1

What Is Measurement?

Overview

This chapter covers the basics of measurement. The chapter aims to provide understanding of measurement and measure development procedures currently discussed in the literature. Measurement error is introduced, and the steps in the measure development process and empirical procedures to assess error are described. The chapter uses many examples based on data to illustrate measurement issues and procedures. Treatment of the material is at an intuitive, nuts-and-bolts level with numerous illustrations.

The chapter is not intended to be a statistics primer; rather, it provides sufficient bases from which to seek out more advanced treatment of empirical procedures. It should be noted that many issues are involved in using appropriately the statistical techniques discussed here. The discussion in this chapter aims to provide an introduction to specific statistical techniques, and the references cited here provide direction for in-depth reading.

What Is Measurement Error?

Measurement is essential to empirical research. Typically, a method used to collect data involves measuring many things. Understanding how any one thing is measured is central to understanding the entire research method. Scientific measurement has been defined as "rules for assigning numbers to objects in such a way as to represent quantities of attributes" (e.g., Nunnally, 1978, p. 3). Measurement "consists of rules for assigning symbols to objects so as to (1) represent quantities of attributes numerically (scaling) or (2) define whether the objects fall in the same or different categories with respect to a given attribute (classification)" (Nunnally & Bernstein, 1994, p. 3).[1] The attributes of objects, as well as people and events, are the underlying concepts that need to be measured. This element of the definition of measurement highlights the importance of finding the most appropriate attributes to study in a research area. This element also emphasizes understanding what these attributes really mean, that is, fully understanding the underlying concepts being measured. Rules refer to everything that needs to be done to measure something, whether measuring brain activity, attitude toward an object, organizational emphasis on research and development, or stock market performance. Therefore, these rules include a range of things that occur during the data collection process, such as how questions are worded and how a measure is administered. Numbers are central to the definition of measurement for several reasons: (a) Numbers are standardized and allow communication in science, (b) numbers can be subjected to statistical analyses, and (c) numbers are precise. But underneath the façade of precise, analyzable, standardized numbers is the issue of accuracy and measurement error.

The very idea of scientific measurement presumes that there is a thing being measured (i.e., an underlying concept). A concept and its measurement can be distinguished. A measure of a concept is not the concept itself, but one of several possible error-filled ways of measuring it.[2] A distinction can be drawn between conceptual and operational definitions of concepts. A conceptual definition describes a concept in terms of other concepts (Kerlinger, 1986; Nunnally, 1978). For instance, stock market performance is an abstract notion in people's minds. It can be defined conceptually in terms of growth in value of stocks; that is, by using other concepts such as value and growth. An operational definition describes the operations that need to be performed to measure a concept (Kerlinger, 1986; Nunnally, 1978). An operational definition is akin to rules in the definition of measurement discussed earlier in the chapter, and refers to everything that needs to be done to measure something. The Dow Jones average is one measure of stock market performance. This operational definition involves tracking the stock value of a specific set of companies. It is by no means a perfect measure of stock market performance. It is one error-filled way of measuring the concept of stock market performance.

The term *construct* is used to refer to a concept that is specifically defined for scientific study (Kerlinger, 1986). In *Webster's New World Dictionary, construct* means "to build, form or devise." This physical meaning of the word *construct* is similar to the scientific meaning: Constructs are concepts devised or built to meet scientific specifications. These specifications include precisely defining the construct, elaborating on what it means, and relating it to existing research. Words that are acceptable for daily conversation would not fit the specifications for science in terms of clear and precise definitions. "I am going to study what people think of catastrophic events" is a descriptive statement that may be acceptable at the preliminary stages of research. But several concepts in this statement need precise description, such as "think of," which may separate into several constructs, and "catastrophe," which has to be distinguished from other descriptors of events. Scientific explanations are essentially words, and some of these words relate to concepts. Constructs are words devised for scientific study.[3] In science, though, these words need to be used carefully and defined precisely.

When measuring something, error is any deviation from the "true" value, whether it is the true value of the amount of cola consumed in a period of time, the level of extroversion, or the degree of job satisfaction.[4] Although this true value is rarely known, particularly when measuring psychological variables, this hypothetical notion is useful to understand measurement error inherent in scientific research. Such error can have a pattern to it or be "all over the place." Thus, an important distinction can be drawn between consistent (i.e., systematic) error and inconsistent (i.e., random) error (Appendix 1.1). This distinction highlights two priorities in minimizing error. One priority is to achieve consistency,[5] and the second is to achieve accuracy.

Simple explanations of random and systematic error are provided below, although subsequent discussions will introduce nuances. Consider using a weighing machine in a scenario where a person's weight is measured twice in the space of a few minutes with no apparent change (no eating and no change in clothing). If the weighing machine shows different readings, there is *random error* in measurement. In other words, the error has no pattern to it and is inconsistent. Alternatively, the weighing machine may be off in one direction, say, consistently showing a reading that is 5 pounds higher than the accurate value. In other words, the machine is consistent across readings with no apparent change in the weight being measured. Such error is called *systematic error* because there is a consistent pattern to it. It should be noted, though, that on just one reading, the nature of error is not clear. Even if the true value is independently known through some other method, consistency still cannot be assessed in one reading. Multiple readings suggest the inconsistent or consistent nature of any error in the weighing machine, provided the weight of the target person has not changed. Similarly, repetition either across time or across responses to similar items clarifies the consistent or inconsistent nature of error in empirical measurement, assuming the phenomenon across time is unchanged.

If the weighing machine is "all over the place" in terms of error (i.e., random), conclusions cannot be drawn about the construct being measured. Random error in measures attenuates relationships (Nunnally, 1978); that is, it restricts the ability of a measure to be related to other measures. The phrase "all over the place" is, in itself, in need of scientific precision, which will be provided in subsequent pages. Random error has to be reduced before proceeding with any further analyses. This is not to suggest no random error at all, but just that such error should be reasonably low. Certainly, a viable approach is to collect multiple observations of such readings in the hope that the random errors average out. Such an assumption may work when random errors are small in magnitude and when the measure actually captures the construct in question. The danger here, though, is that the measure may not capture any aspect of the intended construct. Therefore, the average of a set of inaccurate items remains inaccurate. Measures of abstract concepts in the social sciences, such as attitudes, are not as clear-cut as weighing machines. Thus, it may not be clear if, indeed, a construct, such as an attitude, is being captured with some random error that averages out. The notion that errors average out is one reason for using multiple items, as discussed subsequently.

Measures that are relatively free of random error are called *reliable* (Nunnally, 1978). There are some similarities between the use of reliability in measurement and the common use of the term. For example, a person who is reliable in sticking to a schedule is probably consistently on time. A reliable friend is dependable and predictable, can be counted on, and is consistent. However, there are some differences as well. A reliable person who is consistent but always 15 minutes late would still be reliable in a measurement context. Stated in extreme terms, reliability in measurement actually could have nothing to do with accuracy in measurement, because reliability relates only to consistency. Without some degree of consistency, the issue of accuracy may not be germane. If a weighing machine is all over the place, there is not much to be said about its accuracy. Clearly, there are shades of gray here in that some small amount of inconsistency is acceptable. Waiting for perfect consistency before attempting accuracy may not be as efficient or pragmatic as achieving reasonable consistency and approximate accuracy. Perfect consistency may not even be possible in the social sciences, given the inherent nature of phenomena being studied.

Whether a measure is accurate or not is the realm of *validity,* or the degree to which a measure is free of random and systematic error (Nunnally, 1978). If the weighing machine is consistent, then whether it is accurate becomes germane. If the weighing machine is consistent but off in one direction, it is reliable but not valid. In other words, there is consistency in the nature of error. To say a measure is accurate begs the question, "Accurate in doing what?" Here, validity refers to accuracy in measuring the intended construct.[6]

Random and systematic errors are the two types of errors that are pervasive in measurement across the sciences. Low reliability and low validity have consequences. Random error may reduce correlations between a measure and other variables, whereas systematic error may increase or reduce correlations between two variables. A brief overview of the measure development process is provided next. The purpose here is to cover the basic issues briefly to provide a background for more in-depth understanding of types of measurement error. However, the coverage is markedly different from treatment elsewhere in the literature. Presentation is at a level to enable intuitive understanding. Statistical analyses are presented succinctly with many illustrations and at an intuitive level.

Overview of Traditional Measure Development Procedures

A number of steps have been suggested in the measure development process (Churchill, 1979; Gerbing & Anderson, 1988) that are adapted here and discussed.[7] As in many stepwise processes, these steps are often blurred and iterative. The process is illustrated in Figure 1.1. The steps in the process emphasize that traversing the distance from the conceptual to the operational requires a systematic process. Rather than consider a concept and move directly to item generation, and use of a resulting measure, the distance between the conceptual and the operational has to be spanned carefully and iteratively.

Conceptual and Operational Definitions

The very idea of measurement suggests an important distinction between a concept and its measurement. Hence, the literature distinguishes between conceptual and operational definitions of *constructs* (i.e., concepts specifically designed for scientific study) (Kerlinger, 1986). Scientific research is about constructing abstract devices; nevertheless, the term *construct* is similar in several respects to constructing concrete things. In developing

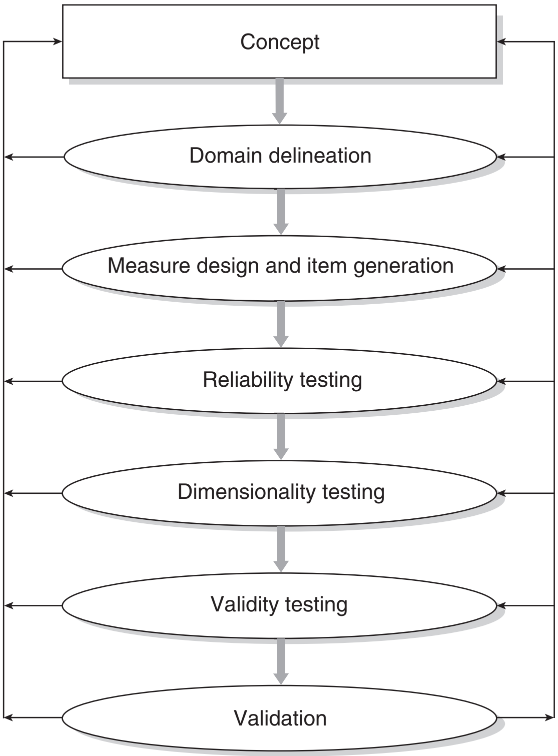


Figure 1.1 Steps in the Measure Development Process

constructs, the level of abstraction is an important consideration. If constructs are too concrete, then they are not as useful for theoretical generalization, although their measurement may be more direct. If constructs are too abstract, their direct measurement is difficult, although such constructs can be used for developing medium-level constructs that are measurable. For example, Freud's concepts of id and superego may not be directly measurable (not easily, anyway), yet they can be used to theorize and derive medium-level constructs. On the other hand, a construct, such as response time or product sales, is more concrete in nature but lacks the explanatory ability of a more abstract construct. Response time is often of interest in cognitive psychology in that it is an indicator of an underlying construct, such as cognitive effort. By itself, response time may not have the same level of theoretical importance.

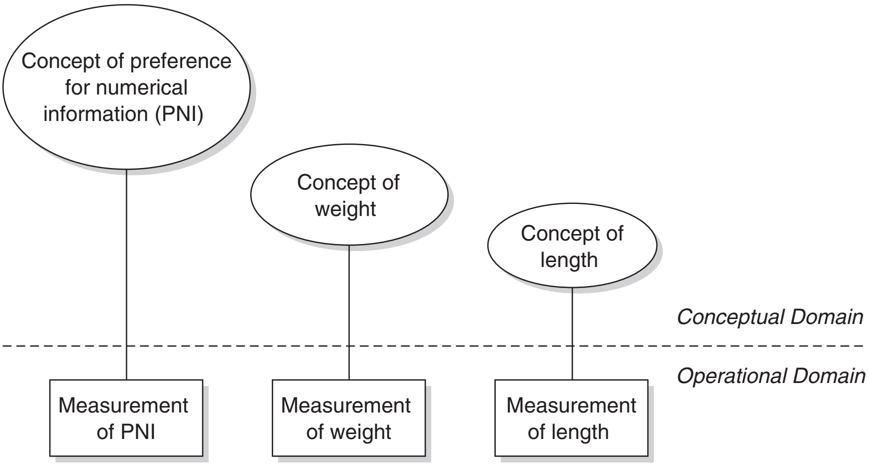


Figure 1.2 Distances Between Conceptual and Operational Domain for Physical and Psychological Constructs and Measures

As the abstractness of a construct increases, the distance between the conceptual and the operational definitions increases (Figure 1.2). Other than a few dimensions such as length (for short distances, not for astronomical ones!) and time (at least in recent times, an issue that will be revisited in a subsequent chapter), the operational definition or measure of a construct is indirect (e.g., weight or temperature) (Nunnally, 1978). For example, length can be defined conceptually as the shortest distance between two points. Measurement of length follows directly, at least for short lengths. The same could be said for time. However, weight or temperature involve more abstract conceptual definitions, as well as larger distances between the conceptual and the operational. Measurement is more indirect, such as through the expansion of mercury for temperature, that is, a correlate of temperature, or gravitational pull on calibrated weights for weight.

In moving to the social sciences, the distance between the conceptual and the operational can be large, for example, in measuring attitudes toward objects or issues. Larger distances between the conceptual and the operational have at least two implications. As the distance increases, so, too, do measurement error and the number of different ways in which something can be measured (Figure 1.3). This is akin to there being several ways of getting to a more distant location, and several ways of getting lost as well!

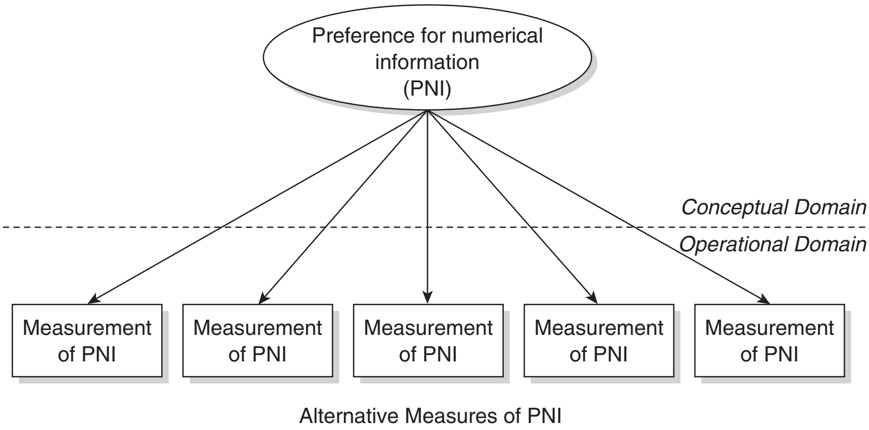


Figure 1.3 Distance Between the Conceptual and the Operational, Potential Operationalizations and Error

Preference for numerical information, or PNI, is used as a sample construct in this chapter. This construct relates to an enduring proclivity toward numbers across several contexts, a relatively abstract construct. In contrast, a less abstract construct may be consumers' preference for numerical information or employees' preference for numerical information, essentially restricting the context of the construct (Figure 1.4). These context-specific constructs may be at a level of abstraction that is useful in deriving theoretical generalizations. In contrast, consider usage of calorie information as a construct (Figure 1.5). Here, the numerical context is further restricted. Although this construct may be useful in the realm of consumption studies, it is relatively concrete for purposes of understanding the use of numerical information in general. The point is that there is less theoretical generalization across different domains when a construct is relatively concrete. Abstract constructs allow for broader generalization. Suppose that a basic preference for numerical information is expected to cause higher performance in numerically oriented tasks. From the perspective of understanding numerical information, this is an argument at an abstract level with broader generalization than one linking usage of calorie information to, say, performance in choosing a brand of cereal. In fact, a modified form of the latter scenario of calorie information and choice of cereals may be a way to test the former scenario involving preference for numerical information and performance in numerically oriented tasks. But such an empirical test would

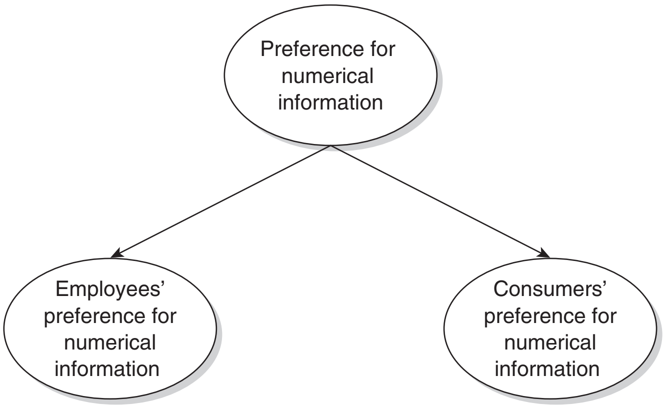

Preference for numerical information
Employees' preference for numerical information
Consumers' preference for numerical information

Figure 1.4 Levels of Abstractness of Constructs

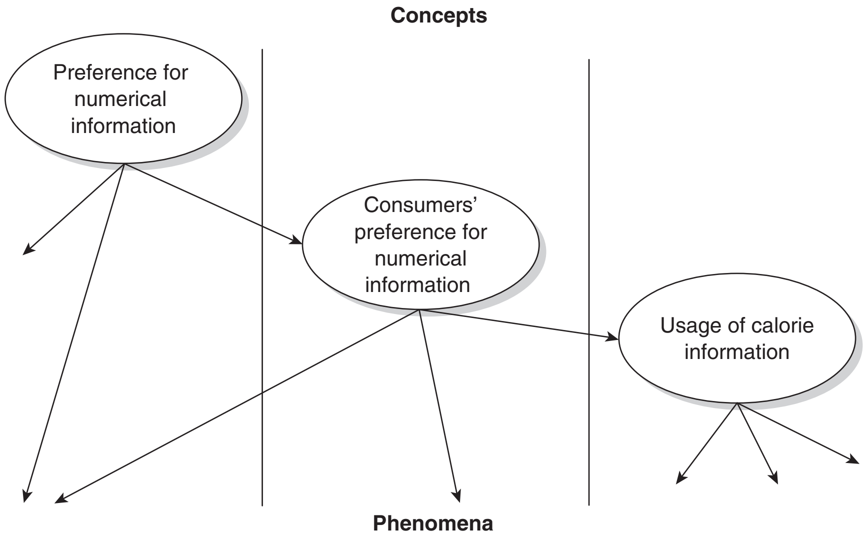

Concepts
Preference for numerical information
Consumers' preference for numerical information
Usage of calorie information
Phenomena

Figure 1.5 Abstractness of Constructs

have to be developed by carefully deriving ways of measuring the concepts being studied and choosing the best way to operationalize them. Empirical tests in specific, narrow contexts ideally should be developed from broader theoretical and methodological considerations, thereby choosing the best setting in which a theory is tested.

In conceptually defining a construct, it should be sufficiently different from other existing constructs. In other words, the construct should not be redundant and should make a sizable contribution in terms of explanatory power. Conducting science requires linking the work to past knowledge. Existing research is an important guide in defining a construct. The words that are used to define a construct have to communicate to a larger scientific audience. Defining constructs in idiosyncratic fashion that are counter to their use in the extant literature precludes or hinders such communication. This is not to preclude the redefinition of existing constructs. In fact, it is necessary to their constantly evaluate construct definitions. However, the point being made here is that redefinitions should be supported with compelling rationale. Existing theory in the area should be used to define a construct to be distinct from other constructs. The distinction should warrant the creation of a new construct. Thus, the hurdle for the development of new constructs is usually high. Conceptually rich constructs enable theoretical generalizations of importance and interest to a discipline.

The distance between the conceptual and operational can also lead to confounding between measurement and theory. Sometimes, discussions of conceptual relationships between constructs and hypotheses about these relationships may confound constructs with their operationalizations, essentially mixing two different levels of analysis (e.g., arguing for the relationship between PNI and numerical ability on the basis of specific items of the PNI scale, rather than at a conceptual level). This highlights the need to keep these two levels of analysis separate while iterating between them in terms of issues, such as conceptual definitions of constructs and rationale for conceptual relationships between constructs. Iterative analysis between these two levels is common and necessary; however, a clear understanding and maintenance of the distinction is critical. Alternatively, measures that aim to assess a specific construct may indeed assess a related construct, either an antecedent or an outcome (say, preference for numerical information measuring numerical ability or preference for qualitative information), thereby confounding constructs. Constructs may also have multiple dimensions, each with a different relationship with other constructs (say, usage of numbers and enjoyment of numbers, with the former having a stronger relationship with a measure of numerical ability) and that may need to be clarified. Such clarification often occurs as research in an area progresses and theoretical sophistication leads to sophistication in measurement and vice versa.

Domain Delineation

The next step in the process of developing a measure of a construct is to delineate its domain (Churchill, 1979; DeVellis, 1991). This step involves explicating what the construct is and is not. Related constructs can be used to explain how the focal construct is both similar to and different from them. This is also a step where the proposed dimensionality of the construct is described explicitly. For instance, a measure of intelligence would be explicated in terms of different dimensions, such as quantitative and verbal intelligence. The domain of the construct is described, thus clearly delineating what the construct is and is not. At its core, this step involves understanding what is being measured by elaborating on the concept, *before the measure is designed and items in the measure are generated.* Careful domain delineation should precede measure design and item generation as a starting point. This argument is not intended to preclude iterations between these two steps; iterations between measure design and item generation, and domain delineation are invaluable and serve to clarify the domain further. Rather, the point is to consider carefully what abstract concept is being measured as a starting point before attempting measure design and item generation and iterating between these steps. Collecting data or using available data without attention to underlying concepts is not likely to lead to the development of knowledge at a conceptual level.

The goal of domain delineation is to explicate the construct to the point where a measure can be designed and items can be generated. Domain delineation is a step in the conceptual domain and not in the operational domain. Therefore, different ways of measuring the construct are not considered in this step. Domain delineation precedes such consideration to enable fuller understanding of the construct to be measured. Several issues should be considered here, including using past literature and relating the construct to other constructs. In other words, the construct has to be placed in the context of existing knowledge, thus motivating its need and clarifying its uniqueness. The construct should be described in different ways, in terms of what is included and what is *not* included by the domain. Such an approach is an effective way of clarifying a construct and distinguishing it from related constructs. For example, preference for numerical information has to be clearly differentiated from numerical ability. This is similar to clarifying the exact meaning of related words, whether it is *happiness* versus *contentment, bluntness* versus *honesty*, and so on. In fact, if anything, scientific research can be distinguished in terms of the precision with which words are used, all the more so with words that denote constructs, the focus of scientific inquiry. Just as numbers are used in scientific measurement because they are standardized and they facilitate communication, words should be used to convey precise meaning in order to facilitate communication. Incisive conceptual thinking, akin to using a mental scalpel, should carefully clarify, distinguish, and explicate constructs well before data collection or even measure development is attempted. Such thinking will also serve to separate a construct from its antecedents and effects. A high hurdle is usually and appropriately placed in front of new constructs in terms of clearly distinguishing them from existing constructs.

