MEASUREMENT OF PSYCHOLOGICAL CONSTRUCTS

To evaluate hypothesized relationships between abstract psychological constructs, researchers must translate the relevant constructs into concrete, observable variables. This is an essential first step in testing hypotheses against external reality because critical observations cannot be made without first specifying the variables to be observed.

Observable variables can be classified along several dimensions, including: the degree of control exerted by the researcher and the status of the variable as independent or dependent. With respect to control, nonexperimental or measured variables are observed passively by the researcher, whereas experimental or manipulated variables are directly controlled by the researcher. Each variable can also be classified as either an independent variable (i.e., predictor) or as a dependent variable (i.e., criterion).

These two classifications overlap somewhat. Criterion variables are always measured (i.e., nonexperimental) variables, such as degree of anxiety, number of words recalled, and psychiatric diagnosis. Predictor variables, on the other hand, can be either measured or manipulated. Predictors are manipulated (i.e., true independent or experimental variables) when the researcher assigns different treatments or conditions (e.g., study time, instructions) to subjects, and measured variables when the researcher examines pre-existing or naturally-occurring differences between subjects (e.g., measures of exposure to environmental stressors, reported spontaneous use of imagery). This chapter reviews issues and techniques involved in measuring psychological constructs, whether these be independent or dependent variables. A later chapter examines issues involved in manipulating experimental independent variables.

BASICS OF MEASUREMENT

Measurement involves the assignment of numbers or labels to reflect the kind or amount of some underlying property. In psychological research, the units being assessed (i.e., the cases) are often human participants or other animals, but the cases can also be non-living objects (e.g., concreteness of words, number of books in home, school size, behavior of single neurons). Measurement begins with the development of an operational definition for the theoretical construct of interest. Operational definitions define constructs in terms of the procedures used to measure the constructs. Several meta-theoretical issues and criticisms with regard to operational definitions are discussed in Chapter 5. This chapter assumes that operational definitions and associated measurements are desirable and describes specific techniques for evaluating and developing psychological measures.

Levels of Measurement

Psychological measures are either numerical or categorical in nature. Numerical variables are those in which the measured trait varies with the magnitude of the numbers assigned to cases. For example, higher scores on a test of anxiety (the numerical measure) indicate greater amounts of anxiety (the underlying construct) than do lower scores; similarly, the greater the number of words recalled (the numerical measure) the greater the memory (the underlying construct). General types of numerical variables include: frequency measures (e.g., number of words recalled, number of arguments by couples), latency measures (e.g., reaction time to name pictures or words, duration of behaviors or mood states), and intensity or strength measures (e.g., magnitude estimation of stimulus intensity, rated liking for a person). Ratings using numerical scales (e.g., 1 to 7) are sometimes referred to as Likert scales.

Psychological studies also involve variables that are categorical rather than numerical. Categorical variables involve classification of cases into distinct groups that usually vary along several (perhaps unspecified) dimensions, rather than a single quantitative dimension. For example, researchers interested in the prediction of depression would classify people as depressed or not depending on the presence of some critical number of symptoms. On the predictor side, such categorical variables as marital status, attachment style, and psychiatric diagnosis are common in psychological research.

Although numbers might be used to label the levels of categorical variables, especially when there are numerous classes (e.g., psychiatric diagnoses), the numbers have limited quantitative meaning and simply provide convenient labels for the distinct groups. Effective categorical variables, especially if they involve subtle distinctions, require considerable attention to the definition and labelling of categories, the development of coding schemes for responses, and clear rules for classification of behaviors.

The boundary between categorical and numerical variables can be fuzzy. In the case of depression, for example, some researchers assign scores reflecting the degree of depression (a numerical variable), whereas others label people as clinically depressed or not (a categorical variable) or define different types of depression. The distinction is also fuzzy because categories often play an important role in the construction of numerical variables. For example, frequency counts of different classroom behaviors require adequate definitions of on-task, out-of-seat, and other classes of behavior that are to be counted by the observers.

Numerical variables can be further divided into three finer types, called ordinal, interval, and ratio scales. These narrower categories are determined by which properties of numbers apply to the scale. Ordinal scales only consider the order of numbers; that is, a score of 8 indicates more of the variable than does 6 which in turn indicates more than 4. Interval scales involve the magnitude of differences between numbers; that is, the difference between scores of 8 and 6 on the scale is the same as the difference between 6 and 4. Ratio scales add an absolute difference which permits assertions that scores of 8 reflect twice as much of the trait as scores of 4. These finer distinctions will not be considered here, but you may come across them in articles or books on methods and statistics. For example, some writers argue that parametric statistics (e.g., t-test, ANOVA) should only be performed on interval or ratio data and that nonparametric statistics (e.g., sign test, Wilcoxen) should be used for ordinal data.

Types of Measures

Psychologists have been creative in the development of measures for theoretical constructs, and there is no neat taxonomy (i.e., classification system) for the diverse measures that have been developed. Nonetheless, several general categories can be used to classify different psychological measures. Such a listing does not preclude the use of alternative methods. Scientific advances often depend on the development of novel ways to measure theoretical constructs, so do not be constrained by the following taxonomy. Any effort spent trying to think of new ways to measure constructs will be well rewarded!

Self-report measures. Many psychological measures fall under the general heading of self-reports. The essential characteristic of self-report measures is that subjects are asked to report directly about internal psychological states or traits. Personality tests and attitude scales ask people whether statements are true for themselves (e.g., I often act without thinking, I would be disturbed if one of my relatives married an oriental person). Many questionnaires and surveys also fall into this category, asking people to report about internal states or events in their lives (e.g., My mother was very strict with me, I voted for the Conservatives in the last election).

