different thing to figure out how to help that specific person develop his or her skills. Or, it is one thing to say that a child or youth is misbehaving at school, it is a different process to determine why. The specific intervention is based on a hypothesis of the why. In the communimetric measurement model, assessment is describing the target of the intervention. The choice of interventions is often based on a hypothesis about a potential cause of the target.

An additional concern that is sometimes raised about action levels is that they are somehow circular. In other words, by defining the ratings based on actions to follow, the assessment is no longer independent of these actions. I would argue that it is true that the ratings are not independent, but that this interdependence is a good thing and not a problem within the context of the human service enterprise. Here is where the business context is different than a scientific perspective that might require that any measurement is independent of all others. It is quite valuable to understand how assessors are conceptualizing needs and strengths from the perspective of the enterprise. The action levels make this possible. A constitutive form of communication in which meaning is made among parties in the transaction through a consensus on the relationship of the level of need to the level of intervention is a major benefit of this approach.

Considering Context

A second unique feature of a communimetric approach is that the person(s) completing the measure is required to consider the larger context in which the measurement is occurring to prevent undue influence of contextual factors on the description of the person or entity under consideration. This characteristic is radically different than traditional scientific measurement, which attempts to control contextual factors methodologically rather than conceptually. Physics measures in a vacuum. Chemistry measures at a set temperature and barometric pressure. Such methodological control is not possible in human service settings. The following are some contextual considerations that might influence the process of establishing action levels.

Services in Place

The purpose of measurement in service delivery is to determine what actions must be taken. If actions are already being taken, that changes the context of the measurement process. That is, you are measuring things that are the targets of the enterprise.

If you are providing business incubation you are measuring factors related to entrepreneurial success. If you are providing health care, you are measuring things related to health status, level of functioning and well-being. If you are providing vocational services, you are measuring things related to job readiness.

If you are in the middle of providing interventions in support of improving targets of the enterprise, then it falls to reason that you would expect change in these targets as an outcome of these activities. That's what transformative offerings are all about. However, many such interventions may work only while they are active. For example, a person may perform adequately at work only when a job coach is present. Remove the job coach and performance deteriorates. Or, a person with a severe mental illness may only be symptom free when they take medication as prescribed. The intervention meets the need, but does not resolve it.

In order to understand the need for ongoing interventions, those that must remain in place to secure success are different from those interventions that have accomplished their objectives and can be ended. Traditional measurement approaches do not make this distinction. They describe the status of the person *regardless* of the service context. A person performing well at work with a job coach is no different than a person performing well with one. A person who is not symptomatic on medication is no different from the person who no longer needs to take his or her medication at all. In traditional measurement, interpretation of the meaning of the measure requires one to consider the service context after the measurement has been completed, as part of the analytic work. This ad hoc interpretation of contextual factors creates all sorts of problems with communication.

Consider the following example. Residential treatment is a common intervention for children and adolescents with severe or complex needs. This form of treatment involves placing the youth into a therapeutic living situation where he or she might stay for treatment from 30 days to several years. The treatment often works and youth get better during the episode of care. The youth then is returned to home or back to the community in a foster home or perhaps even an independent living environment. I have often heard it reported that parents and community providers experience the reported status of children and youth using standard measures as misleading, saying something to the effect of, "The residential provider says that the youth was doing fine, but as soon as they got back home everything began to fall apart again." This miscommunication occurs because the residential provider is describing how the child is doing *in their setting*, which has all sorts of therapeutic components and behavioral controls.

A communimetric measure requires that the communicator represent the child or youth's status independent of the service setting. So instead of describing how the youth is doing in residential treatment, the communicator is instructed to assess how that individual would be expected to function without all the supports inherent in the residential treatment center. Thus, in order to effectively communicate using a structured measure, the residential rater has to distinguish setting effects (improvements that come from living in a structured setting) vs. treatment effects (improvements that transcend the structure setting that will generalize to other environments).

Culture

Over the past several decades, social scientists and service delivery systems have become sensitized to the complexity of addressing cultural issues effectively in practice. There is overwhelming evidence that racial and ethnic disparities exist in many human service systems in the United States (e.g., Smedly, Stith, & Nelson, 2003). Addressing cultural issues is complex, however. There are actually three different strategies that are necessary to effectively address cultural issues in human service enterprises.