A worthwhile approach to distinguishing related constructs is to construct scenarios where different levels of each construct coexist. For instance, if happiness and contentment are two related constructs, constructing examples where different levels of happiness versus contentment coexist can serve to clarify the domain of each at a conceptual level as well, distinguishing a construct from what it is not and getting to the essence of it conceptually. The process of moving between the conceptual and the operational can clarify both.

The continuous versus categorical nature of the construct, as well as its level of specificity (e.g., too broad vs. too narrow; risk aversion vs. financial risk aversion or physical risk aversion; preference for numerical information vs. consumers' preference for numerical information) need to be considered. Constructs vary in their level of abstractness, which should be clarified in domain delineation. The elements of the domain of the construct need to be explicated (e.g., satisfaction, liking, or interest with regard to, say, a construct relating to preference for some type of information). The purpose here is to carefully understand what exactly the domain of the construct covers, such as tendencies versus attitudes versus abilities. Visual representation is another useful way of thinking through the domain. An iterative process where items are developed at an operational level can, in turn, help in domain delineation as well by concretizing the abstract domain.

Domain delineation is an important step that enables understanding the conceptual as it relates to the operational. Delineation may well change the conceptual definition. Such iteration is common in measure development. The key is to allow sufficient iteration to lead to a well-thought-out measurement process. Domain delineation may help screen out constructs that are too broad or narrow. If constructs are too abstract, intermediate-level constructs may be preferable. Figure 1.6 illustrates a narrow versus broad operationalization of the domain of preference for numerical information. Figure 1.6 also illustrates how domain delineation can facilitate item generation by identifying different aspects of the domain. Thus, during the

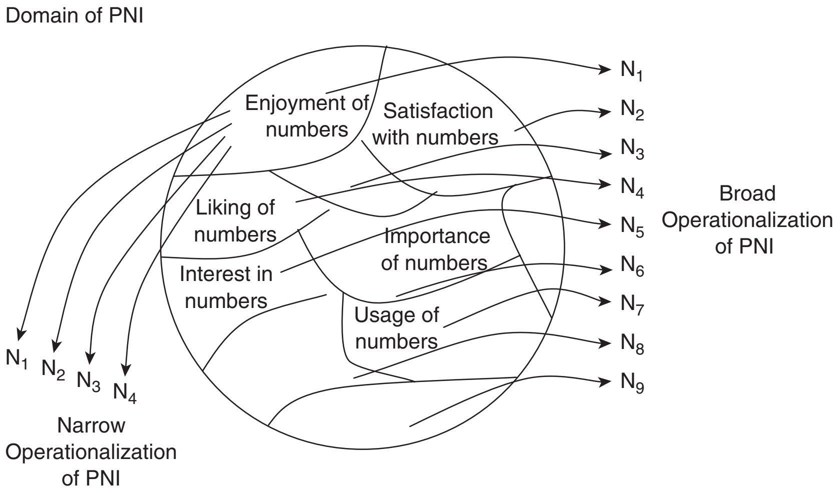


Figure 1.6 Domain Delineation and Item Generation

subsequent step of item generation, items can be generated to cover aspects such as enjoyment and importance. In this way, the distance between the conceptual and the operational is bridged.

Domain delineation is demonstrated using an example of the preference for numerical information (PNI) scale (Exhibit 1.1). Several points are noteworthy. First, PNI is distinguished from ability in using numerical information, and several aspects of preference are discussed. PNI is also distinguished from statistics and mathematics. In addition, PNI is described in terms of its level of abstraction in being a general construct rather than specific to a context such as consumer settings or classroom settings. In this respect, it can be distinguished from other scales such as attitude toward statistics and enjoyment of mathematics. The distinctive value of the PNI construct is thus illustrated by showing that it is a unique construct that is different from potentially related constructs, and that it has the potential to explain important phenomena. As discussed earlier, any proposed new construct has to answer the “So what?” question. After all, any construct can be proposed and a measure developed. There are different ways of motivating a construct: The PNI scale is motivated in a more generic way because of lack of existing theory, but alternatively, a scale could be motivated by placing it in the context of past theory or by identifying a gap in past theory.

Exhibit 1.1 The Preference for Numerical Information Scale: Definition, Domain Delineation, and Item Generation

The errors discussed earlier are illustrated using data from a published scale, the preference for numerical information scale (Viswanathan, 1993). Preference for numerical information is defined as *a preference or proclivity toward using numerical information and engaging in thinking involving numerical information.* Several aspects of this definition need to be noted. First, the focus is on preference or proclivity rather than ability, because the aim here is to focus on *attitude* toward numerical information. Second, the focus is solely on numerical information, rather than on domains such as statistics or mathematics, in order to isolate numerical information from domains that involve the use of such information. Third, PNI is conceptualized as a broad construct that is relevant in a variety of settings by using a general context rather than a specific context such as, say, an academic setting.

Item Generation

Items were generated for the PNI scale in line with an operationalization of the definition of the construct. As mentioned earlier, three aspects of importance in the definition are the focus on preference or proclivity, the focus on numerical information, and the use of a general rather than a specific context. The domain of the construct was operationalized by using parallel terms that represent numerical information, such as *numbers, numerical information,* and *quantitative information.* Proclivity or preference for numerical information was operationalized using a diverse set of elements or aspects, such as the extent to which people enjoy using numerical information (e.g., "I enjoy work that requires the use of numbers"), liking for numerical information (e.g., "I don't like to think about issues involving numbers"), and perceived need for numerical information (e.g., "I think more information should be available in numerical form"). Other aspects were usefulness (e.g., "Numerical information is very useful in everyday life"), importance (e.g., "I think it is important to learn and use numerical information to make well informed decisions"), perceived relevance (e.g., "I don't find numerical information to be relevant for most situations"), satisfaction (e.g., "I find it satisfying to solve day-to-day problems involving numbers"), and attention/interest (e.g., "I prefer not to pay attention to information involving numbers"). The use of information in a general rather than a specific context was captured by wording items to be general ("I prefer not to pay attention to information involving numbers") rather than specific to any context.

A pool of 35 items was generated in line with the operationalization described above in the form of statements with which a respondent could agree or disagree to varying degrees. These items were generated to cover the range of aspects of preference for numerical information listed above (such as satisfaction and usefulness). A total of 20 items were chosen from this pool and inspected in terms of content for coverage of these different aspects, usage of different terms to represent numerical information, and generality of context. The items were also chosen such that an equal number were positively or negatively worded with respect to preference for numerical information. The response format was a 7-point scale labeled at the extremes as *strongly agree* and *strongly disagree*.

Items of the PNI Scale

		Strongly Disagree						*Strongly Agree*
1.	I enjoy work that requires the use of numbers.	1	2	3	4	5	6	7
2.	I find information easy to understand if it does not involve numbers.	1	2	3	4	5	6	7
3.	I find it satisfying to solve day-to-day problems involving numbers.	1	2	3	4	5	6	7
4.	Numerical information is very useful in everyday life.	1	2	3	4	5	6	7
5.	I prefer not to pay attention to information involving numbers.	1	2	3	4	5	6	7
6.	I think more information should be available in numerical form.	1	2	3	4	5	6	7
7.	I don't like to think about issues involving numbers.	1	2	3	4	5	6	7
8.	Numbers are not necessary for most situations.	1	2	3	4	5	6	7
9.	Thinking is enjoyable when it does not involve quantitative information.	1	2	3	4	5	6	7
10.	I like to make calculations using numerical information.	1	2	3	4	5	6	7
11.	Quantitative information is vital for accurate decisions.	1	2	3	4	5	6	7
12.	I enjoy thinking based on qualitative information.	1	2	3	4	5	6	7
13.	Understanding numbers is as important in daily life as reading or writing.	1	2	3	4	5	6	7
14.	I easily lose interest in graphs, percentages, and other quantitative information.	1	2	3	4	5	6	7
15.	I don't find numerical information to be relevant for most situations.	1	2	3	4	5	6	7
16.	I think it is important to learn and use numerical information to make well-informed decisions.	1	2	3	4	5	6	7
17.	Numbers are redundant for most situations.	1	2	3	4	5	6	7
18.	Learning and remembering numerical information about various issues is a waste of time.	1	2	3	4	5	6	7
19.	I like to go over numbers in my mind.	1	2	3	4	5	6	7
20.	It helps me to think if I put down information as numbers.	1	2	3	4	5	6	7

EXHIBIT SOURCE: Adapted from Viswanathan, M., Measurement of individual differences in preference for numerical information, in *Journal of Applied Psychology*, *78*(5), 741–752. Copyright © 1993. Reprinted by permission of the American Psychological Association.

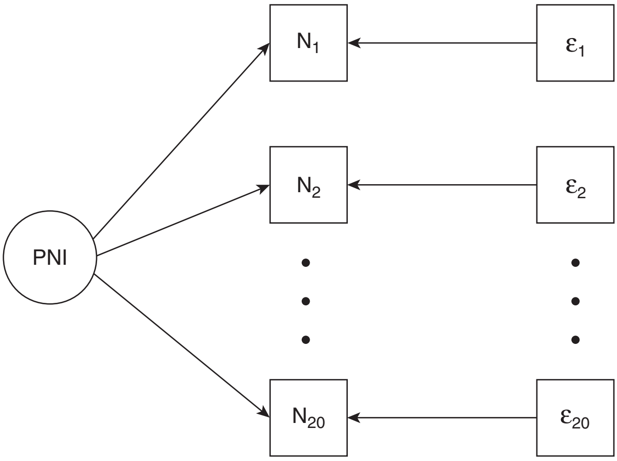


Figure 1.7 Visual Representation for PNI of Latent Construct, Items, and Error

Measure Design and Item Generation

Measure design and item generation follows domain delineation as illustrated for the PNI scale (Exhibit 1.1). Before specific items can be generated, the design of the measure needs to be determined. Important here is the need to think beyond measures involving agreement-disagreement with statements, and also beyond self-report measures. Measures can range from observational data to behavioral inventories. Later in the book, a variety of measures from different disciplines are reviewed to highlight the importance of creative measure design. The basic model that relates a construct to items in a measure is shown in Figure 1.7. A latent construct causes responses on items of the measure. The items are also influenced by error. The error term is the catch-all for everything that is not the construct.

Redundancy is a virtue during item generation, with the goal being to cover important aspects of the domain. Even trivial redundancy—that is, minor grammatical changes in the way an item is worded—is acceptable at this stage (DeVellis, 1991). Measure development is inductive in that the actual effects of minor wording differences are put to empirical test, and items that pass the test are retained. It is impossible to anticipate how people may respond to each and every nuance in item wording. Therefore, items are tested by examining several variations. All items are subject to interpretation. The aim is to develop items that most people in the relevant population would interpret in a similar fashion. Generic recommendations to enable such interpretation include avoiding ambiguous, lengthy, complex, or double-barreled items. Items also should not be worded to be extreme (e.g., I hate numbers), as discussed subsequently.

Unit of analysis is another consideration in item development (Sirotnik, 1980). For instance, items designed to capture individual perceptions about self versus other versus groups suggest different types of wording. Items such as "My workplace is friendly" versus "My colleagues view our workplace as being friendly" versus "Our work group views our workplace as friendly" capture different perceptions—an individual's perception of the workplace, colleagues' perceptions of the workplace (as measured by an individual's perception of these perceptions), and the group's perceptions of the workplace (as measured by an individual's perception of these perceptions), respectively.

Level of abstraction is an important consideration with item generation. Items vary in their level of abstractness. For instance, items could be designed at a concrete level about what people actually do in specific situations versus at a more abstract level in terms of their attitudes. On one hand, concrete items provide context and enable respondents to provide a meaningful response (say, a hypothetical item on the PNI scale such as, "I use calorie information on packages when shopping"). On the other hand, items should ideally reflect the underlying construct and not other constructs, the latter being a particular problem when it occurs for a subset of items, as discussed later. Because PNI is a more global construct that aims to cover different contexts, including the consumer context, an item specific to the consumer context may not be appropriate when other items are at a global level. Such an item also does not parallel other items on the PNI scale that are more global in their wording, the point being that the wording of items should be parallel in terms of level of abstraction. A different approach may well identify specific subdomains (e.g., social, economic) and develop items that are specific to each such subdomain. Thus, the subdomains are based on contexts where preference for numerical information plays out. In such an approach, items at the consumer level may be appropriate. Each of these subdomains essentially would be measured separately, the issue of items worded to be parallel in terms of the level of abstraction being germane for each subdomain. Another approach is to divide up the domain into subdomains, such as enjoyment of numbers and importance of numbers; that is, in terms of the different aspects of preference. Each subdomain could be conceptualized as a separate dimension and items developed for each dimension. Of course, the hypothesized dimensionality has to be tested empirically.

A large pool of items is important, as are procedures to develop a large pool. Several procedures can be employed to develop a pool of items, such as asking experts to generate items and asking experts or respondents to evaluate items in the degree to which they capture defined constructs (Haynes, Richard, & Kubany, 1995). There may be a temptation to quickly put together a few items because item generation seems intuitive in the social sciences. However, the importance of systematic procedures as a way to develop a representative set or sample of items that covers the domain of the construct cannot be overemphasized. These procedures, which take relatively small additional effort in the scheme of things, can greatly improve the chances of developing a representative measure. Moreover, it should be noted that item generation is an iterative process that involves frequently traversing the pathways between the conceptual and the operational. If many items are quickly put together in the hope that errors average out, the average may be of fundamentally inaccurate items, and therefore not meaningful.

The procedures discussed so far relate to the content validity of a measure, that is, whether a measure adequately captures the content of a construct. There are no direct empirical measures of content validity, only assessment based on whether (a) a representative set of items was developed and (b) acceptable procedures were employed in developing items (Nunnally, 1978). Hence, assessment of content validity rests on the explication of the domain and its representation in the item pool. The proof of the pudding is in the procedures used to develop the measure rather than any empirical indicator. In fact, indicators of reliability can be enhanced at the expense of content validity by representing one or a few aspects of the domain (as shown through narrow operationalization in Figure 1.6) and by trivial redundancies among items. An extreme scenario here is, of course, repetition of the same item, which likely enhances reliability at the expense of validity.

Internal Consistency Reliability

After item generation, measures typically are assessed for internal consistency reliability, and items are deleted or modified. Internal consistency frequently is the first empirical test employed and assesses whether items within a set are consistent with each other, that is, covary with each other. Internal consistency procedures assess whether a set of items fits together or belongs together. Responses are collected on a measure from a sample of respondents. Intercorrelations among items and correlations between each item and the total score are used to purify measures, using an overall indicator of internal consistency reliability, coefficient alpha (Cronbach, 1951).

Illustrative descriptions are provided using the PNI scale (Exhibit 1.2). Means for the items in the PNI scale are shown. Means close to the middle of the scale are desirable to allow for sufficient variance and the ability to covary with other items. If an item does not vary, it cannot covary, and measurement is about capturing variation as it exists. (A parallel example is in experimental design, where, say, the dependent variable is recall of information shown earlier, wherein either one or two pieces of information could be shown leading to very high recall and small variance, or a hundred pieces of information shown leading to very low recall and small variance. In both instances, lack of variation precludes appropriate tests of hypotheses.) As an illustration, Item 18 in Exhibit 1.2 (Learning and remembering numerical information about various issues is a waste of time) has a mean of 5.24; it was subsequently found to have low item-to-total correlations in some studies and replaced. The item is worded in somewhat extreme terms, referring to "waste of time." This is akin to having items with words such as "never" or "have you ever" or "do you hate." Items 4 and 5 are additional examples relating to usefulness of numbers and not paying attention to numbers. Item 17 relates to numbers being redundant for most situations—perhaps another example of somewhat extreme wording. An important issue here is that the tone of the item can lead to high or low means, thus inhibiting the degree to which an item can vary with other items. The item has to be valenced in one direction or the other (i.e., stated positively or negatively with respect to the construct in question) to elicit levels of agreement or disagreement (because disagreement with a neutral statement such as "I neither like nor dislike numbers" is ambiguous). However, if it is worded in extreme terms (e.g., "I hate . . ."), then item variance may be reduced. Therefore, items need to be moderately valenced. The name of the game is, of course, variation, and satisfactory items have the ability to vary with other items. Scale variance is the result of items with relatively high variance that covary with other items. This discussion is also illustrative of the level of understanding of the relationship between item characteristics and empirical results that is essential in measure development. Designers of measures, no matter how few the items, have to do a lot more than throw some items together.

Item-to-total correlations are typically examined to assess the extent to which items are correlated with the total. As shown in Appendix 1.1, a matrix of responses is analyzed in terms of the correlation between an item and the total score. The matrix shows individual responses and total scores. The degree to which an individual item covaries with other items can be assessed in a number of ways, such as through examining intercorrelations between items or the correlation between an item and the total scale. Items

(Text continues on page 24)

Exhibit 1.2 Internal Consistency of PNI Scale

Analysis of 20-Item PNI Scale

Results for the 20-item PNI scale (Viswanathan, 1993) are presented below for a sample of 93 undergraduate students at a midwestern U.S. university. Means and standard deviations can be examined. Extreme means or low standard deviations suggest potential problems, as discussed in the chapter.

		Mean	*Std. Dev.*
1.	N1	4.3407	2.0397
2.	N2	3.7253	1.9198
3.	N3	4.3956	1.9047
4.	N4	5.2747	1.6304
5.	N5	5.3297	1.3000
6.	N6	3.9890	1.6259
7.	N7	5.0000	1.6063
8.	N8	4.7363	1.6208
9.	N9	4.3516	1.6250
10.	N10	4.5495	1.9551
11.	N11	5.0769	1.3269
12.	N12	3.4286	1.3755
13.	N13	5.0989	1.3586
14.	N14	4.3297	1.7496
15.	N15	4.8571	1.2163
16.	N16	5.0769	1.3600
17.	N17	5.0220	1.2109
18.	N18	5.2418	1.1489
19.	N19	3.7033	1.7670
20.	N20	4.2527	1.9096

The average interitem correlation provides an overall indicator of the internal consistency between items. It is the average of correlations between all possible pairs of items in the 20-item PNI scale.

Average interitem correlation = .2977

Item-to-total statistics report the correlation between an item and the total score, as well as the value of coefficient alpha if a specific item were deleted. Items with low correlations with the total score are candidates for deletion. Item 12 actually has a negative correlation with the total score and should be deleted. It should be noted that all analyses are after appropriate reverse scoring of items such that higher scores reflect higher PNI.

Item-Total Statistics

	Corrected Item-Total Correlation	*Alpha If Item Deleted*
N1	.7736	.8839
N2	.3645	.8975
N3	.7451	.8857
N4	.5788	.8911
N5	.4766	.8939
N6	.6322	.8896
N7	.7893	.8853
N8	.4386	.8949
N9	.4781	.8938
N10	.7296	.8861
N11	.4843	.8937
N12	−.3603	.9143
N13	.5073	.8931
N14	.6287	.8895
N15	.4617	.8943
N16	.6095	.8904
N17	.5301	.8927
N18	.3947	.8958
N19	.5323	.8924
N20	.6281	.8894

Coefficient alpha is the overall indicator of internal consistency as explained in the chapter.

Alpha = .8975 Standardized item alpha = .8945

The analysis was repeated after deleting Item 12.

Analysis After Deletion of Item 12 (19-item version)
Average interitem correlation = .3562

Item-Total Statistics

	Corrected Item-Total Correlation	*Alpha If Item Deleted*
N1	.7705	.9043
N2	.3545	.9160
N3	.7479	.9052
N4	.5834	.9097
N5	.4823	.9121
N6	.6219	.9088
N7	.7889	.9047

(Continued)

Item-Total Statistics

N8	.4432	.9131
N9	.4811	.9122
N10	.7299	.9056
N11	.5071	.9115
N13	.5175	.9113
N14	.6373	.9083
N15	.4672	.9124
N16	.6237	.9088
N17	.5306	.9111
N18	.3950	.9138
N19	.5381	.9109
N20	.6339	.9084

Alpha = .9143 Standardized item alpha = .9131

The analysis is repeated after deleting another item, Item 2, which has a relatively low item-to-total correlation.

Analysis After Deletion of Item 2 (18-item version)
Average interitem correlation = .3715

Item-Total Statistics

	Corrected Item-Total Correlation	*Alpha If Item Deleted*
N1	.7650	.9063
N3	.7442	.9069
N4	.6025	.9111
N5	.4806	.9140
N6	.6144	.9108
N7	.7916	.9062
N8	.4417	.9151
N9	.4589	.9147
N10	.7290	.9073
N11	.5298	.9129
N13	.5277	.9129
N14	.6453	.9098
N15	.4726	.9142
N16	.6307	.9104
N17	.5156	.9132
N18	.3829	.9159
N19	.5405	.9128
N20	.6374	.9101

Alpha = .9160 Standardized item alpha = .9141

(Continued)

The effect of the number of items in a scale on reliability is noteworthy. Below, 10- and 5-item versions of the PNI scale are presented.

Analysis of 10- and 5-Item Versions of the PNI Scale

Reliability Analysis—10-item version
Average interitem correlation = .3952

Item-Total Statistics

	Corrected Item-Total Correlation	*Alpha If Item Deleted*
N1	.7547	.8411
N2	.3885	.8723
N3	.7489	.8424
N4	.5212	.8611
N5	.4872	.8634
N6	.5702	.8575
N7	.8033	.8402
N8	.3696	.8718
N9	.4761	.8643
N10	.7521	.8417
Alpha = .8688	Standardized item alpha = .8673	

Reliability Analysis—5-Item Version
Average interitem correlation = .3676

Item-Total Statistics

	Corrected Item-Total Correlation	*Alpha If Item Deleted*
N1	.7016	.6205
N2	.3189	.7705
N3	.6888	.6351
N4	.4648	.7198
N5	.4357	.7298
Alpha = .7475	Standardized item alpha = .7440	

that are internally consistent would have high correlation with the total score. Consider a scale that is assumed to consist of a number of internally consistent items. The higher an individual's response on a particular item, the higher the individual's response is likely to be on other items in the scale, and the higher the individual's total score. A person high on PNI would be expected to have a high total score as well as high scores on individual items (once the items have been reverse scored so that higher values reflect higher PNI; for instance, an item such as "I don't like numbers" would be reverse scored so that a response of a 7 is equivalent to a 1, and so on). Items with low correlations with the total score are deleted or modified. Such items are assumed to be lacking internal consistency. They do not covary or are not consistent with the total score or with other items.

Item 12 (Exhibit 1.2 and Figure 1.8) is instructive in that it has a negative correlation with the total score after reverse scoring. The item pertains to qualitative information and is employed here on the assumption that higher PNI may be associated with lower preference for qualitative information. In other words, rather than directly capture preference for numerical information, the item is premised on a contingent relationship between PNI and preference for qualitative information. In fact, for some individuals, high preferences for both types of information may coexist. The key lesson here is the importance of directly relating items to the constructs being measured. Essentially, the item confounds constructs it purports to measure. The negative correlation may also have been the result of some respondents misreading the term *qualitative* as *quantitative*. It should be noted here that the item was appropriately reverse-scored before data analyses, including computing correlations.