Self-reports are also used by cognitive researchers to obtain convergent measures of inferred mental events (e.g., indicate whether or not you had a mental image when you studied each of the following words during learning) or to exclude subjects who might have seen through the purpose of the study (e.g., did you expect the surprise memory test). Many standardized tests, surveys, questionnaires, attitude scales, and so on are self-report instruments.

Ratings by others. Psychologists often ask respondents who are familiar with the subject to provide ratings. With children, for example, parents or teachers might be asked to rate or classify children with respect to sociability, aggression, or some other psychological dimension. There are a variety of rating instruments that have been developed especially for this purpose and for which norms are available. Conners, for example, has developed parent and teacher scales to rate various psychopathologies common in childhood, such as attention-deficit-hyperactivity disorder (ADHD) and conduct disorder.

One variant of the rating method that has been used in developmental, educational, and clinical research with children is the peer nomination technique. Respondents familiar with a group of individuals (e.g., children in their class) are asked to identify (i.e., nominate) those children who best represent some particular category of children (e.g., liked children, disliked children, aggressive children). Each person's score is the number of individuals nominating them; for example, the number of children identifying a particular student as aggressive or as likeable.

Objective tests. Standardized or objective tests provide another kind of frequently used measure, especially in cognitive domains. Mathematical aptitude, reading ability, general intelligence, imagery ability, language comprehension, school achievement tests, motor skills, and diverse other cognitive constructs can be assessed by objective tests in which respondents complete multiple items relevant to the domain being assessed. There are correct answers for the questions and scores are the number of items correct, percentages, or other scores based on number correct (e.g., number correct minus percentage of number incorrect to adjust for guessing).

Laboratory measures. In addition to the paper-and-pencil tests just described, psychologists in such areas as physiology, perception, cognition, and abnormal often use physical equipment to obtain measures related to various psychological traits. Physiological measures include various brain imaging methods (e.g., electroencephalogram or EEG, magnetic resonance imaging or MRI scans), biochemical measures (e.g., quantities of neurotransmitters), and activity of the peripheral nervous system (e.g., muscle tension).

Experimental researchers in perception, cognition, and an increasingly wide range of other areas use various tasks that involve the presentation of stimuli and recording of responses. Scores are based on such measures as the frequency of responses (e.g., number of words recalled, number of stimuli correctly identified) and reaction time (RT) or latency to perform the task. There are several general purpose computer programs (e.g., Micro-Experimental Language or MEL) that help researchers to develop laboratory measures.

Although laboratory measures are generally obtained in laboratory studies, such measures can be adapted to other settings (e.g., group tests). The mental rotations task, for example, involves deciding whether two or more stimuli at different orientations are identical. The task has been used in a laboratory setting, but has also been adapted to paper and pencil tests of spatial ability and intelligence. Similarly, Katz (1979) described a procedure for obtaining RT data from groups of subjects performing cognitive tasks. Subjects perform a task (e.g., stating whether sentences are true or false) as quickly as possible and are stopped after an appropriate period of time. The number of items completed in the allotted time provides an RT measure. Although Katz describes the procedure in the context of classroom demonstrations, the methods would work for group research studies. Computers and computer networks are also making it increasingly easy to automate the administration of laboratory tasks to groups of subjects (e.g., naming latencies, decision RTs) and to incorporate such measures into standardized testing situations.

Observational measures. Researchers can observe directly the behaviors of interest. Such methods have been particularly important on research with children, nonhuman species, and other subjects who might have difficulty providing self-reports. Observational methods play a central role in applied research and has been especially championed by behavioral psychologists (e.g., see the Journal of Applied Behavior Analysis). Researchers interested in behaviors in natural settings also make widespread use of observers. To be effective, observational measures require steps to ensure adequate objectivity, reliability, and validity (e.g., clear definitions of the behaviors, systematic training and monitoring of observers).

Verbal reports or protocol measures. Numerous researchers have made use of written or spoken dialogue as the basis for quantitative or categorical measures. The dialogue might be tape recorded (e.g., tape of therapy sessions or of children in a nursery school), written by the subject (e.g., diaries), or be obtained from archive sources (e.g., essays, letters, speeches, books, articles). Content analysis methods are used to identify and classify particular idea units (e.g., negative self-statements, references to concrete events), and these idea units are used to produce scores related to whatever underlying constructs are of interest (Holsti, 1969). Truax and his colleagues, for example, used tape recordings of therapy sessions to test some of Carl Rogers's hypotheses about empathy, concreteness of language, and other characteristics of effective therapists (Truax, 1961). Verbal reports and like measures play a central role in what are now known collectively as qualitative research methods.

Cognitive researchers interested in problem solving, thinking, and other complex cognitive tasks also make extensive use of verbal protocols. Subjects talk aloud while they try to solve some demanding task (e.g., puzzles such as the Towers of Hanoi). Ericsson and Simon (1984) have proposed a psychological model of the cognitive processes that underlie the production of such protocols. One important consideration is how accessible the sought-after information is to consciousness. Researchers cannot assume that subjects have direct access to all psychological mechanisms that underlie behavior and experience.

Content analysis has a long history in psychology (e.g., Allport, 1942), and contemporary use of the method is increasingly sophisticated and theory-driven. For example, computer programs have been developed to perform some content analyses (e.g., the CHILDES program examines children's language, and Simon has programs that analyze subject protocols from problem-solving sessions).

Verbal reports and content analysis are superficially very similar to introspection, an older and discredited method. Important differences between contemporary use of verbal reports and earlier introspectionism are the use of naive subjects in current research rather than theoretically sophisticated subjects in the earlier literature, and an emphasis on the contents of consciousness rather than having the introspectionist make inferences about underlying processes or mechanisms (e.g., imageless thought). Such considerations help current researchers to avoid some of the problems of introspectionism. Nonetheless, the negative history of introspectionism (e.g., irreconcilable disagreements about whether thoughts were imageless or not) should teach us to use caution in interpreting verbal reports (or any measure for that fact).