Treating Different People Differently

The primary focus of cultural-based initiatives in service delivery has been an effort to teach service delivery systems to treat different people differently. This skill set is often referred to as cultural sensitivity. Some people use the term *cultural competence*, but I would argue that this term is an oxymoron. The opposite of competence is incompetence, and anyone who goes around referring to others as “incompetent” is likely not sensitive to others. Thus, it seems preferable to choose to use the term *cultural sensitivity* to describe the skill of adjusting the human service enterprises to account for relevant variations in culture.

An obvious example of cultural sensitivity comes from mental health. If a person is an active member of a Pentecostal church, he or she may talk in tongues during religious services. This behavior does not make the person psychotic. The same vocalization patterns exhibited by someone walking down the street or being interviewed in an emergency department might be seen as compelling evidence of a symptom of psychosis.

I recently received an e-mail from a colleague about a case of a young woman in Oregon. Her grandfather had died and she had been close to him. He was the *pater familia* and a source of significant support for this adolescent girl. Following his death, she reported talking to her grandfather and her psychiatrist diagnosed her as psychotic and sought to start her on antipsychotic medication due to the presence of delusional thinking. Here’s the problem with this situation. The young lady was Native American. In her culture, speaking to dead ancestors is a traditional way of describing the continuing influence of a lost loved one, just like a religious person may refer to speaking to God.

Traditional measurement approaches try to measure completely independent of cultural influences. So ratings assessing delusions or hallucinations might be defined in a way that a Native American or devoted religious person might respond to in the affirmative. This approach forces cultural sensitivity to occur after the measurement process is complete as a part of interpreting the numbers. While measuring independently from cultural influences is reasonable, and perhaps even optimal, for scientific investigation, it places enormous challenges on information collected in service delivery settings. Without detailed knowledge about the cultures of individuals involved in transactions, it is exceptionally difficult to recreate

potential influences with aggregated data. Thus, it is difficult to know whether disparities exist in assessment or interventions.

Some psychometric measures, such as the Cardiff Anomalous Perception Scale (Bell, Halligan, & Ellis, 2006) address this issue by making all items ipsative; that is, based on the individual's open experience set, e.g., "Do you ever think that food or drink tastes much stronger than it normally would?" This represents a reasonable alternative to considering cultural factors prior to establishing the level of an item. In this model, you allow the individual to correct for cultural influences prior to answering the questions. However, such instructions are never a part of psychometric measurement.

In the traditional model cultural factors become variables that you have to control in order to interpret information. Large sample sizes and/or sophisticated multivariate statistical techniques are required to ensure that standards of cultural sensitivity are met. You can't really even report the frequency with which people report "delusional thinking" without first factoring in the degree to which some cultural factors might influence this rate. Placing this level of interpretative responsibility at the analytical level (following scoring) is inconsistent with effective communication because the raw data collected might be misleading unless specific analytical procedures are first applied. At the individual person level, the implications are more complicated and you are left trying to decide whether or not the information is meaningful. A clearly interpretable rating that does not require scoring is the clearest form of communication.

The traditional alternative to understanding contextual variables analytically is to create different measures for different contexts. This is one of many reasons why so many different measures exist in the human service enterprise. However, the use of culturally specific measurement is limited if you want to be able to draw conclusions about human service enterprises in cross-cultural settings or if you ever want to understand the role of culture in the functioning of these enterprises.

Communimetric measures build the concept of cultural sensitivity directly into the measurement process. Before an action level is determined, culture must be considered. If something is a behavioral norm in an individual's culture, then it is not a need. Family involvement manifests itself in very different ways across ethnic and cultural groups. Consideration of these factors must occur before one could identify actionable family needs or strengths.

An exception to this rule exists when a specific culture has a behavioral norm that is outside the range of nonculture-based behavioral norms. Behaviors such as corporal punishment and female castration are examples of these types of behaviors; normative in some cultures, but widely unacceptable across cultures. For example, a parent beating his or her child would be described as physical abuse in the United States, Canada, and Europe regardless of the culture of the person for which that behavior is described.