Another example from a different research area is illustrative. Consider a measure that attempts to capture the type of strategy a firm adopts in order to compete in the marketplace. A strategy may consist of many different aspects, such as emphasis on new product development and emphasis on research and development. Therefore, self-report items could be generated to measure strategy regarding emphasis on research and development and emphasis on new product development (e.g., "We emphasize new product development to a greater degree than our competition," with responses ranging from *strongly disagree* to *strongly agree*). Suppose also that there is an empirical relationship between the type of environment in which firms operate (e.g., uncertain environments) and the type of strategy in which firms engage; that is, in uncertain environments, firms tend to invest in research and development and new product development. Then, an item generated to measure strategy about how uncertain the environment is (e.g., "Our company operates in an uncertain environment," with

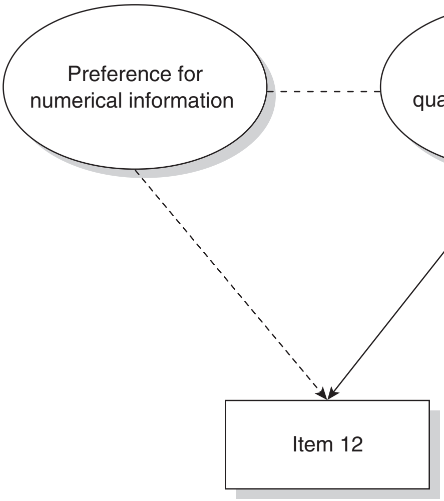

Preference for numerical information
Preference for qualitative information
Item 12

Figure 1.8 Indirect Versus Direct Measurement of Constructs

responses from *strongly disagree* to *strongly agree*) attempts to tap into strategy on the basis of a contingent relationship between the type of strategy and the type of environment. Such an item assumes that firms adopt a particular strategy in a certain environment. As far as possible, items should be developed to directly assess the construct they aim to measure. Otherwise, substantive relationships between constructs are confounded with measures of single constructs. Subtler issues arise depending on the type of indicator that is being developed, as discussed subsequently. Direct measurement does not imply repetitive, similar items. Any item typically captures a construct in some context. Creative measurement (and perhaps interesting items from the respondents' perspective) involves different ways of capturing a construct in different contexts. A key point to note is that ideally, items should not confound constructs but rather use contexts in which the focal construct plays out.

Coefficient alpha is an indicator of the internal consistency reliability of the entire scale. A goal in internal consistency procedures is to maximize *coefficient alpha,* or the proportion of variance attributable to common sources (Cronbach, 1951; DeVellis, 1991). Items are deleted on this basis to achieve a higher coefficient alpha, and this process is repeated until the marginal gain in alpha is minimal. Beyond a point, the marginal increase in alpha may not warrant additional deletion of items. In fact, items with moderate correlations with the total may be retained if they capture some unique aspect of the domain not captured by other items. The average interitem correlation, along with coefficient alpha, provides a summary, and researchers have suggested that its ideal range is between 0.2 and 0.4 (Briggs & Cheek, 1988). The rationale for this guideline is that relatively low correlations suggest that a common core does not exist, whereas relatively high correlations suggest trivial redundancies and a narrow operationalization of one subdomain of the overall construct. For example, for the PNI scale, if all the correlations in Exhibit 1.2 between items are close to 1, this may suggest that all the items in the measure capture only some narrow aspect of the domain of the construct. This could happen if items are merely repetitions of each other with trivial differences. In addition to item deletion on purely empirical grounds, the nature of the item needs to be taken into account. If an item has moderate correlation with the total score, yet captures some unique aspect of a construct not captured by other items, it may well be worth retaining. This is not to maximize reliability but rather to trade off reliability to increase validity, an issue to be discussed subsequently. Purely empirical or purely conceptual approaches are not sufficient in measure development and validation. The iteration between empirical results and conceptual examination is essential.

The definition and computation of coefficient alpha are discussed below (adapted from DeVellis, 1991). Exhibit 1.2 presents coefficient alpha for several versions of the PNI scale. The computation of coefficient alpha is illustrated in Appendix 1.1. A variance covariance matrix is shown for a three-item scale. Considering extreme scenarios in internal consistency is useful in clarifying the typical scenarios that fall in the middle. It is possible that none of the items covaries with each other (i.e., all nondiagonal items are zero). It is also possible that all items covary perfectly with each other (i.e., a correlation of 1 and covariances depending on the unit of measurement). Coefficient alpha is the proportion of total variance that is due to common sources. (Note the plural "sources" for subsequent discussion in factor analysis.) Variation attributable to common sources is indicated by covariances between items. The correction term in Appendix 1.1 standardizes alpha values from 0 to 1 (i.e., 0 and 1, respectively) for the extreme scenarios above. A simple example using three items with perfect intercorrelations illustrates the need for a correction term.

How does coefficient alpha measure reliability (where reliability is the minimization of random error)? Coefficient alpha is the proportion of variance attributable to common sources. These common sources are presumably the construct in question, although this remains to be demonstrated. These common sources have to be assumed to be the construct in question for the moment. Anything not attributable to common sources is assumed to be random error, that is, variation that is not systematically related to the common core that items share. The proportion of variance attributable to common sources represents the square of a correlation; hence, the unit of reliability is the square of the correlation between the scale and the true score. The unit of relevance for reliability is variance, the square of the correlation between a measure and a hypothetical true score.

A qualification is noteworthy here. Internal consistency reliability is essentially about the degree to which a set of items taps into some common sources of variance, presumably the construct being measured. The basic premise here in deleting items that are inconsistent based on their lack of correlation with the total scale is that most of the items in the scale tap into this basic construct. However, whether the scale indeed taps into the construct in question is determined subsequently through procedures for assessing validity. Hence, it is quite possible that a small number of items being deleted indeed represent the construct in question, whereas the majority of items retained tap into an unintended construct. This issue highlights the importance of examining items both conceptually and empirically in conjunction.

As the number of items in a scale increases, coefficient alpha increases. (Exhibit 1.2 reports alphas for 5-, 10-, and 20-item versions.) Why is this happening?[8] Note that this would happen even if the best five items on the basis of item-to-total correlation in the scale are used and compared with the 20-item scale. The mathematical answer to this question is that as the number of items increases, the number of covariance terms ($k^2 - k$) increases much faster than the number of variance terms (k). As the number of items increases, the number of ways in which items covary increases at an even more rapid rate (e.g., if the number of items is 2, 3, 4, 5, 6, 7, 8, 9, and 10, the number of covariance terms is 2, 6, 12, 20, 30, 42, 56, 72, and 90, respectively). But what does this mean intuitively? As the number of items increases, so does the number of ways in which they covary with each other and contribute to a total score with a high degree of variation. Each item covaries with other items and captures aspects of the domain.

During this stage of measure development, the emphasis is on tapping into common sources, presumably the underlying construct. Whether the common sources are multiple sources or a single source is in the purview of dimensionality. However, whether the common sources that are reliably measured are the construct in question is in the purview of validity. Of course, whether any aspect of the true score is captured is an issue in the realm of validity. It is possible that the items in a measure capture no aspect of a construct. In summary, internal consistency assesses whether items covary with each other or belong together and share common sources. The word *sources* is noteworthy because dimensionality addresses the question of the exact number of distinguishable sources. Hence, internal consistency is a distinct form of reliability that hinges on items sharing a common core, presumably the construct in question. Reliability procedures attempt to assess the common variance in a scale. Internal consistency assesses the degree to which items covary together or are consistent with each other. In other words, if a person has a high score on one item, that person should have a high score on another item as well, items being scored such that higher values on them denote higher values on the underlying construct. Dimensionality, as discussed subsequently, assesses whether the common core in question is, indeed, a single core, or if multiple factors reflect the hypothesized number of dimensions with items loading onto their specified factors and covarying more closely within a factor than across. A factor at an empirical level is formed by a subset or combination of items that covary more closely with each other than with other items.

Test-Retest Reliability

Internal consistency reliability is often supplemented with test-retest reliability. Typically, the same scale is administered twice with an interval of a few weeks, with recommendations ranging anywhere from 4 to 6 weeks and higher. More importantly, the interval has to fit the specific research study in terms of allowing sufficient time to minimize recall of responses in the previous administration, just as the weighing machine in the example earlier in the chapter does not have memory of the previous weighing. The logic of test-retest reliability is simply that individuals who score higher (or lower) in one administration should score higher (or lower) in the second, or vice versa. In other words, the ordering of scores should be approximately maintained. A key assumption of test-retest reliability is that the true score does not change between test and retest. For instance, in the weighing machine example, the person does not eat or change clothes before being weighed again. In a sense, test-retest reliability represents a one-to-one correspondence with the concept of reliability, which is centered on replicability. Assessment of reliability involves asking a question: If the measure were repeated, would similar (i.e., consistent) scores be obtained? Test-retest reliability offers direct evidence of such consistency. Internal consistency reliability treats different items as replications of a measure.

Scales (e.g., the PNI scale) are evaluated by examining the overall test-retest correlation. The computation of test-retest reliability is shown in Figure 1.9. As shown in the figure, test-retest reliability is, indeed, the test-retest correlation itself, the square of the hypothetical correlation between a measure and

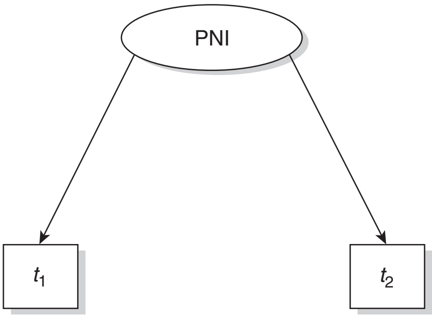


Figure 1.9 Computation of Test-Retest Reliability Formula

t_1 = Score at test at Time 1

t_2 = Score at retest at Time 2

r_{t_1T} = Correlation between score at Time 1 and true score

r_{t_2T} = Correlation between score at Time 2 and true score

$r_{t_1t_2}$ = Test-retest correlation

$r_{t_1T} \cdot r_{t_2T} = r_{t_1t_2}$

$r^2_{t_1T} = r_{t_1t_2}$

$r^2_{t_1T}$ = Proportion of variance attributable to true score

Test-retest reliability = Test-retest correlation

the true score. Individual items could also be examined by this criterion and deleted on this basis. Details for the PNI scale are presented (Exhibit 1.3). Noteworthy here is that some items may perform well on test-retest reliability but not on internal consistency, or vice versa, an issue discussed subsequently. Item 8, for example, has a high item-to-total correlation yet a low test-retest correlation for both the 1-week and 12-week intervals. This could be due to a variety of factors. Inconsistent administration across test and retest or distracting settings could cause low test-retest correlation. Similarly, item wording in Item 8, “Numbers are not necessary for most situations,” may lead to inconsistent interpretations or inconsistent responses across time depending on the “situations” that respondents recall in responding to the item. Means at test versus retest can also be examined.

Dimensionality—Exploratory Factor Analysis

Reliability through internal consistency assesses whether a set of items is tapping into a common core as measured through the degree to which they

Exhibit 1.3 Test-Retest Reliability of a Modified PNI Scale

The PNI scale was administered to 106 undergraduate students at a midwestern U.S. university twice with an interval of 1 week (Viswanathan, 1993). Longer intervals should be used, and the PNI scale has been assessed for test-retest reliability using a 12-week interval (Viswanathan, 1994).

The following are new items that replaced items from the original scale:

2. I think quantitative information is difficult to understand.

12. I enjoy thinking about issues that do not involve numerical information.

18. It is a waste of time to learn information containing a lot of numbers.

Test-Retest Correlations

	1-week interval	*12-week interval*
Total scale	.91**	.73**
Item 1	.87**	.69**
Item 2	.50**	.46**
Item 3	.74**	.52**
Item 4	.62**	.30**
Item 5	.56**	.42**
Item 6	.66**	.56**
Item 7	.65**	.66**
Item 8	.28**	.00
Item 9	.74**	.42**
Item 10	.76**	.73**
Item 11	.58**	.30**
Item 12	.56**	.31**
Item 13	.60**	.39**
Item 14	.60**	.60**
Item 15	.46**	.50**
Item 16	.64**	.45**
Item 17	.42**	.39**
Item 18	.35**	.25**
Item 19	.74**	.56**
Item 20	.60**	.54**
Mean score at test	4.64	4.64
Mean score at retest	4.62	4.56

**$p < .01$.

covary with each other. But whether the common core consists of a specific set of dimensions is in the realm of dimensionality, assessed through factor analysis. At the outset, it should be noted that there are many issues involved in using factor analysis appropriately; the references cited here provide direction for in-depth reading. This discussion aims to provide an introduction to the topic.

Factor analysis is an approach in which variables are reduced to combinations of variables, or factors (Comrey & Lee, 1992; Hair, Anderson, Tatham, & Black, 1998). The assumption in performing factor analysis on a set of items is that an underlying factor or set of factors exists that is a combination of individual items. A correlation matrix for all items in a scale may offer some indication of the outcome of factor analysis and is well worth examining. When subsets of items have sizably higher correlations among each other and lower correlations with items outside of the subset, these items are likely to form a factor. For example, if there are two distinct factors, the matrix of correlations would have two distinct subsets of items with high correlations within items from the same subset and lower correlations among items from different subsets. At an intuitive level, this suggests that if a respondent provided a high (low) rating in response to an item, he or she was more likely to provide a high (low) rating in response to another item within the subset rather than across subsets. Items in a subset or items that belong in a factor covary more closely than do items from different factors. What each factor is has to be determined by the researcher. For example, two factors may be found among items just based on whether they are positively worded or negatively worded. Responses to positively worded items may covary among each other to a much greater degree than they do with negatively worded items. A set of ratings of attributes of grocery stores (e.g., location, times of operation, aisle space, etc.) may reduce to a couple of factors such as atmosphere and convenience. Similarly, ratings of automobiles on attributes such as shoulder room, gas mileage, and so on may reduce to underlying factors such as comfort, safety, and economy. In these examples, factor analysis can be used to understand what consumers are really looking for. Thus, when appropriately used by the researcher, factor analysis can provide insight into responses at the item level (e.g., why someone likes a grocery store in terms of underlying abstract factors, such as convenience, based on responses to concrete items such as checkout speed and location). Clearly, the results of factor analysis reflect the set of items and the content covered, that is, the questions asked and the aspects on which data were collected to begin with. Potential factors are precluded if items exclude specific subdomains.

For a subset to rise to the level of a factor, both empirical outcomes and conceptual examination have to be taken into account. In a sense, when items are generated to represent the domain of a construct, each subdomain may separate into a distinct factor. At an extreme, each item could be considered a separate factor. But such an approach is both an extremely inefficient way of conducting scientific research and very unlikely to lead to conceptually rich theory and abstract generalizations. In practice, whether a set of items rises to the level of a factor has to be determined by both empirical outcomes and conceptual examination of items.

Factor analysis results include correlations or loadings between individual items and factors. If a single factor is expected, then this is likely to be the first factor, with item correlations or loadings for this factor being high. The construct in question, rather than incidental factors being measured, is likely to be dominant and explain the most variation. Individual items can be evaluated by assessing their loading or correlation with the factor considered to represent the overall construct.

Another distinction in factor analysis is in using principal component analysis and common factor analysis (Hair et al., 1998). Essentially, variance for each item can be divided into error variance, specific variance, and common variance shared with other items. In other words, responses to each item reflect some error, some content unique to the item, and some content shared with other items. An item on the PNI scale has some content unique to the item and some content, presumably preference for numerical information, shared with other items. The content unique to an item would depend on the specific context of the item and the wording. In a narrow sense, the item is a measure of some unique and likely very concrete "thing." If one item on the PNI scale is influenced by social desirability in wording and others are not, then social desirability contributes to unique content in an item that does not covary with other items. Each item also has unique content in the way it is worded. Principal components analysis uses all the variance available and is a pure data reduction technique, whereas common factor analysis uses only the common variance shared by variables and is more appropriate for conceptually driven examination of the data. The nature of factor extraction that is meaningful for scales with a common core is one that works off the common variance across items rather than one that is a pure data reduction technique (e.g., principal component analysis). This is because such common variance shared among items reflects the conceptual meaning of a factor that relates to an underlying construct. Using a pure data reduction technique that takes advantage of all variation and chance correlations is inconsistent with the goal of capturing conceptual meaning shared across items within a factor.

In performing common factor analysis, the communality of each item is estimated, which is an indicator of the variance in each item that is shared with other items (Hair et al., 1998; see also Appendix 1.1). Every item has variance, but not all of it is shared with other items or is reflected in covariance between an item and other items. Only the shared portion of an item's variance is employed in common factor analysis. Such an approach corresponds to the conceptualization of a measure as consisting of items that share content. An approach that uses all the variance in an item regardless of such variance being shared is akin to a pure data reduction technique that capitalizes on available variance rather than isolating shared variance.

The results of factor analysis include the number of factors extracted and the variance explained by each factor. Appendix 1.1 presents simple illustrations of exploratory factor analysis. In unrotated factor analysis, the first factor represents the best linear combination of items, the second factor the second best, and so on (Hair et al., 1998). The second factor should be based on the remaining variance after extraction of the first factor in order to be orthogonal to it (i.e., mathematically independent factors or no correlations between factors) (Hair et al., 1998). Unrotated factor matrices are rotated to improve interpretation of the loadings of items on factors. Rotation involves turning the axes on their origin and can lead to improved interpretation of results (Figure 1.10). Rotation redistributes variance from earlier to later factors, whereas earlier factors usually explain considerable variance in unrotated factor analysis. Rotation can be orthogonal or oblique. In oblique rotation, factors may be correlated to each other (Figure 1.10).[9] The purpose of rotation is to facilitate interpretation of factors by simplifying loadings to be closer to 1 or 0. Different types of rotations serve different purposes, each with its limitations. Varimax rotation is one such approach often used when multiple dimensions are anticipated. This type of rotation may be particularly useful when multiple dimensions are present. However, such an approach may lead to multiple factors even when there is a single underlying dimension.

In factor analysis, a judgment has to be made as to the number of meaningful factors underlying the data by comparing the percentage of variance explained by each extracted factor (reflected in eigenvalues) in light of expectations. For instance, a dominant factor would be indicated by the first factor having a much higher eigenvalue (or percent of variance explained) than the second factor. The noteworthy point here is that not all extracted factors are meaningful, and several approaches are used to assess the number of meaningful factors. In essence, the relative variances explained by each factor are compared in deciding how many factors to retain. A scree test plots the variance explained by each factor to identify discontinuities in terms of a drop-off in variance explained (Hair et al., 1998). Later factors are generally thought to contain a higher degree of unique variance (Hair et al., 1998).

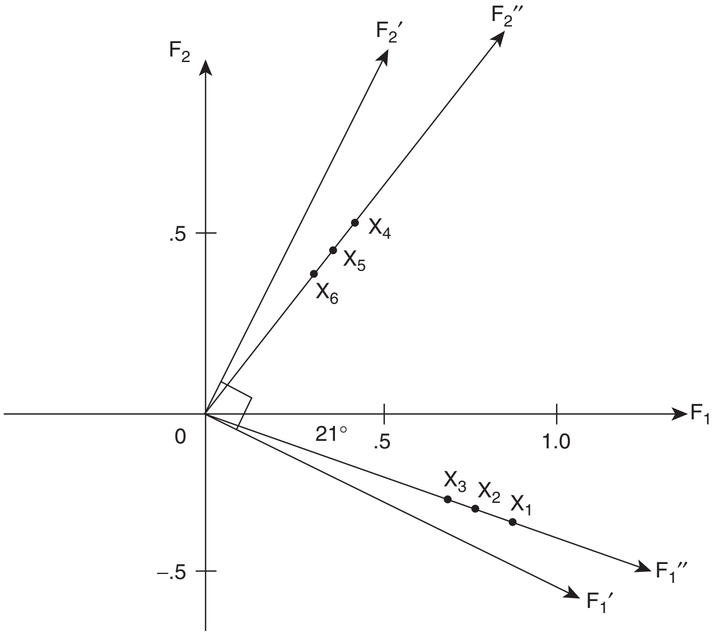


Figure 1.10 Orthogonal Factor Rotation

F_1 and F_2 = orthogonal factors (axes) before rotation;

F_1' and F_2' = orthogonal factors after varimax rotation;

F_1'' and F_2'' = oblique factors after Direct Oblimin rotation. The angle between the two is 66.42 degrees

SOURCE: Kim, J-O., & Mueller, C., *Introduction to factor analysis: What it is and how to do it,* p. 57, copyright © 1978. Reprinted by permission of Sage Publications, Inc.

The scree test plots factors in their order of extraction by eigenvalues (Figure 1.11). Such a plot is usually vertical to begin with and then becomes horizontal, the point at which this occurs being the cutoff for the maximum number of factors to extract. Another approach is to consider any factor with an eigenvalue greater than 1, suggesting that the factor explains more variance than any individual item (Hair et al., 1998).[10]

It should be noted that coefficient alpha and internal consistency relate to whether there are common underlying source*s*, the plural being the key here, and should not be used to make inferences about the dimensionality of a construct. This is illustrated using an example where there are clearly two factors, but coefficient alpha is extremely high (Figure 1.12). Whether there is one or a specific number of underlying sources is in the realm of dimensionality and factor analysis.

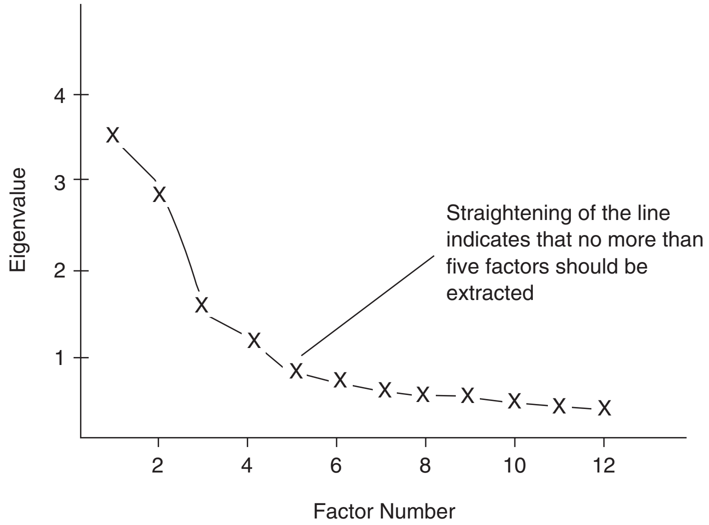


Figure 1.11 Eigenvalue Plot for Scree Test Criterion

SOURCE: Kim, J-O., & Mueller, C., *Introduction to factor analysis: What it is and how to do it,* p. 45, copyright © 1978. Reprinted by permission of Sage Publications, Inc.