These examples of different kinds of measures demonstrate that there will often be multiple ways to measure the same construct. Whenever possible and practical, researchers should use multiple measures in their studies, a practice known as convergent operationism. Researchers should also pay careful attention to the quality of their measures because poor measurement is a common problem in behavioral research. Measurement quality can be assessed in terms of the reliability and validity of the measures.

RELIABILITY

Any measurement procedure should provide reliable information. Reliability refers to the consistency of measurement across items, time, raters, observers, or some other dimension that could add variability to scores. The essential assumption underlying traditional discussions of reliability is that an observed score (y) represents in part the individual's underlying true score (y_t) and in part random variation or error (e); that is, $y = y_t + e$. Sources of random variation include: distractions and other random environmental influences, momentary variations in attention, and idiosyncrasies in items (e.g., whether subjects have particular familiarity with specific items, perhaps because they were previously exposed to those items). Researchers try to minimize these sources of error variability in order to maximize the contribution of true scores to variability in the observed scores.

A basic assumption of this model is that people (or whatever entitity is being measured) possess stable characteristics or traits that persist across time and situations (i.e., the true scores), although distinctions between stable traits and momentary states have been made in several areas (e.g., state vs. trait anxiety). I first consider the reliability of numerical scores, which are amenable to correlational analysis, and then examine some special problems that arise with observational measures that are categorical in nature (e.g., presence or absence of specified behaviors).

Measures of Reliability

The correlation coefficient measures the agreement between two numerical scores and is widely used in the examination of reliability. Reliability is assessed by obtaining two or more measurements using the same instrument on a sample of subjects and then determining the correlation between the resulting scores. Researchers generally seek reliabilities of .80 or better, although a satisfactory value depends somewhat on how the two scores were obtained and on the domain under investigation.

Stability across time. One fundamental aspect of reliability is stability across time; do subjects maintain their relative ranking on the scale when tested on two separate occasions? This type of reliability is measured by test-retest reliability coefficients. To measure the stability of scores across time, the same test or equivalent versions of a test are administered to the same sample of subjects. The correlation between the two sets of scores provides an index of test-retest reliability. In general, the longer the time interval between testings, the lower the correlation. However the effect of the time interval on scores will depend on the stability of the underlying trait as well as on the measure itself. Mood, for example, might be expected to fluctuate from moment to moment, whereas more enduring aspects of personality should be (by definition) more stable.

Split-half measures of internal consistency. Several reliability indices measure the consistency of responses to individual "items" on a test. Although consistency depends somewhat on momentary fluctuations in performance, internal consistency measures of reliability reflect primarily the homogeneity of the test; that is, whether the items on the test assess a single underlying dimension or multiple dimensions. Internal consistency is relevant not only to reliability, but also to construct validity, as discussed later.

One measure of homogeneity is split-half reliability, in which a score based on odd-numbered items is correlated with a score based on even-numbered items (or some other division of the items into two equivalent sets). Most statistical packages permit researchers to generate the scores necessary to determine split-half reliability. For example, the SPSS COMPUTE command and various mathematical operators and functions can be used to produce split-half scores [e.g., COMPUTE odd = MEAN(r1 r3 r5 ...) or COMPUTE odd = r1+r3+r5...]. The MEAN function can include an optional number (e.g., MEAN.10, MEAN.21) to indicate how many of the variables in parentheses must be nonmissing for the computed score to be valid. If too few nonmissing variables occur, then the computed score is coded as missing.

Once calculated, odd and even scores are correlated using standard statistical methods. The resulting correlation coefficient (e.g., .60), however, does not reflect accurately the reliability of the entire test because the two scores being correlated are based on only half of the total number of items on the test. In general, reliability increases as the number of items on the scale increases (Ghiselli, Campbell, & Zedek, 1981); hence, the reliability for the full test is greater than the reliability of the two halves.

The reliability of the entire test can be estimated from the correlation between the two halves using the Spearman-Brown formula. The Spearman-Brown formula for split-half reliability is $2\times r_{oe}/(1+r_{oe})$, where r_{oe} is the correlation between the odd and even halves of the test. For example, if the correlation between odd and even halves of a test is .60, then the split-half reliability for the entire test is $2\times.60/(1+.60) = 1.20/1.60 = .75$.

A more general form of the Spearman-Brown formula is $n\times r/(1+(n-1)\times r)$, where n is the number of times that the test is lengthened or shortened. Substituting 2 for n gives the split-half version of the formula (i.e., $2\times r/(1+(2-1)r) = 2\times r/(1+r)$).

The general version of the Spearman-Brown formula can be used to estimate what the reliability of a test would be if the measure were lengthened or shortened. If a test with 10 items has an estimated reliability of .60, for example, then a comparable test of 40 items would have an internal consistency reliability of $4\times.60/(1+(4-1)\times.60) = 2.40/2.80 = .86$. The value for n is obtained by dividing the length of the estimated instrument by the length of the observed instrument on which the r is based (i.e., 40/10 = 4). Increasing the length of a test has its strongest effects on measures with low initial reliability (see Figure 9.1 in Ghiselli et al., 1981, p. 234) and is one of the most effective methods that researchers can use to improve reliability. Pilot testing of measures allows researchers to determine whether or not additional items should be added to improve reliability, and should be done whenever possible with new measures.

Cronbach's Alpha and internal consistency. An alternative measure of internal consistency is provided by Cronbach's coefficient alpha. Alpha is a common measure of reliability in many areas of psychology and is available in many statistical packages, including SPSS.

Box 1 shows the form of the SPSS RELIABILITY command, which produces Cronbach's alpha, item-total correlations, and various other indices of internal consistency.

```
RELIABILITY VARIABLES = varlist
        [/SCALE(name) = vars] [/SCALE...]
        [/MODEL = {ALPHA**  SPLIT(n)}]
        [/STAT = DESC CORR SCALE ANOVA]
        [/SUMMARY = MEANS VARIANCE CORR
                TOTAL ALL]
```

Box 1. SPSS Reliability Command.