Treating Different People the Same

Cultural sensitivity does not apply to all situations. There are situations in which we must learn to treat different people the same regardless of their cultural differences. Racial disparities in health care and employment are important examples of these

problems. For example, there is substantial evidence that in the United States, African Americans are more likely to be admitted to the hospital and receive lower-quality outpatient treatment than do Caucasians (Smedly et al., 2003). Nobody believes that race should explain the utilization of health care or employment rates and income levels. If a measurement process is to be useful from a cultural perspective in a delivery system, it should be able to be used to identify and address disparities.

Addressing Cultural Needs

The third way in which culture should be addressed within a delivery system is that sometimes specific culture-based needs can be identified. Once identified, the system should be able to address them. Language is an obvious one. If a person of a family member does not speak the primary language in a jurisdiction, then he or she needs help to ensure that effective communication can be accomplished. Without everyone in the process having a full voice, it is impossible to have a fully effective system. Other cultural needs might include access to rituals (e.g., food, holidays, music) or cultural identity and/or stress. Often, families that emigrate to the United States experience complex intergenerational stress in that the children in the family are sometimes more readily affected by U.S. cultural influences, creating tensions with parents.

Development

A third contextual factor can be development. We have dramatically different expectations with regard to behavior and performance based on age, both chronological and developmental. All 3 year olds have anger control problems, so this is irrelevant to any assessment of behavioral health. A 15 year old or a 30 year old who has the anger control skills of a 3 year old would represent a problem, however. We don't expect an infant to be able to toilet himself or herself. By around 2 years old, this becomes a societal expectation, and the failure of an older child to successfully toilet himself or herself is considered an actionable need.

Recreation functioning requires entirely different considerations based on age and development. Children do not engage in the same recreational activities as adolescents. Young adults do not engage in the same recreational activities as older adults. If you want to understand recreational functioning in terms of the need for interventions, it is essential to do it within a developmentally appropriate framework.

Measurement is Descriptive

In the context of measurement in human services, causal relationships are complex and judgments with regard to cause-and-effect is subject to substantial error. For instance, in behavioral health, there is no known pathogen. Therefore, jumping

to a cause of any symptom or behavior is likely to be wrong. So, at least for the majority of items, communimetric tools tend to focus on describing actionable conditions rather than interpreting them within a causal framework.

In trainings, I often use the mantra, “It is about the what, not about the why.” In my experience this aspect of the communimetric measurement facilitates its use in constitutive communication. In many situations within human service enterprises, shame and blame come from the why. Stigma comes from the why. When you focus on the what—the description of what the needs are without initially trying to determine the cause of these needs—it serves as an engagement strategy. The fact that you are homeless is one thing. The reasons you are homeless are a different conversation.

Treatment interventions are almost invariably directed to the theory of why. So the nature of the intervention requires a hypothesis about the why to go along with the description provided in the assessment. This relationship between assessment and treatment allows you to use communimetric tools to pursue person-driven planning. In other words, the assessment process is used to reach a consensus about what is going on (i.e., constitutive communication). The individual or family generates hypotheses as to why these things are happening; then the professional brings in evidence-based approaches to address this proposed cause. If the first intervention doesn’t work, then a new hypothesis is generated.

Use of Time Frames (Windows of Observation)

All measures require a definition of the time frame over which an observation can occur. As a thinking tool, communimetrics has a different philosophy in this regard. Time windows for observations (e.g., 30 days, 24 hours, etc) are recommended, but they exist to remind people using these tools that ratings should be fresh; however, these ratings must be implemented with flexibility. At the end of the day, the role of a measurement process in the human services context in which communimetric tools are used is to establish actionable items. Thus, the action levels take precedence over the time frames. Time frames are only relevant as they inform action levels.

For example, in the Child and Adolescent Needs and Strengths (CANS, see Chap. 5), a 30-day time frame is used unless an item specifies otherwise. However, a rater can change his or her rating based on the specific situation. My favorite example of this procedure is an example of doing an assessment with a young adult who is in the hospital after a car accident. Let’s say for sake of illustration that the young man drank, drove, and crashed his car. As a result of the crash, he ended up hospitalized in a coma for 90 days. If you were charged with planning his treatment post-discharge from the hospital, would you argue that he has been “clean and sober” for 90 days? Of course not. He’s been in a coma. In fact, his substance use need would probably best be described knowing how he was doing prior to his accident, not during his hospital stay.