Dimensionality analyses for the PNI scale are presented in Exhibit 1.4.[11] Key indicators here at this exploratory stage include the variance explained by a dominant factor and factor loadings of individual items. It should be noted that some items have medium to low loadings with the first factor, whereas other items have high loadings with the first factor and high loadings with other factors as well. What are these other factors? They could represent some content domain or some wording or methodological aspect of the measure that is shared by a subset of items. For example, items worded such that higher scores suggest higher PNI (i.e., positively worded items) may be more closely related with each other than with items worded such that lower scores suggest higher PNI (i.e., negatively worded items). Rotated factor analysis can serve to simplify matrices.

Dimensionality—Confirmatory Factor Analysis and Structural Equation Modeling

Exploratory factor analysis provides initial evidence of dimensionality, but confirmatory factor analysis is required for conclusive evidence.

Hypothetical correlation matrix for a 20-item scale with two factors comprising Items 1–10 and Items 11–20, respectively.

1	.9	.9	.9	.9	.9	.9	.9	.9	.9	0	0	0	0	0	0	0	0	0	0
.9	1	.9	.9	.9	.9	.9	.9	.9	.9	0	0	0	0	0	0	0	0	0	0
.9	.9	1	.9	.9	.9	.9	.9	.9	.9	0	0	0	0	0	0	0	0	0	0
.9	.9	.9	1	.9	.9	.9	.9	.9	.9	0	0	0	0	0	0	0	0	0	0
.9	.9	.9	.9	1	.9	.9	.9	.9	.9	0	0	0	0	0	0	0	0	0	0
.9	.9	.9	.9	.9	1	.9	.9	.9	.9	0	0	0	0	0	0	0	0	0	0
.9	.9	.9	.9	.9	.9	1	.9	.9	.9	0	0	0	0	0	0	0	0	0	0
.9	.9	.9	.9	.9	.9	.9	1	.9	.9	0	0	0	0	0	0	0	0	0	0
.9	.9	.9	.9	.9	.9	.9	.9	1	.9	0	0	0	0	0	0	0	0	0	0
.9	.9	.9	.9	.9	.9	.9	.9	.9	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	.9	1	.9	.9	.9	.9	.9	.9	.9	.9
0	0	0	0	0	0	0	0	0	0	.9	.9	1	.9	.9	.9	.9	.9	.9	.9
0	0	0	0	0	0	0	0	0	0	.9	.9	.9	1	.9	.9	.9	.9	.9	.9
0	0	0	0	0	0	0	0	0	0	.9	.9	.9	.9	1	.9	.9	.9	.9	.9
0	0	0	0	0	0	0	0	0	0	.9	.9	.9	.9	.9	1	.9	.9	.9	.9
0	0	0	0	0	0	0	0	0	0	.9	.9	.9	.9	.9	.9	1	.9	.9	.9
0	0	0	0	0	0	0	0	0	0	.9	.9	.9	.9	.9	.9	.9	1	.9	.9
0	0	0	0	0	0	0	0	0	0	.9	.9	.9	.9	.9	.9	.9	.9	1	.9
0	0	0	0	0	0	0	0	0	0	.9	.9	.9	.9	.9	.9	.9	.9	.9	1

$$\alpha = \frac{K}{K-1} \cdot \left(\frac{K\bar{r}}{1 + (K-1) \cdot \bar{r}} \right)$$

$$\bar{r} = \frac{100 \times 0 + 90 \times .9}{190} = 0.426$$

$$\alpha = \frac{20}{20-1} \cdot \left(\frac{20 \cdot (0.426)}{1 + 19 \cdot (0.426)} \right)$$

$$= 0.986$$

Figure 1.12 Coefficient α Versus Factor Analysis

Exploratory factor analysis assumes that all items are related to all factors, whereas confirmatory factor analysis imposes a more restrictive model where items have prespecified loadings with certain factors that also may be

Exhibit 1.4 Exploratory Factor Analysis of the 20-Item PNI Scale

Results for the original 20-item PNI scale are presented below for the same sample as in Exhibit 1.2 (Viswanathan, 1993). Correlations between items provide clues to the likely factor structure that may underlie the data. Below, the factor matrix is presented and is the result of exploratory factor analysis. The correlation between each item and each factor is presented here. A useful way to examine such a matrix is by highlighting the high loadings that each item has on one or more factors. Here, many items have their highest loading on the first factor.

Factor Matrix

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*
N1	.79874	.13506	–.01212	–.09337
N2	.36629	.25747	.27185	.10054
N3	.77987	.27999	–.14529	–.12853
N4	.63339	–.22174	–.29342	.00671
N5	.54453	–.02839	–.03588	.56900
N6	.63067	.21084	.10678	–.00046
N7	.81937	.07605	.03373	.28493
N8	.47641	–.32387	.11877	.01884
N9	.45266	.15707	.05986	.10623
N10	.72986	.21385	–.33176	–.08889
N11	.57628	–.47323	–.27856	.06503
N12	–.31912	.17510	.16366	.10411
N13	.55337	–.24397	.05221	–.23473
N14	.67780	.11840	.09128	–.27664
N15	.51572	–.28529	.30472	–.08219
N16	.67509	–.29761	.00241	–.14682
N17	.57455	–.12036	.56461	.04140
N18	.42251	.15295	.41612	–.05502
N19	.58197	.44522	–.22408	–.06153
N20	.66761	–.08386	–.17123	.13250

The communality of each variable, or the variance it shares with other variables, is reported below.

Variable	*Communality*
N1	.66510
N2	.28447
N3	.72421
N4	.53649
N5	.62237
N6	.45359

(Continued)

Exhibit 1.4 (Continued)

N7	.75948
N8	.34632
N9	.24444
N10	.69640
N11	.63786
N12	.17012
N13	.42356
N14	.55830
N15	.44697
N16	.56587
N17	.66510
N18	.37809
N19	.59091
N20	.49961

Eigenvalues and percent of variance explained by each factor is presented below. Such information is used to decide on the number of meaningful factors as discussed in the chapter.

Factor	*Eigenvalue*	*Pct. of Var.*	*Cum. Pct.*
1	7.32600	36.6	36.6
2	1.17967	5.9	42.5
3	1.13261	5.5	48.0
4	0.66097	3.3	51.3

related to each other. Exploratory factor analysis uses a general model *no matter what the substantively motivated constraints are.* Confirmatory factor analysis allows more precise specification of the relationship between items and factors.

It should be emphasized strongly that, although exploratory factor analysis can be used in preliminary stages of measure development, it should be followed up with confirmatory factor analysis (Gerbing & Anderson, 1988). Confirmatory factor analysis tests carefully specify models of the relationship between items and factors (Gerbing & Anderson, 1988). Confirmatory factor analysis also provides overall indexes of fit between the proposed model and the data, which range in value from 0 to 1. Exploratory factor analysis employs more of a shotgun approach by allowing all items to be related with all factors (Figure 1.13).

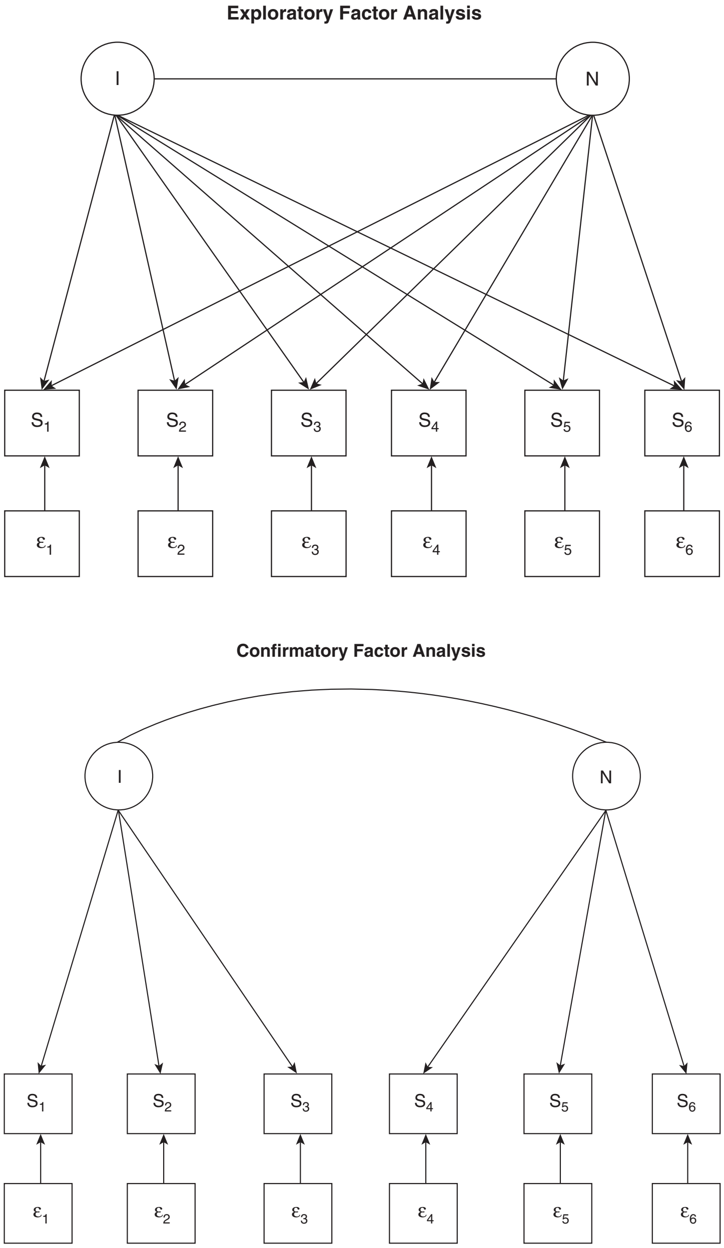


Figure 1.13 Confirmatory Factor Analysis Versus Exploratory Factor Analysis for the Speech Quality Scale

NOTES: I = Intelligibility factor of speech quality; N = Naturalness factor of speech quality; $S_{1–3}$ = Items for intelligibility; $S_{4–6}$ = Items for naturalness.

Whereas exploratory factor analysis considers internal consistency among items of a measure, confirmatory factor analysis considers external consistency across items of different measures or dimensions (Gerbing & Anderson, 1988). As shown in Figure 1.13 and illustrated in Appendix 1.1, the relationship between items from different dimensions is exclusively through the relationship between the dimensions. Therefore, external consistency is assessed by comparing the observed correlation between two items of different dimensions or constructs with the predicted correlation that arises out of the hypothesized relationship between items and measures. Confirmatory factor analysis applies the criterion of external consistency wherein relationships between items across factors are assessed (Appendix 1.1). In essence, the observed correlations between items are compared to the hypothesized correlations in light of the specified model. Moreover, overall fit indexes can be computed with confirmatory factor analysis. Residuals between items denote the degree to which observed relationships deviate from hypothesized relationships. A positive residual between two items suggests that the model underpredicts the relationship between two items and vice versa.

Confirmatory factor analysis is demonstrated using a speech quality scale (Exhibit 1.5). Confirmatory factor analysis allows isolation of items that measure multiple factors that are of substantive importance (see Appendix 1.1 for simplified illustration and Figures 1.13 and 1.14). In a multidimensional measure, each item ideally should measure only one dimension. Two items measuring different dimensions may have an unduly large relationship, suggesting that they are influenced by a common factor, perhaps one of the factors being measured or a different factor. If only one item is influenced by an additional unknown factor, that is usually tolerable. In fact, all items will probably have some unique variance. However, items that measure more than one substantive factor have to be identified and deleted. For instance, in Figure 1.14, reproduced from Gerbing and Anderson (1988), Items 4 and 7 measure more than one factor.

Confirmatory factor analysis falls under the broader umbrella of structural equation modeling (SEM). The caution about appropriate use of statistical techniques and the introductory nature of the material covered here should be reemphasized for structural equation modeling. Structural equation modeling combines econometric and psychometric approaches by simultaneously assessing structural and measurement models; the former deals with relationships between constructs, whereas the latter deals with relationships between constructs and their measures (Figure 1.15). In a typical econometric approach, such as regression, each measure is entered into the analysis without accounting for measurement error. For example,

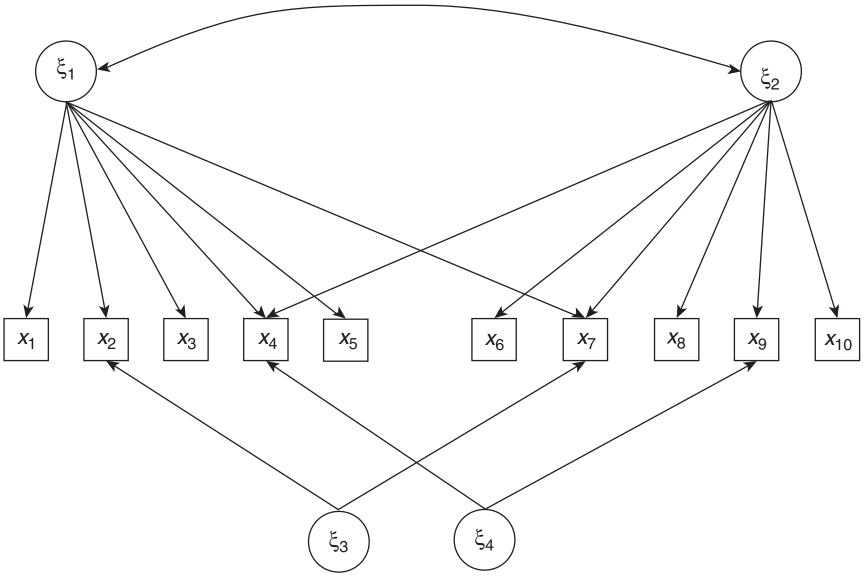


Figure 1.14 Example of Confirmatory Factor Analysis

SOURCE: Adapted from Gerbing, D. W., & Anderson, J. C. An updated paradigm for scale development incorporating unidimensionality and its assessment, in *Journal of Marketing Research*, *25*(5), pp. 186–92, copyright © 1988. Reprinted by permission of the American Marketing Association.

NOTE: ξ_1 and ξ_2 are two moderately correlated factors. x_{1-5} and x_{6-10} are indicators for ξ_1 and ξ_2, respectively. ξ_3 and ξ_4 are additional factors that provide a source of common covariance for two pairs of items across two sets.

considerable random error in one or more measures in regression may decrease the observed relationship between two measures. Nevertheless, the observed relationship is assumed to reflect the true relationship between constructs. SEM combines the econometric and psychometric traditions to evaluate relationships between constructs while accounting for measurement error. At the measurement level, the advantage of SEM is in terms of specifying a precise model and testing it using confirmatory factor analysis (Figure 1.16). At the theoretical level, SEM allows assessment of relationships between constructs while accounting measurement error. For instance, as discussed earlier, random error and unreliability reduce the ability of a measure to correlate with other measures. SEM estimates the relationship between two constructs while taking into account the degree of reliability of their measures.

(Text continues on page 58)

Exhibit 1.5 Confirmatory Factor Analysis on the Speech Quality Scale

The quality of text-to-speech (TTS) systems can be assessed effectively only on the basis of reliable and valid listening tests. The method for these tests must be rigorous in voice sample presentation, respondent selection, and questionnaire preparation. These tests involve preparing several samples of synthesized output from multiple TTS systems, randomizing the system-sentence combinations, and asking listeners to score each output audio. Johnston (1996) notes that opinion tests of speech quality are the basis of speech quality assessment. The Mean Opinion Scale (MOS) has been the recommended measure of text-to-speech quality (ITU-T Recommendation, 1994). It consists of seven 5-point scales that assess overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate, and pleasantness (Lewis, 2001). Items in this scale are presented in the Exhibit 1.5 Figure.

Past research lacks explication of the factors or dimensions of speech quality. Moreover, past research employed exploratory factor analysis to assess factor structure, a procedure appropriate at preliminary stages of measure development that needs to be followed up with testing using confirmatory factor analysis. An item in the MOS asks the respondent to rate the overall quality of the synthesized speech clip on a scale from 1 to 5 (Exhibit 1.5 Figure). The other items relate to various aspects of synthetic speech such as listening effort, pronunciation, speed, pleasantness, naturalness, audio flow, ease of listening, comprehension, and articulation (Exhibit 1.5 Figure). Responses are gathered on the 5-point scales with appropriate phrase anchors. The MOS combines an item on overall sound quality with other items that are more specific and relate to different facets of speech quality.

Several issues are noteworthy with respect to the MOS. At a conceptual level, a central issue is the factor structure of the domain of speech quality. A factor is essentially a linear combination of variables (Hair et al., 1998). In this context, factor analysis is conducted to assess the number of factors that are extracted and to assess the degree to which items correlate with specific factors. In terms of dimensionality, a variety of different results have been reported using exploratory factor analysis, including two factors referred to as intelligibility and naturalness and a separate speaking rate item (Kraft & Portele, 1995; Lewis, 2001) and one factor (Sonntag, Portele, Haas, & Kohler, 1999). More recently, Lewis suggested a revised version of the MOS with modified 7-point response scales. Results suggested two factors, with the speaking rate item loading on the intelligibility factor. Moreover, past research has typically employed exploratory factor analysis and has not followed up with subsequent confirmatory factor analysis, as recommended in the psychometric literature (Gerbing & Anderson, 1988). Confirmatory factor analysis offers a test of factor structure by testing specific models and providing overall indexes of fit.

Also lacking in past research is an explication of the domain of the speech quality construct through a description of underlying factors such as intelligibility. Such conceptual examination should ideally precede item generation and empirical assessment. Intelligibility is related to the extent to which words and sentences can be

(Continued)

Exhibit 1.5 Figure Speech Quality Scale*

Overall Impression	Listening Effort	Pronunciation
How do you rate the quality of the audio you just heard? ○ Excellent ○ Good ○ Fair ○ Poor ○ Very poor	How would you describe the effort you were required to make in order to understand the message? ○ Complete relaxation possible; no effort required ○ Attention necessary; no appreciable effort required ○ Moderate effort required ○ Considerable effort required ○ No meaning understood with any feasible effort	Did you notice anomalies in pronunciation? ○ No ○ Yes, but not annoying ○ Yes, slightly annoying ○ Yes, annoying ○ Yes, very annoying

Speaking Rate	Pleasantness	Naturalness	Audio Flow
The average speed of delivery was: ○ Just right ○ Slightly fast or slightly slow ○ Fairly fast or fairly slow ○ Very fast or very slow ○ Extremely fast or extremely slow	How would you describe the pleasantness of the voice? ○ Very pleasant ○ Pleasant ○ Neutral ○ Unpleasant ○ Very unpleasant	How would you rate the naturalness of the audio? ○ Very natural ○ Natural ○ Neutral ○ Unnatural ○ Very unnatural	How would you describe the continuity or flow of the audio? ○ Very smooth ○ Smooth ○ Neutral ○ Discontinuous ○ Very discontinuous

Ease of Listening	Comprehension Problems	Articulation	Acceptance
Would it be easy or difficult to listen to this voice for long periods of time? ○ Very easy ○ Easy ○ Neutral ○ Difficult ○ Very difficult	Did you find certain words hard to understand? ○ Never ○ Rarely ○ Occasionally ○ Often ○ All of the time	Were the sounds in the audio distinguishable? ○ Very clear ○ Clear ○ Neutral ○ Less clear ○ Much less clear	Do you think that this voice can be used for an interactive telephone or wireless hand-held information service system? ○ Yes ○ No

*Items 1 to 9 of the scale described in the text refer to the consecutive items from listening effort to articulation.

(Continued)

understood; therefore, items should tap into factors that assess listening effort, pronunciation, speaking rate, comprehension problems, and articulation (Exhibit 1.5 Figure). These are aspects of speech that contribute to intelligibility. Therefore, contrary to results that suggest that speaking rate is a separate factor, it belongs conceptually in the intelligibility factor. Naturalness relates to the degree to which speech is similar to natural human speech; hence, items such as naturalness, ease of listening, pleasantness, and audio flow are relevant (see Exhibit 1.5 figure). These are impressions about the speech and the feeling it engenders in respondents. This should be contrasted with specific aspects of respondents' cognition, such as speaking rate, listening effort, and pronunciation. Thus, conceptually, intelligibility and naturalness relate to specific cognitions about aspects of the speech versus broader impressions and feelings about the speech, respectively. Central here from a procedural viewpoint is the importance of explicating the domain of speech quality prior to testing through confirmatory factor analysis.

Another issue with the MOS is the inclusion of an item that is global in nature—assessing overall speech quality—with other items that are more specific to aspects of speech quality, such as articulation and pronunciation. Such an approach is problematic; the scale should consist of either global items or specific items. The global

Exhibit 1.5 Table Results for Confirmatory Factor Analysis on Speech Quality Scale

Dataset	*1*		*2*		*3*		*4*		*5*		*6*	
No. of factors	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*
n	128	128	128	128	128	128	128	128	128	128	128	128
df	27	26	27	26	27	26	27	26	27	26	27	26
Chi-square	72.8	39.1*	49.5	38.3^*	118.2	57.6*	102.8	64.3*	88.0	52.8*	78.3	49.2*
NFI	0.95	0.97	0.96	0.97	0.90	0.95	0.90	0.94	0.92	0.95	0.94	0.97
NNFI	0.96	0.99	0.98	0.99	0.89	0.96	0.90	0.95	0.92	0.96	0.95	0.98
CFI	0.97	0.99	0.98	0.99	0.92	0.97	0.92	0.96	0.94	0.97	0.96	0.98
IFI	0.97	0.99	0.98	0.99	0.92	0.97	0.92	0.96	0.94	0.97	0.96	0.98
SRMR	0.06	0.04	0.05	0.04	0.08	0.06	0.07	0.06	0.07	0.05	0.05	0.05

NOTES: ^ $p > .05$; $p < .05$ for all other chi-square values. *Chi-square values of 1- versus 2-factor models significantly different at .05 level. *n* = sample size; *df* = degrees of freedom; NFI = normed fit index; NNFI = non-normed fit index; CFI = comparative fit index; IFI = incremental fit index; SRMR = standardized root mean square residual. The results for the 3-factor model (*df* = 24) with speaking rate as a separate item were identical to those for the 2-factor model with the following exceptions: Dataset 1: Chi-square = 39.0, NNFI = 0.98; Dataset 2: Chi-square = 36.5; Dataset 3: Chi-square = 57.4; Dataset 4: Chi-square = 64.2, NNFI = 0.94; Dataset 6: Chi-square = 47.6, NNFI = 0.97.

approach essentially asks for overall impressions of speech quality, whereas the specific approach has items representing different aspects of speech quality. For example, if different items relate to different factors such as intelligibility and naturalness, then a global item would relate to both factors, being broader than either. In fact, this argument is supported by inconsistent results that have been obtained. Although this global item was thought to belong to the naturalness factor, Lewis (2001) unexpectedly found that it related to the intelligibility factor. All of these issues were examined through confirmatory factor analysis.