Cronbach's alpha equals the average of all possible split-half reliability coefficients, Spearman-Brown corrected for the length of the halves (Ghiselli et al., 1981, p. 258). Alpha can be calculated as: $(n/(n-1)) \times (1 - (\Sigma s^2) / s^2_t)$, where n is the number of items, Σs^2 is the sum of the variances for individual items, and s^2_t is the variance of the total scale (i.e., scores based on the sum of the individual items). The variance of the total scores depends on the number of items, the standard deviations of the individual items, and the intercorrelations among the individual items (see Ghiselli et al, 1981, p. 158 for formula). For tests with dichotomous items (e.g., right-wrong, yes-no), the Kuder-Richardson version of this formula can be used.

In essence, alpha increases as the intercorrelation among the items increases, as the number of equivalently related items increases, and as the variability of individual increases. Researchers can therefore improve reliability by increasing the number of equivalent items on their measures, creating items that discriminate well among people, and by strengthening relations between items, for example, by rewording items to ensure that they all tap a common underlying construct.

The SPSS reliability command can also be used to obtain the correlations between each of the items and a total score calculated without that item (TOTAL keyword). The item-total correlation provides an item-by-item measure of the relatedness of each individual item to the total scale and is very useful during the construction of a test or pre-testing of measures. Items that do not correlate highly with other items can be eliminated or revised, thus improving the homogeneity of the test. Such correlations can be obtained in SPSS with the RELIABILITY procedure or by using the CORRELATION procedure (e.g., CORR r1 TO r10 WITH total).

```
RELIABILITY VAR = i1 TO i22 /SUMMARY ALL

        # OF CASES =              32.0

ITEM MEANS            MEAN   MINIMUM   MAXIMUM   RANGE     MAX/MIN   VARIANCE
                    3.6506    2.1563    6.0625    3.9063    2.8116     1.2509

ITEM VARIANCES        MEAN   MINIMUM   MAXIMUM   RANGE     MAX/MIN   VARIANCE
                     .9611     .4264    1.4153     .9889    3.3191      .0994

INTER-ITEM
COVARIANCES           MEAN   MINIMUM   MAXIMUM   RANGE     MAX/MIN   VARIANCE
                     .0688   -.4556     .6573    1.1129   -1.4425      .0329

INTER-ITEM
CORRELATIONS          MEAN   MINIMUM   MAXIMUM   RANGE     MAX/MIN   VARIANCE
                     .0682   -.4417     .5464     .9880   -1.2371      .0348

ITEM-TOTAL STATISTICS
              SCALE        SCALE      CORRECTED
              MEAN       VARIANCE       ITEM-        SQUARED        ALPHA
             IF ITEM      IF ITEM       TOTAL        MULTIPLE      IF ITEM
             DELETED      DELETED    CORRELATION   CORRELATION     DELETED
I1           75.2813      51.5635       .0312         .5655         .6385
I2           77.2500      51.1613       .0716         .7602         .6332
I3           76.0938      50.2167       .2022         .7863         .6197
I4           77.8125      54.6734      -.1968         .6210         .6576
I5           77.1563      50.2651       .1912         .8683         .6206
I6           75.5313      49.7409       .1440         .7946         .6265
I7           75.7500      46.5161       .3081         .9207         .6047
I8           77.5938      53.4748      -.1009         .4950         .6465
I9           75.6563      48.0393       .3262         .6818         .6057
I10          76.6563      50.5554       .2100         .8618         .6197
I11          76.0000      49.2258       .1734         .7490         .6230
I12          77.5313      48.0635       .2111         .7851         .6187
I13          78.1563      50.8458       .0539         .6552         .6388
I14          78.1250      43.9194       .5411         .8324         .5730
I15          76.1563      47.4264       .4363         .8518         .5963
I16          77.9688      48.6764       .1905         .8088         .6212
I17          77.1250      48.3710       .2852         .8728         .6099
I18          75.2500      46.2581       .3492         .8205         .5993
I19          77.7813      45.9829       .4123         .8602         .5922
I20          75.9063      48.2812       .2458         .7360         .6138
I21          77.5313      49.6119       .2606         .8390         .6144
I22          74.2500      46.9032       .2831         .7324         .6084

RELIABILITY COEFFICIENTS    22 ITEMS

ALPHA =   .6292          STANDARDIZED ITEM ALPHA =   .6169
```

Box 2. Coefficient Alpha Measure of Reliability.

Box 2 shows an SPSS reliability analysis for a preliminary version of an impulsivity measure developed for a class research project. Briefly, 32 subjects responded on a 7-point Likert scale to 22 items that had been proposed as elements of impulsivity. High ratings indicated higher impulsivity for 11 items and low ratings indicated higher impulsivity for the remaining 11 items. Negative items were reversed (i.e., score = 8 - rating) before being analyzed. Alpha is reported at the bottom of the printout; the observed value of .6292 is modest. Examination of the item-total statistics indicates that items 4 and 8 correlate negatively with the total score and that other items had low rs with the total. Perhaps these items do not belong on the scale or need to be reworded.

Cronbach's alpha is based on the correlations among the individual items. It is possible to get the entire correlation matrix printed out, but the matrix can be difficult to comprehend with large numbers of items. In the example of Box 2, there would be 22 rows and 22 columns in the correlation matrix, for a total of 231 individual correlations. Instead, SPSS RELIABILITY reports the mean correlation and the range of correlations. The mean r in Box 2 is only .0682, with a range from -.44 to +.55. It is the negative and low correlations that produce the modest value for alpha.