Information Integration

Communimetric measurement is an information integration strategy. Information integration refers to the process whereby multiple inputs are combined to generate a measurement. Therefore, communimetrics operates at a higher level of measurement than the direct application of instrumentation. A lab assay applies measurement processes to biological materials. A ruler applies its metric to an observed distance. The direct application of instrumentation to generate measurement is the foundation of science. However, when information is used in human service settings, it is often necessary to measure at what might be called the level of *executive function*. This type of measurement process requires the combination of multiple and potentially competing measurements or observations into a single measure. Psychometrics accomplishes information integration by asking multiple questions to the same source to measure a specific construct. That requirement can be limiting. For example, if a clinician is attempting to measure depression, self-reported symptoms are one input; however, observed mood, physical activity levels, and reports from significant others, are all relevant to that measurement process. The clinical judgment of whether or not depression is evident and to what degree is based on the integration of measurement from multiple sources. Any clinician will tell you that single-source measurement is inherently limited across a cohort of assessments.

In children's mental health, the Child Behavioral Checklist (CBCL; Achenbach, 1991) has versions for parents, teachers, therapists, and youth. The correlations among these versions are generally quite low. Accordingly, these findings demonstrate that working independently, different people describe the same youth differently. However, at some point everyone should come to agreement about what the youth needs and what should be done about it. The disagreement among the multiple sources only prepares you for how much work you will have to do to reach consensus. The consensus is necessary to actually intervene.

Similarly, if a business incubator is attempting to understand a start-up company's market potential, the inputs into that assessment are also multiple. It may include the novelty of the product, its cost, the existence of a known market, and so forth. Each of these factors, all relevant to market potential, require different measurement processes. However, the venture capitalist still must put all of those inputs together to make his or her judgment with regard to a new business's market potential.

Team Decision Making and Strategic Planning

Communimetrics is designed to operate at the level of the person overseeing the implementation of the interventions within a human service enterprise, e.g., the clinician or the venture capitalist. In fact, the design of the communimetric approach is uniquely suited for team decision-making measurement. Any strategic planning

process can be conceptualized as measurement: What do we have? What do we need? What should be done to move forward? These are all higher-order measurements. Teams convene to provide multiple inputs into these planning processes. The contribution of each team member can be conceptualized as a measurement input, and the output of the team similarly can be seen as a measurement. Communimetric measures function well as outputs of team measurement processes. Again, the team is generally engaged in constitutive communication, creating meaning.

Self-Report and Communimetrics

Measurement strategies that have the respondent directly answer questions on a survey are commonly called *self-report*. In many ways, self-report measurement is a field in and of itself, as the nuances of how you obtain accurate and useful information directly from target respondents has received much investigation. Self-report methods of measurement have a number of important advantages:

- They are direct. The target person is the one who responds to the questions or item prompts. There is no interpretive filter by an observing other.
- They provide a certain level of confidentiality; sometimes the illusion is even greater confidentiality than is actually the case.
- They are inexpensive. Generally the target person is not paid to complete the measure, so from the human service enterprise perspective, it is provided at almost no cost.

Getting information directly from the individuals you are seeking to measure makes a great deal of sense. Who knows you better than yourself? As long as the information sought is open to self-observation, then in theory at least, it is accessible to self-report measurement. And, things that are never available to observation (e.g., a feeling state, self-esteem) are only accessible via some form of self-report.

There is a body of research that suggests people are often more comfortable telling secrets to a computer administered survey than to a person-administered approach (Lyons, Howard, O'Mahoney, & Lish, 1997). This suggests that there is something about interacting with a form that is different than interacting with a person. The relational aspects of the presence of the other person might influence how we choose to present ourselves. Relationships can influence differential responses depending on the method of inquiry. This effect appears despite the reality that eventually other people will view the person's responses to the survey questions even if they were provided only to a computer. Consequently, some form of faux confidentiality effect appears to be operating. Perhaps if you don't have to witness the other person's reaction to your responses, you don't worry about those reactions as much as if the other person is sitting with you and you can directly observe her or him as you answer questions.