One hundred and twenty-eight employees of a U.S. company rated six systems on the single item on overall speech quality, the 9-item speech quality scale, and the single item on system acceptability. Because each respondent rated more than one system and the purpose here was to generate independent observations within a dataset, each dataset analyzed related to independent responses to a particular system. Moreover, some systems were slightly different across subsets of respondents. Therefore, the first step was to identify datasets of responses to exactly the same system. This led to four data sets of 64 respondents, each rating identical systems. These four datasets provided the best quality of data of independent observations of identical systems. Six more datasets provided 128 independent observations; however, the systems rated varied slightly across respondents. Sample sizes for these datasets met some of the criteria in past research for factor analysis (i.e., greater than five times the number of items [> 45], or greater than 100 in the case of the larger datasets [Iacobucci, 1994]), although more stringent criteria have been suggested as a function of factors such as the communality of items (MacCallum, Widaman, Zhang, & Hong, 1999). Given the high communality of items of speech quality, less stringent criteria appear to be appropriate here.

Several models were tested through confirmatory factor analysis. Several points are noteworthy with the overwhelming pattern of results. First, the overall levels of fit of both 1- and 2-factor models (i.e., Items 1, 2, 3, 8, and 9 on intelligibility and Items 4, 5, 6, and 7 on naturalness in Exhibit 1.5 Figure) are satisfactory by accepted (e.g., > 0.90) (Bagozzi & Yi, 1988; Bentler & Bonnet, 1980) and even by conservative norms (e.g., > 0.95) (Hu & Bentler, 1998). All individual items had significant loadings on hypothesized factors. Whereas the 2-factor model improved on the 1-factor model, the 1-factor model had a high level of fit. These results suggest that both 1- and 2-factor formulations of the 9-item scale are strongly supported through confirmatory factor analysis. Contrary to the notion that the item on speaking rate loads on a separate factor, here, speaking rate appears to be a part of the intelligibility factor. An alternative model with speaking rate as a separate factor did not improve on the 2-factor fit indexes, although fit levels were so high as to allow little or no room for improvement. A model with speaking rate loading on the intelligibility factor led to superior fit as opposed to a model with speaking rate loading on the naturalness factor, consistent with the argument that speaking rate loads primarily on the intelligibility factor.

(Continued)

(Continued)

Exploratory factor analysis is more of a shotgun approach, where all items can be related to all factors. Confirmatory factor analysis can be used to carefully specify the relationship between items and factors, and between factors. Overall fit indexes can be computed, and alternative models can be compared.

Input and Output From LISREL 8.5 for Confirmatory Factor Analysis

The program statements for confirmatory factor analysis are shown below. The first line states the number of input variables, the number of observations, the number of factors, and the nature of the matrix (i.e., a covariance matrix). The second line specifies the labels for the variables. This is followed by the covariance matrix and the model statement specifying nine variables, one factor, and the label for the factor.

```
DA NI = 9 MA = KM NO =128
LA

s1 s2 s3 s4 s5 s6 s7 s8 s9
KM

1.000000
0.595594  1.000000
0.478988  0.375190  1.000000
0.636631  0.448744  0.401614  1.000000
0.614881  0.567760  0.348624  0.756280  1.000000
0.618288  0.546916  0.410717  0.642101  0.722303  1000000
0.647556  0.523697  0.398602  0.783353  0.798083  0.731203
1.000000
0.582192  0.534391  0.378980  0.398120  0.349791  0.401245
0.394849  1.000000
0.628814  0.561461  0.359818  0.647491  0.651060  0.624736
0.663076  0.483979  1.000000

MO NX = 9 NK = 1 LX = FR PH = ST
LK
sq

OU
```

Edited results are shown below.

```
DA NI = 9 NO = 128 NG = 1 MA = CM
SE
1 2 3 4 5 6 7 8 9 /
MO NX = 9 NK = 1 LX = FU, FI PH = SY, FR TD = DI, FR
LK
sq
FI PH(1,1)
FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1) LX(7,1) LX(8,1) LX(9,1)
VA 1.00 PH(1,1)
PD
OU ME = ML RS XM

TI DA NI = 9 NO = 128

Number of Input Variables 9
Number of Y - Variables 0
Number of X - Variables 9
Number of ETA - Variables 0
Number of KSI - Variables 1
Number of Observations 128

TI DA NI = 9 NO = 128

Covariance Matrix

            s1         s2         s3         s4         s5         s6
        ---------- ---------- ---------- ---------- ---------- ----------
s1        1.00
s2        0.60       1.00
s3        0.48       0.38       1.00
s4        0.64       0.45       0.40       1.00
s5        0.61       0.57       0.35       0.76       1.00
s6        0.62       0.55       0.41       0.64       0.72       1.00
s7        0.65       0.52       0.40       0.78       0.80       0.73
s8        0.58       0.53       0.38       0.40       0.35       0.40
s9        0.63       0.56       0.36       0.65       0.65       0.62

Covariance Matrix

            s7         s8         s9
        ---------- ---------- ----------
s7        1.00
s8        0.39       1.00
s9        0.66       0.48       1.00
```

(Continued)

(Continued)

```
TI DA NI = 9 NO = 128

Parameter Specifications

LAMBDA-X

             sq
         ----------
s1           1
s2           2
s3           3
s4           4
s5           5
s6           6
s7           7
s8           8
s9           9

THETA-DELTA

             s1         s2         s3         s4         s5         s6
         ---------- ---------- ---------- ---------- ---------- ----------
             10         11         12         13         14         15

THETA-DELTA

             s7         s8         s9
         ---------- ---------- ----------
             16         17         18

TI DA NI = 9 NO = 128

Number of Iterations = 9

LISREL Estimates (Maximum Likelihood)
Loadings of each item on the factor are listed below with associated standard errors
and t-values.

LAMBDA-X

             sq
         ----------
s1          0.77
           (0.08)
           10.11
```

```
s2        0.65
        (0.08)
          8.08

s3        0.48
        (0.09)
          5.63

s4        0.84
        (0.07)
         11.54

s5        0.87
        (0.07)
         12.13

s6        0.81
        (0.07)
         10.94

s7        0.89
        (0.07)
         12.61

s8        0.52
        (0.08)
          6.15

s9        0.77
        (0.08)
         10.19

PHI

           sq
       ----------
          1.00

THETA-DELTA

           s1         s2         s3         s4         s5         s6
       ---------- ---------- ---------- ---------- ---------- ----------
          0.41       0.57       0.77       0.29       0.25       0.34
        (0.06)     (0.08)     (0.10)     (0.04)     (0.04)     (0.05)
          7.21       7.58       7.81       6.71       6.38       6.96
```

(Continued)

(Continued)

```
THETA-DELTA

        s7         s8         s9
    ---------- ---------- ----------
      0.21       0.73       0.40
     (0.04)     (0.09)     (0.06)
      6.02       7.78       7.19

Squared Multiple Correlations for X - Variables

        s1         s2         s3         s4         s5         s6
    ---------- ---------- ---------- ---------- ---------- ----------
      0.59       0.43       0.23       0.71       0.75       0.66

Squared Multiple Correlations for X - Variables

        s7         s8         s9
    ---------- ---------- ----------
      0.79       0.27       0.60

Various goodness-of-fit indexes are presented below.
Goodness of Fit Statistics

Degrees of Freedom = 27
Minimum Fit Function Chi-Square = 72.82 (P = 0.00)
Normal Theory Weighted Least Squares Chi-Square = 87.95 (P = 0.00)
Estimated Non-Centrality Parameter (NCP) = 60.95
90 Percent Confidence Interval for NCP = (36.28 ; 93.23)

Minimum Fit Function Value = 0.57
Population Discrepancy Function Value (F0) = 0.48
90 Percent Confidence Interval for F0 = (0.29 ; 0.73)
Root Mean Square Error of Approximation (RMSEA) = 0.13
90 Percent Confidence Interval for RMSEA = (0.10 ; 0.16)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00

Expected Cross-Validation Index (ECVI) = 0.98
90 Percent Confidence Interval for ECVI = (0.78 ; 1.23)
ECVI for Saturated Model = 0.71
ECVI for Independence Model = 6.10

Chi-Square for Independence Model with 36 Degrees of Freedom = 756.24
Independence AIC = 774.24
Model AIC = 123.95
Saturated AIC = 90.00
```

```
Independence CAIC = 808.91
Model CAIC = 193.29
Saturated CAIC = 263.34

Normed Fit Index (NFI) = 0.90
Non-Normed Fit Index (NNFI) = 0.92
Parsimony Normed Fit Index (PNFI) = 0.68
Comparative Fit Index (CFI) = 0.94
Incremental Fit Index (IFI) = 0.94
Relative Fit Index (RFI) = 0.87

Critical N (CN) = 82.91

Root Mean Square Residual (RMR) = 0.061
Standardized RMR = 0.061
Goodness of Fit Index (GFI) = 0.87
Adjusted Goodness of Fit Index (AGFI) = 0.78
Parsimony Goodness of Fit Index (PGFI) = 0.52

TI DA NI =  9 NO = 128

Fitted Covariance Matrix

           s1         s2         s3         s4         s5         s6
       ---------- ---------- ---------- ---------- ---------- ----------
s1        1.00
s2        0.50       1.00
s3        0.37       0.32       1.00
s4        0.65       0.55       0.41       1.00
s5        0.67       0.57       0.42       0.73       1.00
s6        0.62       0.53       0.39       0.68       0.70       1.00
s7        0.68       0.58       0.43       0.74       0.77       0.72
s8        0.40       0.34       0.25       0.44       0.45       0.42
s9        0.60       0.50       0.37       0.65       0.67       0.63

Fitted Covariance Matrix

           s7         s8         s9
       ---------- ---------- ----------
s7        1.00
s8        0.46       1.00
s9        0.69       0.40       1.00
```

(Continued)

(Continued)

```
Fitted Residuals

           s1         s2         s3         s4         s5         s6
       ---------- ---------- ---------- ---------- ---------- ----------
s1        0.00
s2        0.09       0.00
s3        0.11       0.06       0.00
s4       –0.01      –0.10       0.00       0.00
s5       –0.05       0.00      –0.07       0.03       0.00
s6       –0.01       0.02       0.02      –0.04       0.02       0.00
s7       –0.03      –0.06      –0.03       0.04       0.03       0.01
s8        0.18       0.19       0.13      –0.04      –0.10      –0.02
s9        0.03       0.06      –0.01       0.00      –0.02       0.00

Fitted Residuals

           s7         s8         s9
       ---------- ---------- ----------
s7        0.00
s8       –0.07       0.00
s9       –0.02       0.08       0.00

Summary Statistics for Fitted Residuals

Smallest Fitted Residual = –0.10
Median Fitted Residual = 0.00
Largest Fitted Residual = 0.19

Stemleaf Plot

– 1|00
– 0|7765
– 0|4433222111000000000000000
0|12223334
0|6689
1|13
1|89
```

The standardized residuals between pairs of variables are useful to examine. For example, the residual between Items 1 and 8 is positive and high, whereas the residual between Items 2 and 4 is negative and high. It should be noted that Items 4, 5, 6, and 7 form the naturalness factor, and Items 1, 2, 3, 8, and 9 form the intelligibility factor. In a 1-factor model, these residuals are consistent with two underlying factors. Items 1, 2, 3, 8, and 9 tend to have large negative residuals or small residuals with Items 4, 5, 6, and 7. Residuals within each set of items tend to be positive.

```
Standardized Residuals

           s1        s2        s3        s4        s5        s6
        --------  --------  --------  --------  --------  --------
s1         —
s2       2.37        —
s3       2.30      1.06        —
s4      –0.37     –3.08     –0.11        —
s5      –2.25      0.08     –2.05      1.57        —
s6      –0.22      0.49      0.45     –1.71      0.94        —
s7      –1.70     –2.14     –0.97      2.40      2.11      0.64
s8       4.00      3.52      1.96     –1.07     –3.07     –0.53
s9       1.05      1.45     –0.30     –0.09     –0.85     –0.11

Standardized Residuals

           s7        s8        s9
        --------  --------  --------
s7         —
s8      –2.27        —
s9      –1.14      1.80        —

Summary Statistics for Standardized Residuals

Smallest Standardized Residual = –3.08
Median Standardized Residual = 0.00
Largest Standardized Residual = 4.00

Stemleaf Plot

– 3|11
– 2|3311
– 1|77110
– 0|85432111000000000
0|14569
1|11468
2|01344
3|5
4|0
Largest Negative Standardized Residuals
Residual for s4 and s2 –3.08
Residual for s8 and s5 –3.07
Largest Positive Standardized Residuals
Residual for s8 and s1 4.00
Residual for s8 and s2 3.52
```

(Continued)

(Continued)

Edited results for a 2-factor model are presented below for the same data.

```
MO NX = 9 NK = 2 PH = ST

LK

sq1 sq2

FR LX(1,1) LX(2,1) LX(3,1) LX(4,2) LX(5,2) LX(6,2) LX(7,2) LX(8,1) LX(9,1)

OU

LISREL Estimates (Maximum Likelihood)

LAMBDA-X

           sq1          sq2
        ----------   ----------
s1         0.84          —
          (0.07)
          11.22

s2         0.72          —
          (0.08)
           8.97

s3         0.53          —
          (0.09)
           6.09

s4          —           0.85
          (0.07)
          11.74

s5          —           0.88
          (0.07)
          12.47

s6          —           0.81
          (0.07)
          10.85

s7          —           0.91
          (0.07)
          13.04

s8         0.63          —
          (0.08)
           7.59
```

```
s9       0.79        —
        (0.08)
        10.26

PHI

   sq1       sq2
---------- ----------
   sq1      1.00

sq2      0.86        1.00
        (0.04)
        24.26

THETA-DELTA

           s1         s2         s3         s4         s5         s6
       ---------- ---------- ---------- ---------- ---------- ----------
          0.30       0.48       0.72       0.28       0.22       0.35
         (0.05)     (0.07)     (0.09)     (0.04)     (0.04)     (0.05)
          5.60       6.93       7.60       6.47       5.93       6.90

THETA-DELTA

           s7         s8         s9
       ---------- ---------- ----------
          0.18       0.60       0.38
         (0.03)     (0.08)     (0.06)
          5.31       7.32       6.33

Goodness of Fit Statistics

Degrees of Freedom = 26
Minimum Fit Function Chi-Square = 39.10 (P = 0.048)
Normal Theory Weighted Least Squares Chi-Square = 38.56 (P = 0.054)
Estimated Non-Centrality Parameter (NCP) = 12.56
90 Percent Confidence Interval for NCP = (0.0 ; 33.28)

Minimum Fit Function Value = 0.31
Population Discrepancy Function Value (F0) = 0.099
90 Percent Confidence Interval for F0 = (0.0 ; 0.26)
Root Mean Square Error of Approximation (RMSEA) = 0.062
90 Percent Confidence Interval for RMSEA = (0.0 ; 0.10)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.30
```

(Continued)

(Continued)

```
Expected Cross-Validation Index (ECVI) = 0.60
90 Percent Confidence Interval for ECVI = (0.50 ; 0.77)
ECVI for Saturated Model = 0.71
ECVI for Independence Model = 6.10

Chi-Square for Independence Model with 36 Degrees of Freedom = 756.24
Independence AIC = 774.24
Model AIC = 76.56
Saturated AIC = 90.00
Independence CAIC = 808.91
Model CAIC = 149.74
Saturated CAIC = 263.34

Normed Fit Index (NFI) = 0.95
Non-Normed Fit Index (NNFI) = 0.97
Parsimony Normed Fit Index (PNFI) = 0.68
Comparative Fit Index (CFI) = 0.98
Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.93

Critical N (CN) = 149.23

Root Mean Square Residual (RMR) = 0.043
Standardized RMR = 0.043
Goodness of Fit Index (GFI) = 0.94
Adjusted Goodness of Fit Index (AGFI) = 0.89
Parsimony Goodness of Fit Index (PGFI) = 0.54

Summary Statistics for Fitted Residuals

Smallest Fitted Residual = –0.13
Median Fitted Residual = 0.00
Largest Fitted Residual = 0.08

Stemleaf Plot

–12|0
–10|
–8|9
–6|75
–4|725
–2|9753
–0|638764210000000000
0|6936
```

```
2|2246
4|457711
6|95
8|0
Standardized Residuals

           s1        s2        s3        s4        s5        s6
       ---------- ---------- ---------- ---------- ---------- ----------
s1        —
s2      –0.33       —
s3       1.11     –0.08       —
s4       0.75     –2.00      0.33       —
s5      –0.86      0.63     –1.18      0.43       —
s6       1.02      1.12      0.86     –2.06      0.53       —
s7      –0.32     –1.14     –0.31      1.09     –0.18     –0.08
s8       1.81      1.91      0.83     –1.50     –3.28     –0.82
s9      –1.91     –0.22     –1.45      2.06      1.69      2.01

Standardized Residuals

           s7        s8        s9
       ---------- ---------- ----------
s7        —
s8      –2.66       —
s9       1.71     –0.46       —

Summary Statistics for Standardized Residuals

Smallest Standardized Residual = –3.28
Median Standardized Residual = 0.00
Largest Standardized Residual = 2.06

Stemleaf Plot

– 3|3
– 2|710
– 1|95421
– 0|98533322110000000000
0|3456889
1|01117789
2|01
Largest Negative Standardized Residuals
Residual for s8 and s5 –3.28
Residual for s8 and s7 –2.66
```

EXHIBIT SOURCE: Adapted from *Computer Speech and Language, 19*(1), Viswanathan, M., & Viswanathan, M., Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale, pp. 55–83, Copyright 2005. Reprinted with permission from Elsevier.

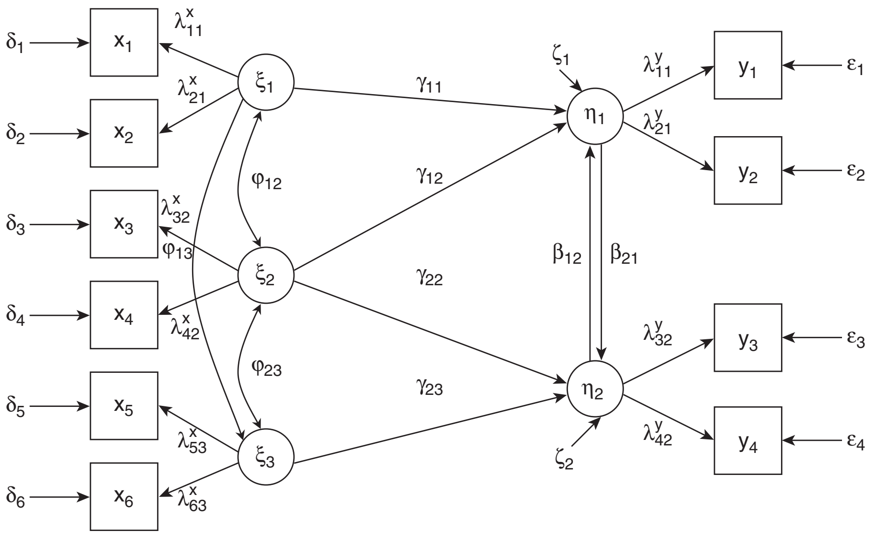


Figure 1.15 Combined Measurement Component and Structural Component of the Covariance Structure Model

SOURCE: Long, J. S., *Covariance structure models: An introduction to LISREL*, p. 18, copyright © 1983. Reprinted by permission of Sage Publications, Inc.

NOTES: The measurement model for independent variables consists of the following:
ξ = latent variable
x = observed variable
δ = unique factor or error
λ = loading of x on ξ
$x_1 = \lambda^x_{11}\xi_1 + \delta_1$.
The measurement model for dependent variables consists of the following:
η = latent variable
y = the observed variable
ε = unique factor or error
λ = loading of y on η
$y_1 = \lambda^y_{11}\eta_1 + \varepsilon_1$.
The structural model relating independent to dependent latent variables consists of the following:
η related to ξ by γ, with error represented by ζ
$\eta_1 = \gamma_{11}\xi_1 + \gamma_{12}\xi_1 + \beta_{12}\eta_2 + \zeta_1$
The xs are independent observed variables related to the dependent variables by the slope coefficients β_1 and β_2.

As illustrated in Figures 1.15 and 1.16, and as explained in the adjoining notations and equations, the structural model relates latent variables to each other, whereas the measurement model relates to the operationalizations of

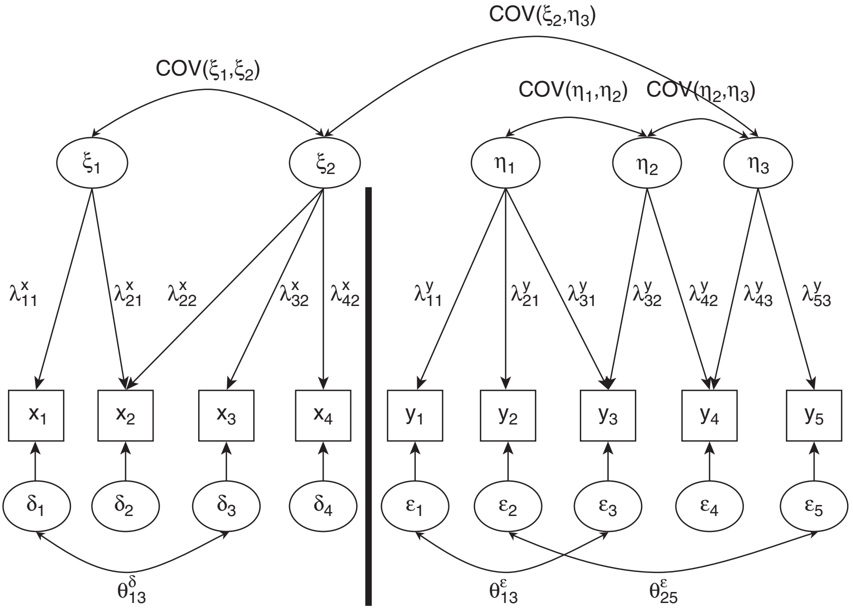


Figure 1.16 The Measurement Component of the Covariance Structure Model

SOURCE: Long, J. S., *Covariance structure models: An introduction to LISREL*, p. 18, copyright © 1983. Reprinted by permission of Sage Publications, Inc.

latent variables through observed variables. In specifying the measurement model, items could be aggregated into parcels or subsets. Method factors could be specified. For instance, if specific methods are used to collect data on specific items, they could be incorporated into the model as separate factors in addition to the latent variables. For example, in Figure 1.14, ξ_3 could be considered as a method factor for items x_2 and x_7. Thus, the effect of specific methods is explicitly incorporated. Correlated errors between items also can be specified, as shown in Figure 1.16 between x_1 and x_3 or between y_1 and y_3. Considerable caution and strong rationale are necessary to specify method factors and correlated error terms. Otherwise, this approach essentially can be misused to fit the model to the data.