SPSS RELIABILITY also reports a standardized alpha, .6169 in Box 2. The unstandardized alpha depends on the means and variances of individual items (RELIABILITY reports averages and ranges for these statistics), as well as on their intercorrelation. The standardized alpha shows how reliable the measure would be if all items were standardized before being summed or averaged. When items vary dramatically in terms of means and variances, standardized measures can be more appropriate and more reliable than unstandardized measures.

Reliability and Between-measure Correlations

In addition to its importance for measurement purposes, reliability plays an important role in the interpretation of relations between variables (e.g., between x and y). The strength of a relation between two variables can be limited by the reliabilities of the individual measures. For example, the maximum possible correlation between two variables with reliabilities of .5 is .5, and the maximum correlation between variables with reliabilities of .6 and .4 is .49. Ghiselli et al. (1981, p. 243) summarize maximum intercorrelations between variables with different degrees of reliability.

A correction for attenuation is sometimes used to estimate the correlation between true scores from observed correlations between variables with known reliabilities. The formula is $r_{xy(true)} = r_{xy}/SQRT(r_{xx}r_{yy})$, where r_{xx} and r_{yy} are the reliabilities for x and y, respectively. Given a correlation of .4 between measures with reliabilities of .4 and .5, $r_{xy(true)} = .4/SQRT(.4\times.5) = .4/.45 = .89$, a considerable improvement. The denominator is the maximum value for the correlation, so the correction divides the observed r by the maximum possible r given the observed levels of reliability. Caution should be used in interpreting correlations that have been corrected for attenuation.

One lesson from the relationship between reliability and between-measure correlations is that researchers should put considerable effort into selecting and developing reliable measures. Using measures that have poor reliability makes it very difficult or impossible to obtain reasonable correlations between the variables of interest, even if the underlying relations are consistent with theoretical expectations.

Measuring Reliability in Observational Studies

In the case of observation measures, reliability refers to the degree of agreement across observers about the occurrence or strength of the behavior. Measuring reliability of observational studies has been controversial and surprisingly complex.

Interobserver agreement. One simple measure that was once used widely is the percentage of interobserver agreement. This is the frequency of agreements divided by the total number of opportunities for an agreement. Across 100 observation intervals, for example, observers one and two might agree on 5 occurrences of the behavior and on 85 non-occurrences. The percent agreement is 100 × (5+85)/100 = 90%.

One serious limitation of % agreement is that the chance % is highly sensitive to the relative frequency with which different behaviors are judged to have occurred. That is, very high % agreement scores can be obtained just by chance. Chance varies between 100% and 50%, making interpretation of observed agreement statistics problematic.

Box 3 illustrates the problem. In this example, observers 1 and 2 have each judged that the specified behavior occurred 90 times out of 100 opportunities. They agreed on 81 of the occurrences and 1 of the non-occurrences for a % agreement of 82%. This seems to indicate a reliable measure.

		Observer 1		
		0	1	
Observer 2	0	1	9	10
	1	9	81	90
		10	90	100

% Agreement = 82%

Box 3. Chance and interobserver agreement.

The problem is that this level of

agreement is exactly what is expected by chance given the individual frequencies of occurrence. The proportion agreement about occurrence by chance is 90/100 × 90/100 = .81, which translates into an expected frequency of .81 × 100 = 81, the observed value. The proportion agreement about non-occurrence by chance is 1/100 × 1/100 = .01, which gives an expected frequency of .01 × 100 = 1, the observed value. Note that these levels of agreement occur by chance, meaning that this amount of agreement would occur if the two observers acted completely independently of one another (i.e., zero consistency).

To demonstrate this issue for yourself, toss two coins 100 times and record the outcomes for coin 1 (e.g., a nickel) and coin 2 (e.g., a penny). Record Tail or Head for each coin, perhaps using a table as in Box 3. After 100 trials, count the number of agreements (# HH + # TT) and calculate percent agreement (# agreements/100). What you will observe, with varying degrees of random deviation, is that each coin produced approximately 50 heads and 50 tails, and that there was approximately 50% agreement in the outcomes of the two coins just by chance (25% for heads and 25% for tails). Repeating this exercise with two die and recording whether a 7 occurred or not would produce approximately 17 occurrences (1/6 × 100 = 16.666) and 83 non-occurrences (5/6 × 100 = 83.333) for each die. Chance agreement between the two die would now be 72%, approximately 3 matches for occurrences (1/6 × 1/6 × 100 = 2.777) and 69 matches for non-occurrences (5/6 × 5/6 × 100 = 69.400).

Hopkins and Hermann (1977, p. 124) plot chance agreement as a function of the percent of intervals in which behavior is recorded. Generally, the more extreme the response proportions the greater the level of chance agreement. So be cautious about interpreting high levels of agreement as indicating respectable reliabilities. Other measures must be used, especially when chance agreements are high, and it is often possible to calculate these statistics even when the researchers do not provide them.

Other measures of observer reliability. To compensate for chance, researchers have used various statistics that include an adjustment for chance. There are in fact numerous statistical measures of relations between such categorical variables as occurrence-nonoccurrence. Two commonly used measures are the Phi coefficient and Cohen's Kappa (Hartmann, 1977).

Box 4 illustrates the calculation of the Phi coefficient (Greek φ), a form of correlation coefficient. The numerator of the equation in Box 4 is the difference between the proportion agreements as to the occurrence of the behavior (25/100 = .25) and the proportion agreements expected by chance (30/100 × 60/100 = .18). The denominator is the square root of the product of all the marginal proportions (70/100, 30/100, 40/100, and 60/100).

```
             Obs 1
Obs 2      0     1
      0    35     5     40
      1    35    25     60

           70    30    100

% Agreement = 60%

         (.25-.3×.6)
Phi = --------------- = .3118
      SQRT(.3×.7×.6×.4)
```

Box 4. Calculation of Phi Coefficient.