Once you consider self-report from a communication perspective, it shifts how you think about self-report methods and may lead you to consider whether it really

is a separate method at least in human service enterprise applications. Table 2.2 inventories three basic approaches to self-report. Instructions only is the type of measurement process in which you simply give the respondent the survey, with instructions written on the survey, and ask him or her to complete it independently. With support involves working with the respondent to make sure he or she understands the instructions and what each question is attempting to measure. Collaboration means that the respondent and a professional sit down together and talk through the survey so that the respondent can fill it out. Which strategy you choose will depend on a variety of factors, including the difficulty of the construct measured and the age, developmental stage, and reading level of the respondent.

Table 2.2 also contains three basic methods for interview. Open-ended interviews are simply discussions. They have no required structure. Semistructured interviews provide some basic structural guidelines in terms of topics and general questions, but limit the structure to more global topics than specific questions. Structured interviews, on the other hand, are fully elaborated. Questions are provided to the interviewer, who is expected to ask them verbatim, and the respondent is given closed-ended response options and asked to endorse one (or more) for each question.

If you consider the options in Table 2.2, you will see there is hardly any difference between a collaborative model of self-report and the structured interview technique. The difference may be only who wields the pencil (or access to the keypad) to actually answer the questions. In self-report, the respondent generally completes the form, while in a structured interview the interviewer does.

One could actually make a similar interpretation of the other pairs of methods. In some ways (although not all), self-report with support and semistructured interviews are similar in that they both give a bit more leeway for the person completing the form to interpret the information herself or himself without the input of others. And, only self-report and open-ended interviews both give the person completing the form a great deal of freedom to interpret the measure in any fashion. The only difference is in the range of response options. Generally, an interview has more

Table 2.2 Basic Methodological Approaches to Collecting Self-Report and Interview Information

Self-report

Instructions only. Informant is given form and completes it independently

Support. Informant completes form but is allowed to ask questions and seek assistance as needed

Collaboration. The form is completed as the informant works through the questions with someone to read and clarify the questions and possible responses.

Interview

Open-ended. Interviewer asks general questions and allows informant to determine the direction of the interview

Semistructured. Interviewer asks a set of defined questions but allows the informant to deviate somewhat based on the content of the interview

Structured. Interview follows strict order of questions and requests that the informant answer the questions in order

response options than a self-report questionnaire (although this is not an absolute requirement).

Exploring Myths in Measurement

Merriam-Webster defines a myth as “(a) a popular belief or tradition that has grown up around something or someone; (b) an unfounded or false notion.” Based on research using psychometric theories and research samples, there are some myths that have become accepted truths of measurement. Primary among these beliefs are the following notions:

- Add or subtract an item from a scale and you change the reliability and validity.
- Change the order of the items and you change the reliability and validity.
- Single-item scales are not likely to be reliable or valid.
- All measures must be “normed.”

As may be obvious from the prior description of measurement of communimetrics, this theory of measurement questions these four beliefs. Since these ideas come close to reaching the perceived level of “truth” in the field of measurement, it is worth discussing why a communimetrics perspective does not accept the truth of these assertions.

Item Inclusion and Sequence

In order to understand these first two beliefs, it is important to consider the context of most measurement research in social sciences. The vast majority of measurement research is accomplished by psychologists in university settings. These researchers balance their need to both teach and engage in productive research by establishing subject pools, often through introductory classes. For example, most introduction to psychology classes provide a very low-grade Sophie’s Choice—either write a paper or participate in a research study. Not surprising, most students choose the research participation over writing an additional paper. Often psychologists are sometimes seen as tricky or manipulative in their research, so these naïve subjects are likely wondering what the purpose of the study in which they are participating. In this context, of course, the inclusion or exclusion of an item makes a difference, or the order of the items shifts how subjects might respond to different questions. They are likely attempting to guess what the experimenter is looking for and tailoring their communication consistent with an emerging (and possibly shifting) theory. This logic does not apply to people seeking assistance in the human service enterprise.