A key issue in structural equation modeling is identification, or whether model parameters can be uniquely determined from the data (Bollen, 1989; Kaplan, 2000). Another issue relates to procedures used to estimate the model, such as maximum likelihood—used in the example on speech quality in Exhibit 1.5—and generalized least squares (Kaplan, 2000). A considerable amount of work has focused on the nature of statistical tests and overall fit indexes, and their vulnerability to factors such as sample size. The chi-square statistic reflects the level of mismatch between the sample and fitted covariance matrices (Hu & Bentler, 1998); hence, a nonsignificant chi-square is desirable. However, this statistic is influenced by a number of factors, including sample size. On the other hand, fit indexes quantify the degree of fit. They have been classified in several ways, such as absolute (i.e., directly assessing goodness of a model) versus incremental (assessing goodness of a model in comparison with a baseline model, such as a model in confirmatory factor analysis where each item is a distinct factor). Specific fit indexes, such as the goodness-of-fit index, the adjusted goodness-of-fit index, the normed fit index (ranging from 0 to 1), and the non-normed fit index, are each vulnerable to various factors, such as sample size (e.g., Bollen, 1986; Mulaik et al., 1989). Fit indexes have been compared on criteria such as small sample bias and sensitivity to model misspecification (Hu & Bentler, 1998). The standardized root mean square residual and the root mean square error of approximation are other indexes used to gauge fit. Considerable literature has compared various indexes of fit. As an exemplar of such literature, Hu and Bentler (1998) recommend the root mean square residual along with one of several indexes, including the comparative fit index (Hu & Bentler, 1998). The comparative fit index currently is the most recommended index (Bagozzi & Edwards, 1998). Important here is the usage of several indexes, including those currently recommended in the literature and generally produced in programs such as LISREL 8.5, as well as consideration and presentation of full information on the results, ranging from chi-square to multiple fit indexes to variance extracted and standardized root mean square residuals. Computations for some fit indexes are shown in Appendix 1.1. What specific value constitutes acceptable fit is another area with a wealth of literature. Some guidelines are provided in Chapter 5.

In comparing SEM versus the traditional approach to measurement, the traditional approach is a necessary step to developing measures and assessing reliability, dimensionality, and validity. In using SEM, a key issue is that many different models may fit the data. Therefore, SEM could be used for confirmatory factor analysis after some initial work has been completed on a measure such as the PNI scale. Preliminary empirical and conceptual work would serve to purify measures before using SEM. SEM can also be used for assessing different types of validity, as discussed subsequently. If confirmatory factor analysis is used in an exploratory fashion, the findings have to be confirmed with new data. If this procedure is used to test alternative models, then the chosen model should be tested with new data. Such testing is extremely important, because many different models can lead to good fit. Furthermore, a model could be modified in many ways to achieve fit (for example, in a simple unidimensional model,

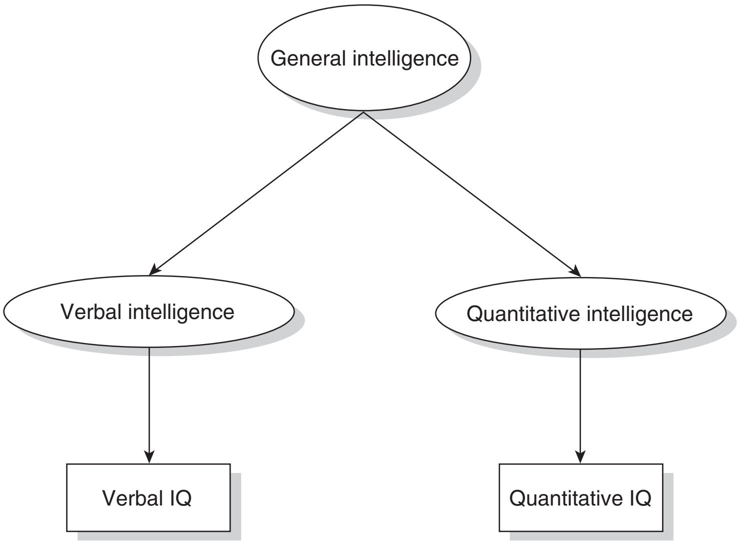


Figure 1.17 Hierarchical Factor Structure

correlated errors could be used to achieve fit). If two items of the same construct or of different constructs are influenced by an identifiable factor (say, the same measure at different points in time), then using a model with correlated error may be useful. Noteworthy here is that there are many ways to improve fit, and plausible reasoning should precede modeling. Furthermore, SEM ideally should not be used during the early stages of measure development for item purification.

It should be noted that hierarchical factor analyses can also be performed using structural equation modeling, wherein a set of factors loads onto higher level factors, such as, say, quantitative and verbal intelligence loading onto general intelligence. Thus, second-order factor analysis would involve an intermediate level of factors that loads onto higher order factors (Figure 1.17).

Validity

Whether a measure captures the intended construct, or whether the core tapped by a measure is the intended core, is the purview of validity. Assessing validity is like searching for an object at the bottom of the ocean with a searchlight, not knowing what the object looks like! Such searches may employ certain predetermined criteria (i.e., target areas on which to focus that are akin to representative domain, and content validity in terms of what the object might look like and indicators to look for in the ocean). When multiple searchlights and search crews are used and converge on the same location (i.e., multiple measures), then one type of evidence of validity is provided. When an unseen object has to be located, different types of evidence are needed to locate it.

The distinction between reliability and validity, and how one can actually be unrelated to the other, is brought out by the following stark example modified from Nunnally (1978). Consider an exercise where a stone of a certain weight is thrown as far as possible over 20 trials with sufficient rest between trials and under conditions with no wind and comfortable weather. Is this a reliable measure of intelligence? Yes, if the same measure a week later yields the same average throwing distance. In the social sciences, the criterion for consistency is often the relative standing of a set of people at test versus retest as reflected in a correlation, the requirement of the exact throwing distance between test and retest being stringent for the nature of the phenomenon being studied. Note that multiple trials and rest between trials may minimize variations because of arm weariness and may average out errors. Whether throwing stones captures intelligence is the purview of validity. This is an extreme scenario because a reliable measure is typically likely to have some degree of validity as a result of procedures employed to capture the content of a construct. However, the scenario is instructive in illustrating that a reliable measure is nothing more than a replicable measure. Hence, in colloquial language, a reliable friend who is late but consistently late by the same amount of time would still be considered reliable in the measurement world.

Several types of validity need to be considered (Churchill, 1979; Cronbach & Meehl, 1955; Nunnally, 1978). Very brief descriptions are provided below, demonstrated in Exhibit 1.6, and listed in Figure 1.18.

1. Content validity (subjective judgment of content, following proper procedures to delineate content domain, using a representative set of items, assessing if items make sense)

2. Face validity (Does the measure look like it is measuring what it is supposed to measure?)

3. Known-groups validity (Does the measure distinguish between groups that are *known* to differ on the construct, such as differences in scores on measures between people with or without specific medical conditions?)

4. Predictive validity (Does the measure predict what it is supposed to predict, such as an external criterion, say GRE or university entrance exam and grades in college?)

5. Convergent validity (Does the measure correlate or converge with another measure of the *same* construct?)
6. Discriminant validity (Is the measure of a construct not correlated with measures of constructs to which it is not expected to be related?)
7. Nomological validity (Does the measure of a construct relate to measures of other constructs with which it is theoretically expected to be correlated; that is, considering a nomological or theoretical network of constructs, does the measure behave in theoretically expected ways?)
8. Construct validity (Does a measure measure what it aims to measure; does a measure or operationalization correspond to the underlying construct it is aiming to measure?)

Exhibit 1.6 Validity Analysis of the PNI Scale

Correlations Between PNI Scale and Other Scales

	PNI	MEN	MVL	MTH	ATF	ATC
MEN	0.67					
MVL	0.56	0.41				
MTH	0.74	0.93	0.73			
ATF	0.57	0.42	0.46	0.50		
ATC	0.51	0.65	0.41	0.66	0.49	
ATS	0.61	0.64	0.49	0.69	0.79	0.92

NOTES: All correlations were significant at the 0.01 level. MEN = Enjoyment of mathematics scale; MVL = Value of mathematics scale; MTH = Total attitude toward mathematics scale; ATF = Attitude toward statistics field scale; ATC = Attitude toward statistics course scale; ATS = Total attitude toward statistics scale. All scales scored such that higher scores indicate more positive attitudes toward statistics, mathematics, and so on.

	PNI
AMB	–0.24**
NFC	0.30**
SOC	0.03

NOTE: Scales scored such that higher scores indicate more tolerance for ambiguity and higher need for cognition.

**p < .01.

Scale Descriptions

Attitude Toward Mathematics Scale (MATH). This scale is divided into (a) an Enjoyment of Mathematics scale (MATHEN)—"a liking for mathematical problems and a liking for mathematical terms, symbols, and routine computations"

(Continued)

(Continued)

(Aiken, 1974, p. 67); sample item: "Mathematics is enjoyable and stimulating to me" (Aiken, 1974, p. 68); and (b) a Value of Mathematics scale (MATHVAL), which relates to "recognizing the importance and relevance of mathematics to individuals and to society" (Aiken, 1974, p. 67); sample item: "Mathematics is not important for the advance of civilization and society" (Aiken, 1974, p. 68).

Intolerance for Ambiguity Scale (AMB). This scale defines "a tendency to perceive or interpret information marked by vague, incomplete, fragmented, multiple, probable, unstructured, uncertain, inconsistent, contrary, contradictory, or unclear meanings as actual or potential sources of psychological discomfort or threat" (Norton, 1975, p. 608). A sample item is "I do not believe that in the final analysis there is a distinct difference between right and wrong" (Norton, 1975, p. 616).

Attitude Toward Statistics Scale (STAT). This scale is a measure of attitudes held by college students toward an introductory course in statistics (Wise, 1985). This scale is divided into Attitude Toward Course (STCOURS) and Attitude Toward the Field (STFIELD) subscales.

Need for Cognition (NFC). This scale defines "the tendency of an individual to engage in and enjoy thinking" (Cacioppo & Petty, 1982, p. 116).

Results

The relationship between the PNI scale and the Social Desirability scale (Crowne & Marlowe, 1964) was assessed to examine whether responses to items indicating a higher (or lower) preference for numerical information may be partially explained by a motive to appear socially desirable, and to provide evidence for the discriminant validity of the PNI scale. A possible explanation for such responses may be based on a perception that it is more socially desirable to indicate greater preference for numerical information. This may be particularly likely given the composition of the sample, which consisted of undergraduate students at a midwestern U.S. university. Having taken quantitative courses, students may be likely to indicate a greater preference for numerical information as a means of appearing socially desirable. PNI had no significant correlations with the 33-item Social Desirability scale ($r = 0.03$). Therefore, it appears that social desirability is not a significant factor in explaining responses to items on the PNI scale, with the result providing evidence for the discriminant validity of the PNI scale.

The relationship between PNI and Need for Cognition—the "tendency of an individual to engage in and enjoy thinking" (Cacioppo & Petty, 1982, p. 116)—was assessed to provide evidence of the nomological validity of the PNI scale. Because the PNI scale is argued to tap individuals' preference for engaging in thinking involving numerical information (as captured by aspects such as enjoyment and liking), a positive relationship was expected between PNI and Need for Cognition. Individuals who have a tendency to enjoy thinking may be more likely to enjoy thinking based on a particular type of information (i.e., numerical information) than individuals who do not enjoy thinking. PNI had a positive correlation with Need for Cognition ($r = 0.30$, $p < .01$). The significant correlation provides evidence for the claim that the PNI scale taps proclivity toward thinking based on one type of information (i.e., numerical information). However, the size of the correlation suggests that a tendency to enjoy thinking per se is not strongly related to a tendency to enjoy thinking based on information in a numerical form, perhaps because numerical information is just one of several types of information that could be used in thinking. This is a form of evidence for nomological validity in tapping enjoyment of thinking based on one type of information.

Given the numerical content in statistics and mathematics, positive correlations between preference for numerical information and attitudes toward statistics and mathematics were expected in order to provide evidence of the nomological validity of the PNI scale. Positive correlations were obtained between the PNI scale and the Enjoyment of Mathematics scale ($r = 0.67$, $p < .01$); the PNI scale and the Value of Mathematics scale ($r = 0.56$, $p < .01$); and the PNI scale and the total Attitude Toward Mathematics scale ($r = 0.74$, $p < .01$). Positive correlations were also obtained between the PNI scale and the attitude toward statistics course scale ($r = .57$, $p < .01$); the PNI scale and the attitude to statistics field scale ($r = 0.51$, $p < .01$); and the PNI scale and the total statistics scale ($r = 0.61$, $p < .01$). The results suggest that PNI has moderate to strong relationships with the various subscales relating to mathematics and statistics, thereby providing evidence for the nomological validity of the PNI scale due to the overlap between these scales in terms of numerical content. The PNI scale had comparable or higher correlations with the subscales of attitude toward mathematics (statistics), such as the Enjoyment of Mathematics and Value of Mathematics scales, than the correlations between these subscales, possibly because PNI overlaps with both subscales in terms of numerical content, whereas the subscales overlap in terms of mathematical (statistical) content.

EXHIBIT SOURCE: Adapted from Viswanathan (1993).

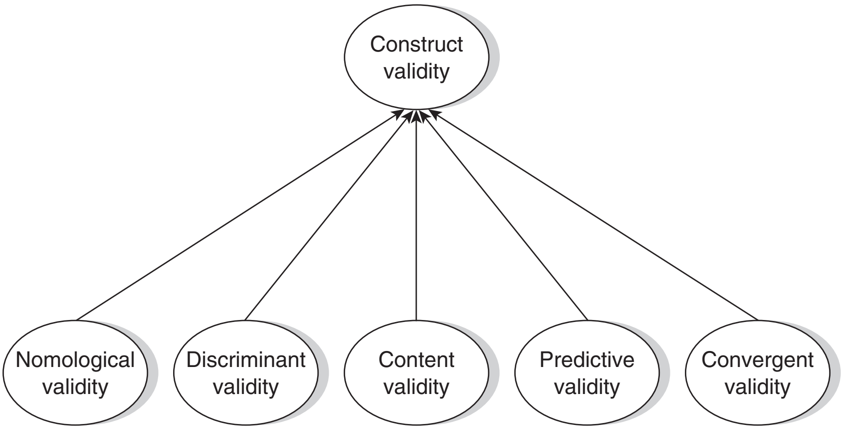


Figure 1.18 Types of Validity

Content validity relates to whether a measure is representative of a domain and whether appropriate procedures have been employed for the development of items. Content validity comes into play if a measure is construed to represent just one or a few aspects of the domain. Assessments of content validity focus on whether aspects of the domain of a construct have been excluded and aspects of the domain of a distinctly different construct have been included. Content validity is a subjective assessment of the representativeness of a measure and the procedures used to develop the domain. Whereas most other validity assessments are mainly empirical in nature, content validity is not and is also a central indicator of the representativeness of a measure. For the PNI scale, content validity evidence resides on documentation of procedures for domain delineation and item generation (Exhibit 1.1). Face validity relates to whether a measure looks like it is measuring what it is supposed to measure, a very preliminary judgment about validity.

Several other types of validity depend on empirical evidence. Convergent validity tests assess the degree to which two measures of a construct converge (i.e., are highly correlated or strongly related). Convergent validity between two measures is somewhat akin to internal consistency between items of a measure. The logic of convergent validity is that a measure of a construct should converge with another validated measure of the same construct, assuming such a measure is available. However, convergence alone is not definitive evidence of validity because both measures could be invalid. No two measures in practice are likely to be exactly identical and lead to perfect convergence. In fact, such similar measures may not serve to test convergent validity if they are really trivially different from each other. The aim here is to attempt different approaches to measuring the same construct, which generally may translate to less-than-perfect correlation among alternative measures. Two different ways of measuring the same thing may not, in practice, be identical. Just as items get at different aspects of a domain, different ways or methods of measuring something may get at different aspects of a construct. Different approaches also differ methodologically, contributing to the less-than-perfect correlation to be expected between two measures of the same construct. In fact, as knowledge in an area progresses, researchers may well conclude that what were originally considered two measures of the same construct are really measures of different constructs. But this is one way in which knowledge in an area progresses, different ways of attempting to measure constructs being central in advancing such knowledge. Convergence between measures of the same construct is a matter of degree for a variety of reasons. Note that convergent validity is not involved when two measures of constructs are shown to be related. Rather, it is useful to restrict this type of validity to the relationship between two measures of the *same* construct. Relationships between measures of related constructs are captured under nomological validity.

Nomological validity demonstrates that a measure is related to a construct to which it is theoretically expected to be related. Discriminant validity assesses the degree to which a measure is not related to constructs to which it is not supposed to be related. Hence, it is the empirical counterpart of the notion that domain delineation spell out what a construct is not. Significant relationships with constructs to which a measure is expected to be related provide evidence of nomological validity, whereas nonsignificant relationships with constructs to which a measure is expected to be unrelated provide evidence of discriminant validity. However, such a statement is polarized in that evidence of small relationships could also provide evidence of either type of validity (e.g., the relationship between PNI and need for precision) (see Exhibit 1.6). For example, if the expected theoretical relationship is small and such a pattern is found, then it could provide evidence of either nomological or discriminant validity.

Discriminant validity and nomological validity are two sides of the same coin. A small relationship between a measure and another measure could provide evidence of both and could well be considered nomological validity; that is, are measures of constructs behaving in theoretically expected ways in nonrelationships or in small relationships? A simple version of discriminant validity is to show that a measure of a construct is unrelated to a measure of a construct to which it is not supposed to be related. For example, a measure of intelligence could be shown to have discriminant validity by showing a nonrelationship with stone throwing, a measure of arm strength. A measure of preference for numerical information could be shown to be unrelated to a measure of extroversion. Clearly, a completely unrelated measure could be selected from an infinite set to demonstrate discriminant validity. Often, measures are shown to be unrelated to individual differences in social desirability in order to demonstrate that the measure in question is not eliciting responses based on respondents' needs to appear socially desirable. Such assessments play an important role in showing that response tendencies unrelated to content drive scores on measures.

However, showing that a measure is unrelated to measures of other constructs may be a weak form of evidence of discriminant validity. After all, a construct would likely be unrelated to numerous constructs from completely unrelated domains. The need for a reasonable test arises only when there is some possibility of a relationship and a specific psychometric purpose in assessing the relationship, such as showing that a measure is not influenced by social desirability. For the PNI scale (Exhibit 1.6), its relationship with social desirability is assessed to show that responses are not being influenced by social desirability—a plausible alternative for students rating preference for numerical information. In a similar vein, showing that a measure is unrelated to a few measures from a large pool of potentially unrelated measures in itself does not provide sufficient evidence about the discriminant validity of the measure. Perhaps stronger evidence can be provided by showing that a measure has a small relationship with a measure of a construct to which it is likely to be related; that is, these related measures are indeed tapping into different constructs. For the PNI scale (Exhibit 1.6), its relationship with need for cognition is assessed to show a significant but small relationship. A subtler form of discriminant validity is to show a differential relationship between two related measures and a third measure. Examples of each of these types of evidence are presented below for the PNI scale (Exhibit 1.6). Here, the PNI scale and a measure of a construct to which it is strongly related are shown to have a differential relationship with a third measure. Thereby, the PNI scale is shown to be distinct from a measure of a construct to which it is strongly related.

Predictive validity is related to whether a measure can predict an outcome, and it has relevance in practical settings. In descriptive research, predictive validity is often subsumed under nomological validity. Known-groups validity is a form of predictive validity wherein a measure is shown to distinguish between known groups in terms of scores (e.g., differences in scores on measures between people with or without specific medical conditions in clinical settings, or differences in PNI scores for undergraduate majors in mathematics vs. art).

Construct validity is an umbrella term that asks the basic question, Does a measure measure the construct it aims to measure? There is no empirical coefficient for construct validity, just increasing amounts of evidence for it. Therefore, there is no correlation coefficient for construct validity, only degrees of evidence for it using all the types of validity listed above. Construct validity is the most important and most difficult form of validity to establish. It is akin to establishing causality between variables in substantive research, only causality is between a construct and its measure.

Although Exhibit 1.6 presents correlations between measures, it should be noted that structural equation modeling can be used to assess different types of validity while accounting for measurement error, such as unreliability in measures. Relationships between a focal measure and measures of other constructs in nomological validity tests can be assessed while accounting for measurement error. Discriminant validity can be shown by differential relationships between a target measure of a construct and a measure of a related construct and a third variable (Judd, Jessor, & Donovan, 1986). Discriminant validity could also be demonstrated by specifying a one- versus two-factor model of measures of two different constructs and showing that the two-factor model has superior fit.

An example of the entire measure development process for a somewhat different construct is presented in Exhibit 1.7. This construct is, in some ways, more complex than PNI and is used to illustrate a different scenario.

Exhibit 1.7 Developing a Scale of Consumer Literacy: A Different Type of Scale

Whereas the scales used as examples earlier relate to attitudes, other scales, such as intelligence tests, may be ability-related. Some scales may be on the margin between ability and attitude, such as the consumer literacy scale for people with low levels of literacy.

First Phase: Literature Review and Exploratory Interviews

In the first phase, exploratory research is undertaken to (a) examine parallel measures of ability and skill through a literature review; (b) assess adult education tests and textbooks for examples in consumer contexts; (c) examine assessment tools in literacy, adult education, and, more broadly, education, such as the National Adult Literacy Survey (1993); and (d) interview educators in adult education. The aim of this phase is to develop comprehensive grounding before developing the measure and conceptualizing the construct.

Second Phase: Conceptualization and Domain Delineation

In the second phase, the domain of consumer literacy is carefully delineated. In conceptually defining a construct, it should be sufficiently different from other existing constructs and should make a sizable contribution in terms of explanatory power. This step involves explicating what the construct is and what it is not. This is also a step where the proposed dimensionality of the construct is described explicitly. Keeping in focus the need to have a measure at the low end of the literacy continuum, this delineation of domain needs to list the basic skills necessary to complete fundamental tasks as consumers. A matrix of basic reading, writing, and mathematical skills versus consumer tasks should be created to provide a complete delineation. This listing of skills and associated consumer tasks needs to be examined by several individuals with different types of expertise, such as consumer researchers, adult education teachers, education researchers, and students at adult education centers who have progressed through to completion of the GED and further education. A team of experts can be formed consisting of teachers/directors at adult education centers, researchers in consumer behavior, and researchers in educational measurement. Experts can be provided with conceptual definitions of various dimensions of consumer literacy and associated listings of skills and consumer tasks and asked to evaluate the definitions and the domain delineation for completeness and accuracy. Such careful delineation of the domain is very important in any measure development process, all the more so for a construct as complex as consumer literacy.

(Continued)

(Continued)

Third Phase: Measure Construction and Content Validity Assessment

In the third phase, the measure should be constructed through item generation. In this phase, a large pool of items (i.e., consumer tasks) is developed to identify specific aspects of the domain explicated in the previous phase. A large pool of items is important, as are appropriate procedures to develop individual items. Redundancy is a virtue at this stage in the process, with the goal being to cover important aspects of the domain (DeVellis, 1991). Researchers have documented a host of issues that need to be addressed in developing items (DeVellis, 1991; Haynes et al., 1995). Several such procedures should be employed, such as asking experts to generate items on the basis of definitions of specific dimensions of consumer literacy, and asking experts to evaluate items in light of conceptual definitions. Thus, item generation draws on the team of experts mentioned in Phase 2. Through these means, the ever-pertinent issue of understanding of items by low-literate individuals is also addressed. Thus, items and usage conditions are assessed together and modified. The procedures discussed so far relate to the content validity of a measure, or whether a measure adequately captures the content of a construct. There are no empirical measures of content validity, only assessments based on whether (a) a representative set of items was developed and (b) appropriate procedures were employed in developing items (Nunnally, 1978). Hence, assessment of content validity rests on the explication of the domain and its representation in the item pool.