Although calculated in an unusual manner, the Phi coefficient is equivalent to a standard correlation coefficient between two dichotomous variables. That is, Phi equals the correlation (r) between the two sets of 0s and 1s, where 0 indicates judged non-occurrence and 1 indicates occurrence of the behavior, and each column represents one of the observers. Phi calculated in this manner is shown in Box 5.

```
     Observer    n       CP
    1     2
    0     0     35    -.3×-.6
    0     1     35    -.3×+.4
    1     0      5    +.7×-.6
    1     1     25    +.7×+.4
```

M_{O1}=.3 M_{O2}=.6

SS_{O1}=21 SS_{O2}=24 SCP = 7

r = 7/SQR(21×24) =.3118 = φ

Box 5. Phi Calculated as r.

The value of .31 for the phi coefficient is much lower than the percentage agreement measure, suggesting that much of the latter was due to chance agreement. Calculation of Kappa would show a similar attenuation. The statistical significance of Phi and Kappa can be determined by appropriate inferential statistics, but significant effects are a minimal standard with respect to reliability. Generally, values for Kappa of .4 to .6 might be considered fair reliabilities, values of .60 to .75 moderate, and values over .75 excellent (Fleiss, 1981).

As with other measures of reliability, modern statistical packages provide commands to calculate various measures of relationship for categorical data. In SPSS, the statistical procedure CROSSTABS can be used to perform reliability analyses for categorical data. CROSSTABS can produce Phi, Kappa, and numerous other statistics for agreement between categorical variables. Such methods for determining the reliability of observers and the preceding discussion are also relevant to other studies with dichotomous variables (i.e., variables with only two levels, such as Yes-No, Right-Wrong).

Factors affecting observer reliability. Factors that contribute to unreliable observations include: periodic lapses of attention, temporary blocking of the observer's field of view, occasional inclusive or exclusive coding errors, and lack of agreement on criterion for responses. Many of these and other problems can be handled by appropriate training and administration of observational measures (e.g., avoiding excessively long observation periods, ensuring adequate viewing conditions).

Kazdin (1977) has identified several additional factors that influence the reliability of observations. Reactivity of reliability assessment has been observed in several studies; that is, reliability estimates are higher when observers know that their reliability is being measured than when it is not known. Unobtrusive reliability checks or checks based on randomly selected videotapes of behavior can help to optimize the accuracy of observations. However, these techniques are not always feasible in applied settings or may be prohibitively expensive or time-consuming.

Bias due to observer drift can also influence reliability, either positively or negatively. Drift occurs when behaviors are "redefined" across the time-course of the study. Scoring tapes in random order and bringing in new observers permits the assessment and control of drift.

A third factor affecting reliability is the complexity of the behavior and the observational coding scheme, with more complex systems generally leading to lower reliability. Complexity might also contribute to reactivity effects inasmuch as observers use fewer codes when they think reliability is being checked. One factor that does not appear to affect reliability is observer expectancies about the behavior being observed, unless scores consistent with the expectancies are explicitly reinforced by those monitoring the observers.

Knowledge of observers about the person assessing reliability is also problematic and can inflate reliabilities (Kent et al., 1977). Apparently, idiosyncratic characteristics of the assessor can be developed during the training of observers and these characteristics serve to increase reliability.

Published vs. Study-Specific Reliability

Researchers must examine whether past evidence for reliability is sufficient for their purposes or whether additional evidence should be gathered on reliability, either prior to or as part of the planned study. Careful consideration should be given to the question of whether published reliabilities can confidently be generalized to the present study. In general, well-established standardized tests and other measurement instruments have demonstrated adequate reliability in previous studies and will not require direct new evidence for reliability. Generalizations from previous research might be questionable, however, for studies that involve radically different populations, unusual testing conditions, or other special factors relevant to reliability.

The story is quite different, however, for observational measures. Even when procedures are well-established, observers play such a central role in measurement that reliability should normally be assessed in each study and should not be based solely on prior studies with the same instruments, as is common in the psychometric domain. Observers generally change from study to study and can have a profound influence on the nature of the measures obtained, including their reliability.

Reliability is a minimal condition for measuring human behavior. Unless there is reliable measurement, researchers have little evidence that they are obtaining evidence with respect to the constructs of interest. But reliability alone is not sufficient to conclude that the measures being used are of high quality. High degrees of reliability are of little benefit if the measures are assessing some construct other than that of interest. Measures must be valid as well as reliable.

VALIDITY

Validity refers to the extent to which measurements truly reflect the underlying construct of interest; that is, does the test or measurement procedure actually measure what it was designed to measure? Traditionally validity has been assessed in several different ways, although there is increasing appreciation that alternative methods are intimately related to one another.

Kinds of Validity

Measures of validity can be roughly classified into three types: content validity, criterion validity, and construct validity. All measures of validity involve agreement between alternative judgments about the construct. Agreement is often reflected as a correlation coefficient; when it is, correlations of .50 are thought to reflect adequate levels of validity.

Content validity. Content or face validity refers to the apparent relevance of the material included in the test. A measure of impulsivity, for example, could be validated by seeing if the items directly implicate impulsive behavior (e.g., I do things without thinking.). Sometimes judgments about content validity are relatively straightforward (e.g., a spelling item would be irrelevant on a test of mathematical knowledge), but at other times judgments are complicated by uncertainty about the nature of the underlying construct or constructs being measured. Content validity is important in the early stages of developing a test when researchers are trying to generate items that they think will reflect the construct of interest.

Content validity has also been an important consideration in the issue of cultural biases in testing. Intelligence tests that ask about "busses" and "subways," for example, are inappropriate for use with children in remote regions (e.g., Inuit children), or at least they were inappropriate prior to the availability of television. Along similar lines, several items on the widely-used Weschler tests of intelligence have been modified for use with Canadian subjects.