Although there are a large number of studies exploring these issues, they share the same basic method: the study of college undergraduates. As an example, Dahlstrom,

Brooks, and Peterson (1990) demonstrate that scrambling items on the Beck Depression Inventory results in a higher estimated level of depression than does ordering them by severity (as is the standard approach with this measure). Of course the subjects for this study were undergraduate women at the University of North Carolina at Chapel Hill, not people seeking treatment for their depression. Knowles 1988 with a sample of 120 undergraduate psychology and human development students at the University of Wisconsin at Green Bay demonstrated that the later an item occurred in a sequence, the higher its correlation with the total score. This finding was used to posit that over the course of the experiment, the subject was becoming increasingly self-aware (i.e., activated self-schema), and thus the subject is more accurate and reliable over time.

However, there is evidence even with college samples in support of the communimetric perspective. Hamilton and Shuminsky 1990 followed the Knowles study to demonstrate that contextual differences can influence the importance of the serial position of an item. In their study of 242 college undergraduates at the University of Colorado at Colorado Springs, these researchers use Fenigstein and Levine's 1984 story writing method to induce either an internal (self-awareness) or external focus. The subjects participating in the internal focus group were not affected by the serial position of items. Subjects in the external focus group replicated the findings of Knowles (1988).

It is an easy argument that people seeking help from the human service enterprise would be far more likely to be self-aware regarding the reasons that they are seeking help than your average college undergraduate participating in a study that is not necessarily relevant to them other than helping them avoid writing a paper. In fact, I would argue that often self-awareness of need is what actually brings an individual in contact with the human service enterprise in the first place. While this is not always true (e.g., court-mandated treatment for mental health or substance abuse), it is generally true. Regardless, people are coming to human service enterprises for help. These individuals are fundamentally different from college freshman participating in an experiment. There are very few people who would argue that the best way to get accurate information from people in need is to force them to answer a standard set of questions in a standard format. Reliability and validity of a measure are just technical aspects of accuracy. Most experienced human service providers learn that they need to let people tell their stories to start—however they tell it. This builds the type of relation that is required to get accurate information. So in fact, a standard battery of questions that may or may not be relevant to the person seeking assistance is potentially off-putting.

Individual Items

The potential reliability of single items has been demonstrated multiple times in the field of medicine. From the Apgar forward, most clinimetric measures are single items that result in reliable and valid information. Therefore, this myth does not

reflect the existing literature. That being said, it remains the case that linear combinations of variables are generally more reliable than single variables (within that set). But it is a non sequitur to argue that a linear combination of relatively unreliable items is more reliable and, therefore, valid than a well-constructed single item. Anderson et al. (2003) demonstrated that item reliability can be obtained prospectively and with chart audit across a range of 45 different items of a communimetric tool.

Norms

Creating norms for various measures has been a tradition within psychometrics for a long time. The primary purpose of a norm is to try to give meaning to an otherwise arbitrary metric. By creating a standard scale with an identified and known mean and standard deviation it is possible to create clear expectations about the placement of an individual relative to all other individuals in the distribution of scores from that measure. We know that an IQ of 100 is perfectly normal (i.e., average) because 100 is the defined mean of the normative IQ score. Further, we know that an IQ of 130 is two standard deviations above the mean, indicating that 2.5% of the population has IQ scores at this level or higher. Similarly, an IQ of 85 is one standard deviation below the mean, indicating that only about 14% of people have an IQ lower than this one. Norming a measure makes the values more readily interpretable. Sometimes, but not often, that means they are more readily linked to real-world implications (Blanton & Jaccard, 2006). More likely, it gives a quicker sense of where an observation lies in a distribution of scores without telling us anything about its meaningfulness relative to external (real-world) implications of the score.

Communimetrics seeks to evolve many of the “rules” of psychometric measurement in the design phase. However, as discussed in the chapters that follow, when multiple items are combined to create scale scores, a number of psychometric considerations return as requirements for effective measurement in human service enterprises. It is primarily in the design phase that communimetrics represents a different theory of measurement.



<http://www.springer.com/978-0-387-92821-0>

Communimetrics

A Communication Theory of Measurement in Human
Service Settings

Lyons, J.S.

2009, VII, 224 p., Softcover

ISBN: 978-0-387-92821-0