Fourth Phase: Reliability and Validity Assessment

In the fourth phase, empirical studies should be conducted to assess the reliability, dimensionality, and validity of the consumer literacy measure. Convenience samples can be drawn for all studies from students at adult education centers. This method also allows for access to the records of participants on important and relevant tests. The consumer literacy measure is broad in scope and ideally applies to low-literate individuals in general. In this regard, students at adult education centers should be distinguished from other functionally low-literate consumers in their motivation to become functionally literate. Nevertheless, the choice of students at adult education centers greatly facilitates efficient access and recruitment of a group that is very difficult to sample and provides a sound starting point. Convenience sampling is suited for these studies rather than probabilistic sampling because the aim is not to establish population estimates, but rather to use correlational analysis to examine relationships between items and measures. Several pilot tests should be conducted in which students at adult education centers complete the measure. Through observation and explicit feedback from students and teachers, the measure can be modified as needed. Such pilot testing is expected to be carried out in many stages, each using a small set of respondents, with the measure being adjusted between stages. Improvements in the measure during pilot testing, both in changing individual items and in addressing administration procedures, are likely to be considerable and to go a long way toward minimizing sources of measurement error in the final measure. Four to five larger scale studies are required in this phase, each employing sample sizes of 100–150. The first two to three large-scale studies should aim to assess and purify the scale through assessment of the reliability of individual items, and through initial assessment of validity and dimensionaity. Subsequent studies with a purified measure are more confirmatory and assess dimensionality as well as test validity in a detailed manner.

After item generation, measures typically are assessed for internal consistency reliability, and items are deleted or modified. Internal consistency frequently is the first empirical test employed and assesses whether the items in a set are consistent with each other or belong together. Such internal consistency assessment is pertinent in evaluating items within each dimension of consumer literacy for consistency in response.

An important form of reliability in this context is test-retest reliability, where the measure is completed twice by the same individuals with an interval of, typically, a few weeks. In a sense, test-retest reliability represents a one-to-one correspondence with the concept of reliability, which is centered on replicability or consistency. Test-retest reliability offers direct evidence of such consistency over time.

Exploratory factor analysis should be employed for preliminary evaluation of the dimensionality of the measure by assessing the degree to which individual items are correlated with respective dimensions or factors. However, confirmatory factor analysis is required for conclusive evidence (Gerbing & Anderson, 1988). Confirmatory factor analysis imposes a more restrictive model, where items have prespecified relationships with certain factors or dimensions.

Item-response theory (Hambleton, Swaminathan, & Rogers, 1991) offers another scaling technique that is particularly relevant here, because sets of items may require similar levels of skill. Therefore, responses to such sets of items may be similar. Item-response theory can specify the relationship between a respondent's level and the probability of a specific response. This approach can be used to assess measurement accuracy at different levels of ability and also to construct measures that have comparable accuracy across different levels of ability. The use of item-response theory for the National Adult Literacy Survey (1993) provides guidance in this regard.

Tests of different types of validity take measure development to the realm of cross-construct relationships. Whereas reliability establishes consistency and the lack of random error, validity pertains to whether a measure measures what it purports to measure (i.e., the lack of random and systematic error). Several types of validity need to be considered, and very brief descriptions are provided below along with examples of possible tests.

Content validity relates to a subjective judgment of content, based on adherence to proper procedures to delineate content domain and on generation of a representative set of items. For example, content validity could be assessed and enhanced by a team of experts reviewing the conceptual definition, domain delineation, and item generation. *Convergent validity* relates to whether the measure correlates or converges with another measure of the same construct. This type of validity is not directly applicable here because no other measure of consumer literacy is available.

(Continued)

(Continued)

However, a proxy measure can be created from current functional literacy measures using tasks oriented toward consumer economics. Other types of validity include *known-groups validity*, where a measure is shown to distinguish between known groups, such as individuals with or without an illness in clinical settings. In this context, this form of validity can be evaluated by assessing differences in the consumer literacy measure for students with 0 to 4th- versus 5th- to 8th- versus 9th- to 12th-grade reading levels. *Predictive validity* is related to whether a measure can predict a criterial outcome. In this context, one way in which predictive validity can be assessed is by relating the measure to performance in specific consumer tasks. *Nomological validity* demonstrates that a measure is related to a construct to which it is theoretically expected to be related. Examination of the relationship between consumer literacy and traditional functional literacy tests is an example of a nomological validity test. *Discriminant validity* assesses whether the measure of a construct correlates weakly or not at all with measures of constructs to which it is expected to relate weakly or not at all, respectively. Small or moderate correlations between a measure of consumer literacy and basic measures of reading literacy would provide a form of evidence of discriminant validity.

The aim of all of the measurement procedures described above is to provide evidence of *construct validity*, an umbrella term that asks the basic question, “Does a measure measure the construct it aims to measure?” There is no empirical coefficient for construct validity, just increasing amounts of evidence for it.

In all phases of measure assessment, both individual items and administration procedures need to be adjusted carefully. This is critical for the proposed measure; both the items of a measure and the conditions of their use have to be carefully assessed to minimize measurement error.

It should be noted that validity tests rest on several assumptions (Figure 1.19). To validate a measure of a construct by relating it to another measure of a different construct (say, in a test of nomological validity), it is important to use a reliable and valid measure of this different construct. It is also important to have a sufficient support for the hypothesized relationship between two constructs. Thus, only one of the three branches in Figure 1.19 is being tested and results can be interpreted clearly.

Looking at the broad picture of the relationship between reliability and validity, reliability is a necessary but not sufficient condition for validity. Although reliability and validity are both a matter of degree, generally speaking, a valid measure is reliable, but a reliable measure is not necessarily valid. Reliability is about random error, and if random error has not been reduced, the issue of validity does not arise.[12] If a measure is reliable, it

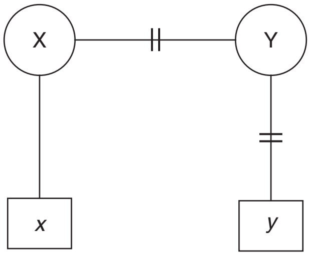


Figure 1.19 Assumptions of Validity Tests

Assumptions

– Construct X is related to Construct Y

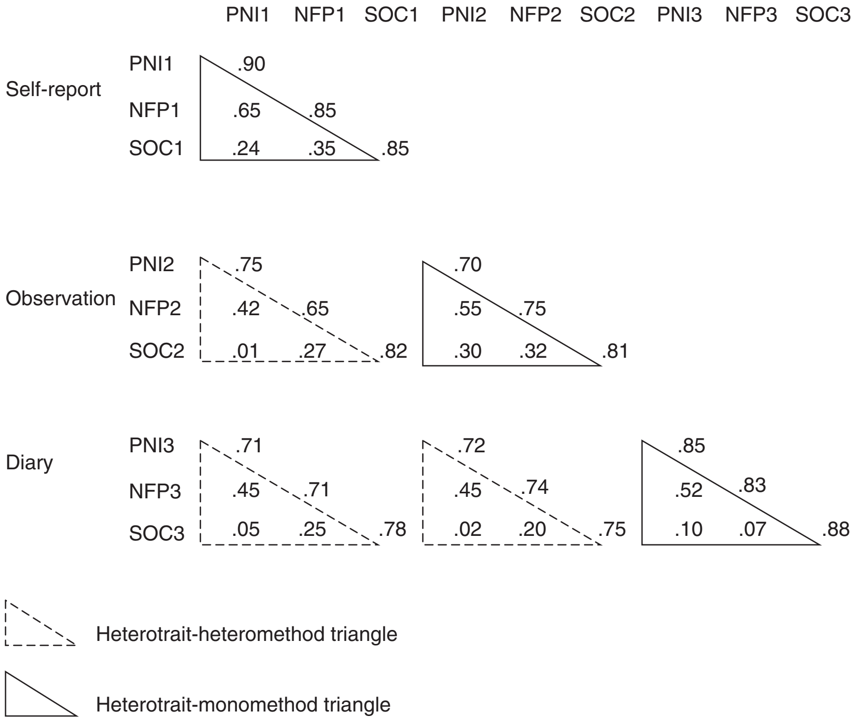
– y is a reliable and valid measure of Y

$\therefore$ r_{xy} can be used to assess r_{xX}

$r_{xy} = r_{xX} \cdot r_{XY} \cdot r_{Yy}$

means that a large proportion of the total variance is due to common sources. Then, it is pertinent to discuss validity. Even if a measure looks face valid, if it is not reliable, it is not valid. Reliability and validity have been viewed as being the ends of a continuum where reliability relates to agreement between the same or similar methods, whereas validity relates to agreement between different methods (Campbell & Fiske, 1959). This view is consistent with the difference between reliability and convergent validity discussed here.

The multitrait-multimethod approach systematically examines the effect of using identical versus different methods (Campbell & Fiske, 1959). It can be illustrated using the PNI construct, essentially adapting similar matrices presented in past research (Campbell & Fiske, 1959) (see Figure 1.20). Consider three different ways of measuring preference for numerical information: (a) a self-report scale as shown earlier (self-report), (b) a rating by observers based on a half-hour of observation of behavior in a controlled setting on the degree to which individuals spend time with numerically oriented tasks (observation), and (c) a diary that individuals complete of numerically oriented tasks in which they engaged every day for a few weeks (diary method). Consider also two other traits that are measured through these three methods: Need for precision (NFP) (Viswanathan, 1997) and social desirability (SOC). A hypothetical pattern of correlations is shown in Figure 1.20 and represents data from the same individuals on these different methods and traits.



		PNI1	NFP1	SOC1	PNI2	NFP2	SOC2	PNI3	NFP3	SOC3
Self-report	PNI1	.90								
	NFP1	.65	.85							
	SOC1	.24	.35	.85						
Observation	PNI2	.75			.70					
	NFP2	.42	.65		.55	.75				
	SOC2	.01	.27	.82	.30	.32	.81			
Diary	PNI3	.71			.72			.85		
	NFP3	.45	.71		.45	.74		.52	.83	
	SOC3	.05	.25	.78	.02	.20	.75	.10	.07	.88

Heterotrait-heteromethod triangle
Heterotrait-monomethod triangle

Figure 1.20 Multitrait-Multimethod Matrix

The reliability diagonal refers to the reliability of the measures and is essentially the correlation of a measure with itself. This is consistent with indicators of reliability, such as coefficient alpha, being the proportion of variance attributable to common sources and test-retest reliability being the test-retest correlation. Reliability is the relationship of a measure with itself. The validity diagonal (that is, heteromethod-monotrait) lists correlations between measures of the same construct using different methods. This is convergent validity, the relationship between two different measures of the same construct. The heterotrait-monomethod triangle shows the relationship between measures when using the same method. These correlations may have been inflated by the use of the same method, say, capturing individuals' tendencies to respond in certain ways to certain methods, such as in using the positive end of self-report scales. This is an example of a type of systematic error—consistent differences across individuals over and above the construct being measured—that may inflate or deflate correlations. Whereas random error reduces observed correlations, systematic error may inflate or deflate observed correlations. A simplified example of systematic error was illustrated with the weighing machine example. Subsequent chapters will introduce a nuanced discussion of systematic error. To examine the effect of method, these correlations have to be compared to the heterotrait-heteromethod triangle. The effect of using the same method can be gauged by comparing correlations between measures using different methods versus the same method. For example, the correlation between PNI and NFP is considerably lower when using observation for NFP and self-report for PNI (0.42) than when using self-report for both (0.65). Therefore, the differences in correlations using the same versus different methods provide an estimate of the effect of a common method. In contrast, the diary method does not have as much of a common method effect as self-report (0.52 vs. 0.45). Social desirability has a correlation of 0.24 and 0.35 with PNI and NFP, respectively, using self-report. However, for PNI, the corresponding correlations in the heteromethod-heterotrait triangle are considerably lower (0.01 and 0.05) but remain small but sizable for NFP (0.27 and 0.25). This pattern points to problems in the self-report NFP scale in terms of being associated with social desirability, an undesirable characteristic in this context. On the other hand, the diary method of assessing NFP does not have this problem.

The multitrait-multimethod approach outlined above suffers from several problems (Bagozzi & Yi, 1991). One problem is the absence of clear standards for ascertaining when any particular criterion is met, because only rules of thumb are available. Another problem is the inability to assess separate amounts of trait, method, and random error in the data, all confounded by examining only the raw correlations. Several assumptions are made, such as no correlations between trait and method factors, all traits being equally influenced by method factors, and method factors being uncorrelated (Bagozzi & Yi, 1991). The use of structural equation modeling for multitrait-multimethod matrices has several advantages, such as providing measures of the overall degree of fit and providing information about whether and how well convergent and discriminant validity are achieved (Bagozzi & Yi, 1991). Figure 1.21 illustrates such an approach where method factors are incorporated into the model. Other approaches to modeling the relationships include using correlated error terms between items that share the same method.

General Issues in Measurement

One issue about constructs is noteworthy: The strength and relevance of constructs may vary across people. The notion of a metatrait has been used

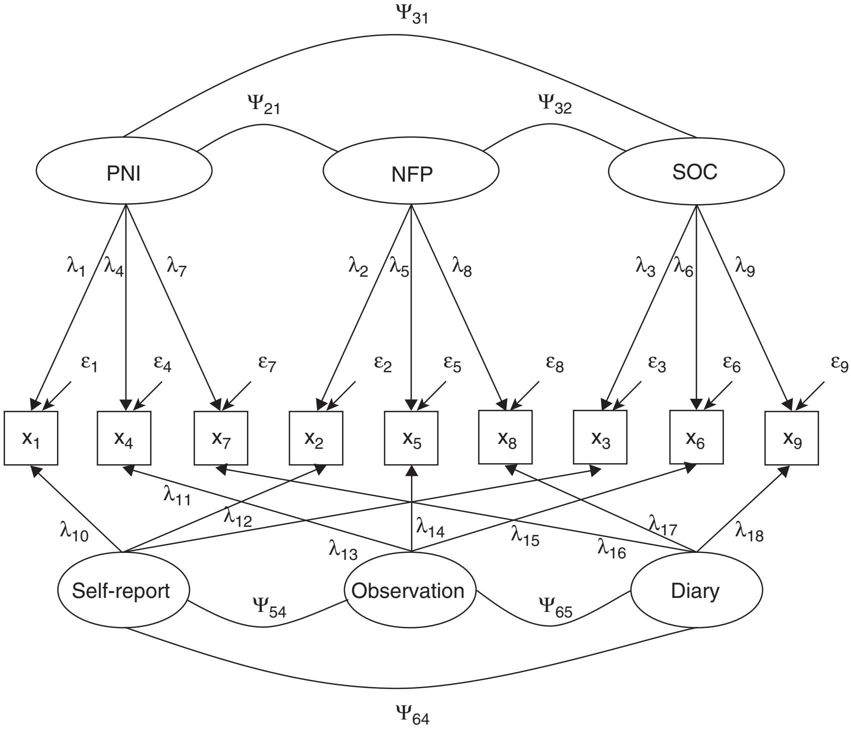


Figure 1.21 Using Structural Equation Modeling for Multitrait-Multimethod Data

to refer to possessing or not possessing a trait (Britt, 1993). Moreover, many questions used in a research design are really just that and do not purport to assess an underlying abstract construct. For instance, they may assess specific behaviors, such as spending on entertainment. Nevertheless, examination of the intended thing being measured using the measure development process described can be very beneficial. Such examination can lead to insights, precise wording of the question to enable consistent interpretation, and a broadened examination of multiple issues about this phenomenon through several questions. For example, with entertainment spending, pertinent questions relate to what entertainment is and what is included or excluded.

A few general points are noteworthy with regard to psychometrics. First, relatively large sample sizes are essential for various psychometric procedures (Guadagnoli & Velicer, 1988; Nunnally, 1978). Small samples lead to large sampling errors and add new forms of uncertainty into psychometric estimates of reliability and validity. Coefficient alpha has a confidence interval associated with it as well that is based on sampling error (Appendix 1.1). Second, lack of sufficient variation inhibits the ability of an item to covary or correlate with other items and with the total measure. Covariations, reflected statistically in correlations, are the underpinning of psychometric analyses, and lack of sufficient variation inhibits the ability of an item to covary with other items. Third, although reliability and validity are the key concepts in measurement, dimensionality is discussed separately earlier. However, it is subsumed within validity, the aim being to understand what a measure is capturing both conceptually and empirically. Fourth, and more generally, measure validation is an ongoing process whereby measures are refined; items are added, modified, or deleted; dimensions are expanded; and constructs are redefined. Therefore, no single study is definitive in terms of measure validation. Fifth, empirical testing of measures rests on relative ordering, correlations being largely determined by relative standing between variables (Nunnally, 1978). Relationships examined in substantive studies in the social sciences are correlational in nature as well, reflecting the nature of the phenomenon and the precision with which relationships can be specified. Sixth, an attenuation formula to allow for unreliability in studying the relationship between two variables is available in the literature and presented in Appendix 1.1. However, there is no substitute for well-designed, reliable, valid measures to begin with. More generally, data analysis may not be able to account for problems in research design. Finally, generalizability studies may be a useful alternative by formalizing the effects of occasions and items (Cronbach, Rajaratnam, & Gleser, 1963; Rentz, 1987). Generalizability studies assess measurement error across facets or conditions of measurement, such as items, methods, and occasions. This approach relates to generalizing from observations to a universe of generalization consisting of the conditions of measurement of facets. Effects of facets and interactions between them are analyzed in this approach. The generalizability approach blurs the distinction between reliability and validity in that methods could be a facet of generalization (Rentz, 1987). However, this approach may be complex to design and implement (Peter, 1979).

Summary

Accurate measurement is central to scientific research. There are two basic types of measurement error in all of scientific research—random error and systematic error. Measures relatively free of random error are reliable, and measures relatively free of random and systematic error are reliable and valid.[13] The measure development process consists of a series of steps to develop reliable and valid measures. It starts out by carefully understanding what is being measured, through definition of the construct and delineation of its domain. Rather than proceed directly from an abstract construct to its concrete measurement, the distance between the conceptual and the operational has to be traversed carefully and iteratively. Internal consistency reliability, test-retest reliability, dimensionality, and validity tests are performed on measures, with items being added, modified, or deleted along the way.

Measure development should ideally combine empirical assessment with conceptual examination. A purely empirical approach may neglect the content of items, whereas a purely conceptual approach neglects empirical reality and how respondents are actually interpreting an item. A purely conceptual approach that neglects empirical results rests on the notion that the measure or individual items are reliable and valid no matter what the outcomes say. A purely empirical approach rests on the assumption that, somehow, item content and domain representation are irrelevant once data are collected. For instance, an item that has moderate item-to-total correlation may still be worth retaining if it is the only item that captures a certain aspect of the construct's domain. Even an item with low item-to-total correlation may be worth editing if its content is uniquely capturing some aspect of a construct's domain. Similarly, conceptual reasoning for an item cannot overrule poor empirical outcomes, which are essentially suggesting problems in the item's interpretation.

Low reliability and low validity have consequences. Random error may reduce correlations between a measure and other variables, whereas systematic error may increase or reduce correlations between two variables. If there is a key theme in all of this, it is that a construct and its measure are not the same, and imperfect measurement of any construct has to be taken explicitly into account.

A metaphor that can be used to understand measurement is to consider the universe and the location of specific planets or stars, which are like abstract concepts. Although we are unable to see them, they have to be located through paths from Earth to them, akin to a path from a measure to the underlying concept. Reliability refers to whether the paths (measures) can be reproduced across time. Convergent validity refers to whether two different paths (measures) converge on the same planet (construct). Discriminant validity refers to whether the path (measure) to one planet (construct) is different from a path (measure) leading to a different planet (construct). Nomological validity refers to whether the path (measure) relates to other known paths (measures) to different planets (constructs) in expected ways in light of where the target planet (construct) is expected to be. Construct validity refers to whether the path (measure) indeed leads to the intended planet (construct).

The book has so far provided a brief overview of traditional measure development procedures. The rest of the book will provide more in-depth understanding of measurement error leading to important insights for measure development and methodological design. Thus, the rest of the material will break new ground in understanding measurement error and methodological design.

Notes

1. Nunnally (1978) is cited throughout the book wherein both the Nunnally (1978) book and the later edition by Nunnally and Bernstein (1994) can be cited.

2. The term *measure* is sometimes used interchangeably with the term *scale* in this book.

3. Many words used in science, of course, refer to things other than constructs, such as umbrella terms referring to topic areas, stimuli, or artifacts, such as a not-for-profit organization. Any stimulus object could be viewed in terms of numerous underlying constructs, with the specific object having specific levels of magnitude on these constructs. A not-for-profit organization, for instance, needs to be differentiated from other organizations by isolating the key dimensions of distinction. A construct can be viewed as having levels that are either along a quantitative continuum or are qualitatively distinct.

4. Whether a true value really exists is a relevant issue. When dealing with psychological phenomena, responses are often crystallized where questions are posed. The notion of a true value that exists when all the factors of a response situation are controlled for is somewhat like the notion of infinity, a useful concept. What usually matters in the study of relationships in the social sciences is the accurate relative ordering of individuals or stimuli.

5. Consistency is a central element of scientific research. A key characteristic of science is replicability.

6. Can measurement error be categorized in some other way? Perhaps error can be classified further in terms of having constant versus variable effects. Subsequent discussion will address this issue.

7. Researchers have suggested alternative approaches to the measure development process. For example, the C-OARS-E procedure (Rossiter, 2002) emphasizes content validity. The steps in this procedure—reflected in the acronym—are Construct definition (where the construct is defined initially in relation to object, attribute, and rater entity, e.g., IBM's [object] service quality [attribute] as perceived by IBM's managers [raters]), Object classification (where raters are interviewed to classify object as concrete or abstract and items are developed accordingly), Attribute classification (where raters are interviewed to classify attributes similarly as objects), Rater identification (where the rater is identified), Scale formation (where the scale is formed from object and attribute items, response categories are chosen, and items are pretested), and Enumeration.

8. As the number of items in a measure increases, it could be argued that there is a greater likelihood of picking items to cover different aspects of a domain, akin to randomly picking states in the United States and covering the geographic area. In practice, though, the opposite may happen sometimes, where items are picked from one or a few subdomains. Moreover, the issue of representing different aspects of the domain is not one of reliability but one of validity. In fact, representing a broad domain may actually lower reliability while enhancing validity.

9. For oblique rotations, a factor pattern matrix and a factor structure matrix are reported in factor analysis. The former, which is easier to interpret and usually reported, shows the unique loadings of items on factors, whereas the latter shows correlations between items and factors that include correlations between factors (Hair et al., 1998).

10. Another issue in factor analysis is the use of factor scores, which are estimates of the values of common factors. Problems with factor scoring procedures suggest that careful evaluation of factor scores is needed before using them (Grice, 2001).