Criterion validity. Criterion validity is determined by correlating scores on the new measure with scores for the same people from already-accepted measures of the construct (i.e., measures that are accepted as valid). Criterion validity for a new test of impulsivity could be determined by correlating impulsivity scores with self-ratings, ratings by peers or others familiar with the subjects, or performance on the Matching Familiar Figures Task, one widely-used measure of impulsivity.

When performance on the criterion task is obtained much later than the target measure, the term predictive validity is sometimes used. An example of predictive criterion validity would be correlating aptitude test scores obtained at the start of grade six with final grades in grade six. When performance on the criterion task and the new measure are obtained at the same time, then the term concurrent validity can be used. Correlating scores on two tests of anxiety that were administered at the same time illustrates concurrent criterion validity.

Construct validity. Construct validity is a sophisticated view of validity that subsumes other types of validity (Messick, 1981).   To determine construct validity, researchers examine whether test scores can be linked to measures of other constructs in ways that are consistent with the theoretical network on which the target construct is based. Scores on a measure of impulsivity, for example, should correlate with other variables (e.g., frontal lobe injuries) in accordance with the theory of impulsivity used to develop the measure. These theoretical relations include differences in anxiety between pre-existing groups as well as variation in anxiety as a function of experience or controlled situations (i.e., experiments).

Construct validity is confirmed when a measure correlates as expected with theoretically related measures (i.e., convergent validity) and when it fails to correlate with theoretically unrelated measures (i.e., discriminant validity). Convergent validity is related to the notion of internal consistency, except that different measures are correlated, rather than items within a measure. To illustrate discriminant validity, a cognitive ability measure, such as multiplication skill, should not correlate highly with measures of unrelated, emotional traits (e.g., depression).

Contemporary views of validity recognize that construct validity subsumes other types of validity and some aspects of reliability. Construct validity implicates internal consistency measures of reliability because inter-item correlations and related statistics (e.g., split-half correlations or alpha) provide evidence about the conceptual homogeneity of the items (i.e., convergent validity). High intercorrelations among items indicate a relatively pure measure of some construct.

The current emphasis on construct validity also puts theory development at the heart of measurement (as do other developments in testing that are discussed later). Researchers must develop their constructs and theories at least to the point that meaningful evaluation of measures can be made. Unless a well-defined construct is situated firmly in a theoretical network of other constructs, it is difficult to evaluate construct validity. That is, measures of a construct X cannot be validated without knowing how X is related theoretically to Y, Z, and a host of other variables. One useful tool in construct validation is factor analysis. This statistical method identifies the number of “factors” (i.e., constructs) that underlie correlations among some set of measurements (e.g., individual responses to a questionnaire). In the area of intelligence testing, for example, some analyses suggest two overlapping factors (a verbal factor and a performance or nonverbal factor). For a brief introduction to Factor Analysis, see the Appendix.

Threats to Validity

Numerous factors can compromise the validity of psychological measures. In essence, validity is reduced by any variable that is irrelevant to the construct supposedly being measured and yet affects the scores of subjects. For further discussion of the following factors, as well as other issues related to validity, see Cook and Campbell (1979).

Inadequate preoperational explication of constructs. One of the most important and all-too-often neglected aspects of validity is the initial articulation of the constructs to be studied. Pay careful attention to what constructs you want to measure, defining each as fully as possible. Development or selection of the actual measures should be done in a way that will maximize the fit between the theoretical constructs being examined and the actual measures used in the research.

Limited implementation of constructs. Researchers too often limit themselves to a single measure of the target constructs. Single measures make it impossible to determine the validity of the measures in this particular study and leave the variables unnecessarily exposed to contaminating variables. Multiple operationism means that researchers identify and implement multiple measures for their constructs, enabling them both to minimize threats to validity that can affect single measures, and to actually validate their measures in the current study.

Response biases. Subjects often respond on the basis of general biases or response tendencies, rather than the construct of interest. If all the items on a test require a "Yes" response, for example, then people who tend to agree more than others will obtain somewhat higher scores than people who tend to agree less, irrespective of the trait actually being measured. This response bias can be minimized by having high scores result from agreement for half the items and from disagreement for the other half.

Other responses biases may be less easy to eliminate or reduce. Social desirability refers to the tendency to endorse items in a socially desirable way, perhaps to present a positive image of yourself (rather than the true image of interest to the researcher). One strategy test-developers use to deal with social desirability is to measure both the trait of interest and social desirability, and eliminate items that correlate more highly with desirability than with the target construct.

Reactivity. Reactivity refers to the effect of measurement on the trait being measured and was mentioned earlier in the context of observational measures. That is, subjects might behave differently because they are being observed or measured than when they are not being assessed. Reactivity depends on the conspicuousness of the observer, personal characteristics of subjects and observers, and the rationale provided for the observing. Reactivity is particularly problematic when the behavior being measured is highly sensitive (e.g., parent-child interactions, value-laden beliefs). When an observer is used, a lengthy familiarization period can reduce the effect of observation on the behavior.

One solution to the problem of reactivity is to use nonreactive or unobtrusive measures, which are measures that cannot be influenced by the measuring procedure. Archival or historical records constitute one illustration of nonreactive measures. School records, for example, were created in the past and will not change when the researchers score them for some measure (e.g., references to aggressive behavior). Webb, Campbell, Schwartz, Sechrest, and Grove (1981) describe many examples of nonreactive measures, including a number of very creative approaches to the measurement of behavior. Some potential unobtrusive measures, however, my be inappropriate for ethical reasons.

Misremembering. Memory distortions can invalidate measures obtained by retrospective reports. Robbins (1963) examined the accuracy of parental reports of the nature and timing of various landmark events during infancy (e.g., age child walked alone, whether the infant sucked thumb). Accuracy could be determined because the parents who provided the reports had participated in a longitudinal diary study during the period of the events queried for the retrospective reports.