11. This type of factor analysis is referred to as R factor analysis and is distinguished from Q factor analysis, where individuals load on factors and correlations are computed across stimuli (Hair et al., 1998; McKeown & Thomas, 1988). For example, in a Q sort method, 53 respondents ranked 60 statements on morality by ordering them from −5 to +5 (i.e., most unlike to most like respondents' views) (McKeown & Thomas, 1988). The rankings were then correlated and factor analyzed. Groups of individuals loaded on specific factors. The Q factor analysis approach is not used often because of problems with computation (Hair et al., 1998).

12. One way in which reliability and validity have been illustrated is through bullet holes in a shooting target, wherein the bull's-eye is analogous with the construct that a measure aims to measure. A set of bullet holes close together in proximity suggests consistency and reliability, whereas a set of scattered bullet holes suggests unreliability. A set of bullet holes close together around the bull's-eye suggests reliability and validity.

13. These statements describe reliability and validity in a mathematical sense in terms of the types of error. It is possible for a measure to be perfectly reliable and valid in capturing an unintended construct. Therefore, validity relates conceptually to whether a measure measures the construct it purports to measure. Measurement of a different construct reliably and validly would not, of course, meet this requirement. For instance, the stone-throwing exercise may be a reliable measure of intelligence, and perhaps a valid measure of arm strength, but it is not a valid measure of intelligence because it lacks content validity. Therefore, the mathematical description of reliability and validity in terms of error, in some sense, works when it is assumed that the intended construct is captured along with error. In other words, the observed score is the sum of the true score and error. But in extreme scenarios, lacking the true score term to begin with because of measurement of the unintended construct, the mathematical expression does not fully reflect the concept of validity. Observed scores on an unintended construct may be viewed as reflecting scores on the intended construct. For instance, in the case of the stone-throwing exercise, consistency may be achieved internally; however, the pattern of responses would be unrelated to true scores on intelligence. Noteworthy is that the issue of reliability is internal to a measure.

Appendix 1.1

Selected Metrics and Illustrations in Measurement

Random and Systematic Errors

$X_o = X_t + X_r + X_s$
X_o = observed score
X_t = true score
X_r = random error
X_s = systematic error

SOURCE: Adapted from Churchill, 1979.

Illustration of Internal Consistency

Sample Item Responses for the PNI Scale After Reverse Scoring Items

	Items							
Respondents	*N1*	*N2*	*N3*	*N4*	*N5*	*N6*	*N7*	*Total*
1	1	2	2	1	2	2		31
2	3	3	4	2	3	3		58
3	4	4	5	3	3	3		83
99	5	6	6	4	4	5		102
100	7	7	6	5	6	7		122

NOTE: Responses are reverse scored so that higher values denote higher preference for numerical information.

Covariance:

$$\sigma_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{n-1}$$

Correlation$_{xy}$:

$$\frac{\text{Covariance}_{xy}}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}}$$

$\sigma_x^2 =$ x true score variance

$\sigma_y^2 =$ y true score variance

Item-to-total correlation = Cov (N1, Total)/Sq rt [Var(N1)Var(Tot)]

Cov = Covariance
Var = Variance
Sq rt = Square root

Computation of coefficient alpha (DeVellis, 1991; Nunnally & Bernstein, 1994, p. 232)

Coefficient alpha = $k\bar{r}/[1 + (k - 1)\bar{r}]$
k = number of items in a measure
r = average interitem correlation

The expression of alpha presented above is based on a variance-covariance matrix of all the items in a measure. Alpha is the proportion of variance due to common sources. The elements in a covariance matrix can be separated into unique variances represented by diagonal elements and covariances between items represented by off-diagonal elements. A three-item example is shown below.

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

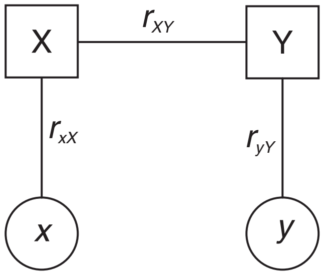
Alpha = $[k/(k - 1)][1 -$ (Sum of unique elements/Sum of covariance elements)]
$k/(k - 1)$ is a correction term applied to restrict the range of alpha from 0 to 1.

For example, coefficient alpha = $k\bar{r}/[1 + (k - 1)\bar{r}]$

For a 3-item scale:

If $r = 0$, then alpha = 0; If $r = 1$, then alpha = $3 \cdot 1/[1 + (3 - 1)1] = 1$.

Attenuation Formula for Reliability



x is a measure of construct X
y is a measure of construct Y
r_{xy} = observed correlation between x and y
r_{XY} = true correlation between X and Y
r_{xX} = correlation between measure x and construct X
r_{yY} = correlation between measure y and construct Y

$$r_{xy} = r_{xX} \cdot r_{XY} \cdot r_{Yy}$$

r^2_{xX} = reliability of x; r^2_{yY} = reliability of y
(i.e., the proportion of variance attributed to the true score)

$$\therefore\ r_{xy} = \sqrt{\mathrm{Rel}_x}\ .\ r_{XY}\ .\ \sqrt{\mathrm{Rel}_y}$$

Rel_x = Reliability of measure x; Rel_y = Reliability of measure y
If $r_{XY} = .5$, $\mathrm{Rel}_x = .5$, and $\mathrm{Rel}_y = .5$,
Then $= r_{xy} = \sqrt{.5} \cdot .5 \cdot \sqrt{.5} = .25$

SOURCE: Nunnally and Bernstein (1994), p. 241.

Kuder-Richardson Formula 20 for Reliability of Dichotomous Item Scale

$$r_{xx} = \frac{N}{N-1}\left(\frac{S^2 - \sum pq}{S^2}\right)$$

N = number of items
S^2 = variance of the total score
p = proportion passing each item; $q = 1 - p$

SOURCE: Nunnally and Bernstein (1994), p. 235.

Standard Error of Measurement (SEM) for Reliability

$$SEM = \sigma_x \cdot \sqrt{1 - r_{xx}}$$

σ_x = Standard deviation of the distribution of total scores on the test
r_{xx} = reliability of the test

SOURCE: Nunnally (1978), p. 239.

Illustration of Test-Retest Reliability

	Items							
Respondents	*N1*	*N2*	*N3*	*Total*	*RN1*	*RN2*	*RN3*	*RTotal*
1	1	2	2	31	2	2		29
2	3	3	4	58	3	3		59
3	4	4	5	83	3	3		77
99	5	6	6	102	4	5		98
100	7	7	6	122	6	7		127

NOTES: Responses on items N1, N2, and so on are at test, and RN1, RN2, and so on are at retest. Items are reverse scored so that higher values denote higher preference for numerical information. Total and RTotal are total scores at test and retest, respectively.

A Simplified Illustration of Exploratory Factor Analysis

The following are equations and computations showing the relationships between two uncorrelated factors and four items.

F = factor
x = item
l = loading or correlation between an item and a factor
e = error term

Numbers refer to different factors or items.

$x_1 = l_{11}F_1 + l_{12}F_2 + e_1$
$x_2 = l_{21}F_1 + l_{22}F_2 + e_2$
$x_3 = l_{31}F_1 + l_{32}F_2 + e_3$
$x_4 = l_{41}F_1 + l_{42}F_2 + e_4$

In matrix form:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \\ l_{31} & l_{32} \\ l_{41} & l_{42} \end{bmatrix} \times \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

Communality of $x_1 = l_{11}^2 + l_{12}^2$.

More generally, communality of $x_i = \Sigma_{j=1}^{n} l_{ij}^2$, where i is the ith item and j is the jth factor with n factors.

Variance explained by F_1 in eigenvalues $= l_{11}^2 + l_{21}^2 + l_{31}^2 + l_{41}^2$.

More generally, variance explained by F_j in eigenvalues, $F_j = \Sigma_{i=1}^{m} l_{ij}^2$, where i is the ith item and j is the jth factor with m items.

A Simplified Illustration of Confirmatory Factor Analysis

In confirmatory factor analysis, rather than every item being allowed to be related to every factor (as shown in the equations earlier), specific models are tested. Consider a model where the first two of four items are hypothesized to load on the first factor and the last two items are hypothesized to load on the second factor.

$$\begin{aligned} x_1 &= l_{11}F_1 + 0F_2 + e_1 \\ x_2 &= l_{21}F_1 + 0F_2 + e_2 \\ x_3 &= 0F_1 + l_{32}F_2 + e_3 \\ x_4 &= 0F_1 + l_{42}F_2 + e_4 \end{aligned}$$

In matrix form:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & 0 \\ 0 & l_{32} \\ 0 & l_{42} \end{bmatrix} \times \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

Any correlation between items from different factors—say, x_1 and x_3—occurs only because of a relationship between F_1 and F_2. Confirmatory factor analysis assesses the degree to which the proposed model is consistent with the data in terms of relationships between items belonging to the same factor (say, x_1 and x_2) and the relationship between items belonging to different factors (say, x_1 and x_3). For instance, if an item actually measures two factors, then its relationship with an item from the second factor will be high, leading to a poor fit for the overall model. Confirmatory factor analysis assesses internal consistency across items from the same factors and external consistency across items from different factors.

Internal and External Consistency in Confirmatory Factor Analysis

The product rule for internal consistency for correlation between two indicators i and j of construct ξ, where t is the true score, is

$$\rho_{ij} = \rho_{i\xi}\rho_{j\xi}.$$

The product rule for external consistency for correlation between two indicators i and p, p being an indicator of another construct ξ*, is

$$\rho_{ip} = \rho_{i\xi}\rho_{\xi\xi^*}\rho_{p\xi^*}.$$

SOURCE: Gerbing and Anderson (1988).

Expressions in Structural Equation Modeling

Reliability of item (Bagozzi, 1991)

$$\rho_{x_i} = \frac{\lambda_{x_i}^2}{\lambda_{x_i}^2 + \mathrm{var}(\delta_i)}$$

Reliability of measure with p items (Bagozzi, 1991)

$$\rho_c = \frac{\left(\sum_{i=1}^{p} \lambda_{x_i}\right)^2}{\left(\sum_{i=1}^{p} \lambda_{x_i}\right)^2 + \sum_{i=1}^{p} \mathrm{var}(\delta_i)}$$

Average variance extracted for measure with p items (Fornell & Larcker, 1981)

$$\rho_{vc} = \frac{\sum_{i=\mathrm{i}}^{p} \lambda_{x_i}^2}{\sum_{i=1}^{p} \lambda_{x_i}^2 + \sum_{i=1}^{p} \mathrm{var}(\delta_i)}$$

ξ = latent variable
x = observed variable
δ = error
λ = loading of x on ξ
var = variance

Alternative Overall Fit Indexes

Normed fit index (Bentler & Bonnett, 1980)

$$\frac{T_b^2 - T_n^2}{T_b^2}$$

T_b^2 = chi-square for baseline model
T_n^2 = chi-square for hypothesized model

Incremental fit index (Bollen, 1989)

$$\frac{T_b^2 - T_n^2}{T_b^2 - df_n}$$

df_b = degrees of freedom for baseline model
df_n = degrees of freedom for hypothesized model

Comparative fit index (Bentler, 1990)

$$\frac{(T_b^2 - df_b) - (T_n^2 - df_n)}{T_b^2 - df_b}$$

Tucker-Lewis index (Tucker & Lewis, 1973)

$$\frac{T_b^2/df_b - T_n^2/df_n}{T_b^2/df_b - 1}$$

Appendix 1.2

Sample Scales (Response Categories Adapted to 7-Point Scales for Convenience)

Consumer independent judgment-making (Manning, Bearden, & Madden, 1995[1])

	Strongly Disagree						*Strongly Agree*
Prior to purchasing a new brand, I prefer to consult a friend that has experience with the new brand.	1	2	3	4	5	6	7
When it comes to deciding whether to purchase a new service, I do not rely on experienced friends or family members for advice.	1	2	3	4	5	6	7
I seldom ask a friend about his or her experiences with a new product before I buy the new product.	1	2	3	4	5	6	7
I decide to buy new products and services without relying on the opinions of friends who have already tried them.	1	2	3	4	5	6	7
When I am interested in purchasing a new service, I do not rely on my friends or close acquaintances that have already used the new service to give me information as to whether I should try it.	1	2	3	4	5	6	7
I do not rely on experienced friends for information about new products prior to making up my mind about whether or not to purchase.	1	2	3	4	5	6	7

Consumer novelty (Manning et al., 1995[1])

	Strongly Disagree						*Strongly Agree*
I often seek out information about new products and brands.	1	2	3	4	5	6	7
I like to go to places where I will be exposed to information about new products and brands.	1	2	3	4	5	6	7
I like magazines that introduce new brands.	1	2	3	4	5	6	7
I frequently look for new products and services.	1	2	3	4	5	6	7
I seek out situations in which I will be exposed to new and different sources of product information.	1	2	3	4	5	6	7

I am continually seeking new product experiences.	1	2	3	4	5	6	7
When I go shopping, I find myself spending very little time checking out new products and brands.	1	2	3	4	5	6	7
I take advantage of the first available opportunity to find out about new and different products.	1	2	3	4	5	6	7

Material values—Defining success (Richins & Dawson, 1992[2])

	Strongly Disagree						*Strongly Agree*
I admire people who own expensive homes, cars, and clothes.	1	2	3	4	5	6	7
Some of the most important achievements in life include acquiring material possessions.	1	2	3	4	5	6	7
I don't place much emphasis on the amount of material objects people own as a sign of success.	1	2	3	4	5	6	7
The things I own say a lot about how well I'm doing in life.	1	2	3	4	5	6	7
I like to own things that impress people.	1	2	3	4	5	6	7
I don't pay much attention to the material objects other people own.	1	2	3	4	5	6	7

Material values—Acquisition centrality (Richins & Dawson, 1992[2])

	Strongly Disagree						*Strongly Agree*
I usually buy only the things I need.	1	2	3	4	5	6	7
I try to keep my life simple, as far as possessions are concerned.	1	2	3	4	5	6	7
The things I own aren't all that important to me.	1	2	3	4	5	6	7
I enjoy spending money on things that aren't practical.	1	2	3	4	5	6	7
Buying things gives me a lot of pleasure.	1	2	3	4	5	6	7
I like a lot of luxury in my life.	1	2	3	4	5	6	7
I put less emphasis on material things than most people I know.	1	2	3	4	5	6	7

Material values—Pursuit of happiness (Richins & Dawson, 1992[2])

	Strongly Disagree						*Strongly Agree*
I have all the things I really need to enjoy life.	1	2	3	4	5	6	7
My life would be better if I owned certain things I don't have.	1	2	3	4	5	6	7
I wouldn't be any happier if I owned nicer things.	1	2	3	4	5	6	7
I'd be happier if I could afford to buy more things.	1	2	3	4	5	6	7
It sometimes bothers me quite a bit that I can't afford to buy all the things I'd like.	1	2	3	4	5	6	7

Value consciousness (Lichtenstein, Ridgway, & Netemeyer, 1993[3])

	Strongly Disagree						*Strongly Agree*
I am very concerned about low prices, but I am equally concerned about product quality.	1	2	3	4	5	6	7
When grocery shopping, I compare the prices of different brands to be sure I get the best value for the money.	1	2	3	4	5	6	7
When purchasing a product, I always try to maximize the quality I get for the money I spend.	1	2	3	4	5	6	7
When I buy products, I like to be sure that I am getting my money's worth.	1	2	3	4	5	6	7
I generally shop around for lower prices on products, but they still must meet certain quality requirements before I will buy them.	1	2	3	4	5	6	7
When I shop, I usually compare the "price per ounce" information for brands I normally buy.	1	2	3	4	5	6	7
I always check prices at the grocery store to be sure I get the best value for the money I spend.	1	2	3	4	5	6	7

Price consciousness (Lichtenstein et al., 1993[3])

	Strongly Disagree						*Strongly Agree*
I am not willing to go to extra effort to find lower prices.	1	2	3	4	5	6	7

I will grocery shop at more than one store to take advantage of low prices.	1	2	3	4	5	6	7
The money saved by finding lower prices is usually not worth the time and effort.	1	2	3	4	5	6	7
I would never shop at more than one store to find low prices.	1	2	3	4	5	6	7
The time it takes to find low prices is usually not worth the effort.	1	2	3	4	5	6	7

Coupon proneness (Lichtenstein et al., 1993[3])

	Strongly Disagree						*Strongly Agree*
Redeeming coupons makes me feel good.	1	2	3	4	5	6	7
I enjoy clipping coupons out of the newspaper.	1	2	3	4	5	6	7
When I use coupons, I feel that I am getting a good deal.	1	2	3	4	5	6	7
I enjoy using coupons regardless of the amount I save by doing so.	1	2	3	4	5	6	7
Beyond the money I save, redeeming coupons gives me a sense of joy.	1	2	3	4	5	6	7

Sale proneness (Lichtenstein et al., 1993[3])

	Strongly Disagree						*Strongly Agree*
If a product is on sale, that can be a reason for me to buy it.	1	2	3	4	5	6	7
When I buy a brand that's on sale, I feel that I am getting a good deal.	1	2	3	4	5	6	7
I have favorite brands, but most of the time I buy the brand that's on sale.	1	2	3	4	5	6	7
I am more likely to buy brands that are on sale.	1	2	3	4	5	6	7
Compared to most people, I am more likely to buy brands that are on special.	1	2	3	4	5	6	7

Consumer ethnocentrism (Shimp & Sharma, 1987[4])

	Strongly Disagree						*Strongly Agree*
American people should always buy American-made products instead of imports.	1	2	3	4	5	6	7

Only those products that are unavailable in the U.S. should be imported.	1	2	3	4	5	6	7
Buy American-made products. Keep America working.	1	2	3	4	5	6	7
American products, first, last, and foremost.	1	2	3	4	5	6	7
Purchasing foreign-made products is un-American.	1	2	3	4	5	6	7
It is not right to purchase foreign products.	1	2	3	4	5	6	7
A real American should always buy American-made products.	1	2	3	4	5	6	7
We should purchase products manufactured in America instead of letting other countries get rich off us.	1	2	3	4	5	6	7
It is always best to purchase American products.	1	2	3	4	5	6	7
There should be very little trading or purchasing of goods from other countries unless out of necessity.	1	2	3	4	5	6	7
Americans should not buy foreign products, because this hurts American business and causes unemployment.	1	2	3	4	5	6	7
Curbs should be put on all imports.	1	2	3	4	5	6	7
It may cost me in the long run but I prefer to support American products.	1	2	3	4	5	6	7
Foreigners should not be allowed to put their products in our markets.	1	2	3	4	5	6	7
Foreign products should be taxed heavily to reduce their entry into the U.S.	1	2	3	4	5	6	7
We should buy from foreign countries only those products that we cannot obtain within our own country.	1	2	3	4	5	6	7
American consumers who purchase products made in other countries are responsible for putting their fellow Americans out of work.	1	2	3	4	5	6	7

Need for cognition (Perri & Wolfgang, 1988[5])

	Strongly Disagree						*Strongly Agree*
I would prefer complex to simple problems.	1	2	3	4	5	6	7
I like to have the responsibility of handling a situation that requires a lot of thinking.	1	2	3	4	5	6	7

Thinking is not my idea of fun.	1	2	3	4	5	6	7
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.	1	2	3	4	5	6	7
I try to anticipate and avoid situations where there is likely chance I will have to think in depth about something.	1	2	3	4	5	6	7
I find satisfaction in deliberating hard and for long hours.	1	2	3	4	5	6	7
I only think as hard as I have to.	1	2	3	4	5	6	7
I prefer to think about small, daily projects to long-term ones.	1	2	3	4	5	6	7
I like tasks that require little thought once I've learned them.	1	2	3	4	5	6	7
The idea of relying on thought to make my way to the top appeals to me.	1	2	3	4	5	6	7
I really enjoy a task that involves coming up with new solutions to problems.	1	2	3	4	5	6	7
Learning new ways to think doesn't excite me very much.	1	2	3	4	5	6	7
I prefer my life to be filled with puzzles that I must solve.	1	2	3	4	5	6	7
The notion of thinking abstractly is appealing to me.	1	2	3	4	5	6	7
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.	1	2	3	4	5	6	7
I feel relief rather than satisfaction after completing a task that required a lot of mental effort.	1	2	3	4	5	6	7
It's enough for me that something gets the job done; I don't care how or why it works.	1	2	3	4	5	6	7
I usually end up deliberating about issues even when they do not affect me personally.	1	2	3	4	5	6	7

Consumer susceptibility to interpersonal influence (Bearden, Netemeyer, & Teel, 1989[6])

	Strongly Disagree						*Strongly Agree*
I often consult other people to help choose the best alternative available from a product class.	1	2	3	4	5	6	7

If I want to be like someone, I often try to buy the same brands that they buy.	1	2	3	4	5	6	7
It is important that others like the products and brands I buy.	1	2	3	4	5	6	7
To make sure I buy the right product or brand, I often observe what others are buying and using.	1	2	3	4	5	6	7
I rarely purchase the latest fashion styles until I am sure my friends approve of them.	1	2	3	4	5	6	7
I often identify with other people by purchasing the same products and brands they purchase.	1	2	3	4	5	6	7
If I have little experience with a product, I often ask my friends about the product.	1	2	3	4	5	6	7
When buying products, I generally purchase those brands that I think others will approve of.	1	2	3	4	5	6	7
I like to know what brands and products make good impressions on others.	1	2	3	4	5	6	7
I frequently gather information from friends or family about a product before I buy.	1	2	3	4	5	6	7
If other people can see me using a product, I often purchase the brand they expect me to buy.	1	2	3	4	5	6	7
I achieve a sense of belonging by purchasing the same products and brands that others purchase.	1	2	3	4	5	6	7

Notes

1. From Manning, K. C., Bearden, W. O., & Madden, T. J., Consumer innovativeness and the adoption process, in *Journal of Consumer Psychology*, *4*(4), copyright © 1995, pp. 329–345. Reprinted by permission.

2. From Richins, M. L., & Dawson, S., Materialism as a consumer value: Measure development and validation, in *Journal of Consumer Research*, *19*, pp. 303–316, copyright © 1992. Reprinted by permission of the American Marketing Association.

3. From Lichtenstein, D. R., Ridgway, N. M., & Netemeyer, R. G., Price perceptions and consumer shopping behavior: A field study, in *Journal of Marketing Research*, *30*, pp. 234–245, copyright © 1993. Reprinted by permission of the American Marketing Association.

4. From Shimp, T. A., & Sharma, S., Consumer ethnocentrism: Construction and validation of the CETSCALE, in *Journal of Marketing Research*, *24*,

pp. 280–289, copyright © 1987. Reprinted by permission of the American Marketing Association.

5. From Perri, M., & Wolfgang, A. P., A modified measure of need for cognition, in *Psychological Reports*, *62*, pp. 955–957, copyright © 1988. Reprinted by permission.

6. From Bearden, W. O., Netemeyer, R. G., & Teel, J. E., Measurement of consumer susceptibility to interpersonal influence, in *Journal of Consumer Research*, *15*, pp. 473–481, copyright © 1989. Reprinted by permission.