Parents were questioned when their children were approximately three years old. Mothers and fathers were questioned separately at the same time. Interviewers were counterbalanced across parents; that is, each of the two interviewers interviewed half mothers and half fathers.

Retrospective reports were compared to the earlier diary records. Accurate reports were those that agreed for such qualitative dimensions as sucking thumb and those that were within one month, half a pound, or half an inch for quantitative dimensions. Most of the quantitative dimensions were time-based; hence accuracy was generally a retrospective report that was within one month of the the diary data.

Figure 1 shows the percentage of mothers and fathers who were accurate on the 5 qualitative dimensions and the 13 quantitative dimensions according to these criteria.

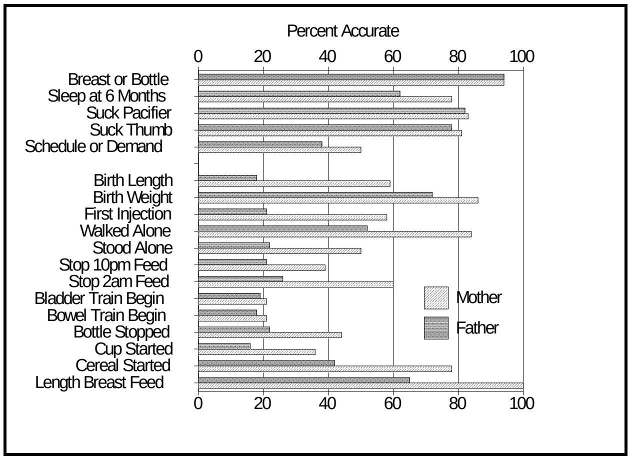

Figure 1. Accuracy of Parent Reports (Robbins, 1963)

Although there were a few areas in which high percentages of parents were accurate (e.g., whether their babies were breast or bottle fed, length of breast feeding), many more of the questions were answered inaccurately on the retrospective reports. Only 20% of the parents remembered to within one month the time that bowel training began. Freudian psychologists should probably be cautious about testing their theories against parental reports, at least if accuracy of reporting is an important consideration.

These findings and related studies indicate that retrospective reports must be examined very carefully for accuracy. It is imprudent to assume that even highly educated and motivated individuals can accurately remember information from even a few years back. Similar precautions hold when asking individuals to recall events in their own lives. Various cognitive and emotional factors can produce serious distortions in the data. However, there may also be ways of reducing such distortions. Loftus, Klinger, Smith, and Fiedler (1990), for example, found that repeating questions with different reference periods improved memory for medical procedures, which could be verified against medical records.

Future Developments in Validity

This chapter has described some basic and traditional aspects of measurement, primarily drawn from the areas of assessment and statistics. Increasingly, however, measurement is being studied and modelled by researchers in various areas who are interested in the psychological mechanisms that underlie measurement. These researchers bring the theories and processes of cognitive psychology, social psychology, and other disciplines to bear on the question of how subjects behave in testing. These perspectives emphasize the development of theoretical models for the performance of different tasks and add a new dimension to traditional views of validity.

Cognitive models. Cognitive psychologists typically study the mental processes that underlie performance on memory, linguistic, and other tasks. But these are the same processes that subjects use to perform any psychological task, including completing a questionnaire, making judgments about others or themselves, and other measurement procedures. It has been proposed that cognitive analysis can shed light on many aspects of test performance, perhaps especially on intellectual tests (e.g., Sternberg, 1981). The basic idea is that understanding the mechanisms that underlie task performance will shed light on the psychological differences between high and low scorers. To illustrate, subjects might do poorly on a picture-naming test, not because of vocabulary or other linguistic difficulties, but because they are unable to suppress interfering responses to test items. Similar factors may determine the difficulty of items on picture-naming tests (Johnson & Clark, 1988).

Elizabeth Loftus and her colleagues (Loftus, Fienberg, & Tanur, 1985) have also demonstrated how cognitive psychology can improve survey measurement methods. For example, using concrete landmarks (e.g., eruption of Mt. Helen's, New Year's) reduces the incidence of reporting events as occurring more recently than they did (i.e., "forward telescoping;" Loftus & Marburger, 1983). Much of the laboratory and natural work that Loftus has done on the permanence and malleability of human memory is relevant to many psychological measurement procedures.

Context Recreation: Recall enhanced by recreating event stimuli (physical and psychological). Ask interviewee to think back to original event, recalling physical (time of day, workspace, etc.) as well as emotional (rushed, bored, etc.) surroundings.
Focused Concentration: Distractions deteriorate memory retrieval process. Interviewee may close eyes to minimize distractions; interviewer avoids interrupting or other intrusions to sessions.
Extensive Retrieval: Recall increased by increasing number of retrieval attempts. Interviewer does not let interviewee stop after cursory search of memory, but encourages multiple attempts.
Varied Retrieval: Recall may be activated by different probes. Events are commonly recalled in chronological order, from an egocentric perspective. Ask for recall of details in reverse order, or starting from middle and working to either end. Asking interviewee to recall event from perspective of third party witnessing event may elicit additional details previously unrecalled.
Multiple Representations: Events may be stored and recalled in multiple forms. Have interviewee recall details considered unusual, humorous, etc. (i.e., those that share a theme. Ask interviewee to use multiple senses (sounds, tactile representations, etc.) when attempting to recall.

Box 6. Cognitive Interview Techniques.

Other cognitive work along these lines has resulted in the development of guidelines for eye-witness, clinical, and research interviews. One such method is the Cognitive Interview (Moody et al, 1998), which addresses interview implications of five aspects of memory performance (see Box 6). Note that some of these techniques remain controversial (i.e., may implant false memories).

Social information processing. A number of other factors that influence the results of surveys have been conceptualized as involving social information processing. Schwarz (1994) provides several illustrations of how social and cognitive principles can help us to understand the results of surveys and other psychological measures.

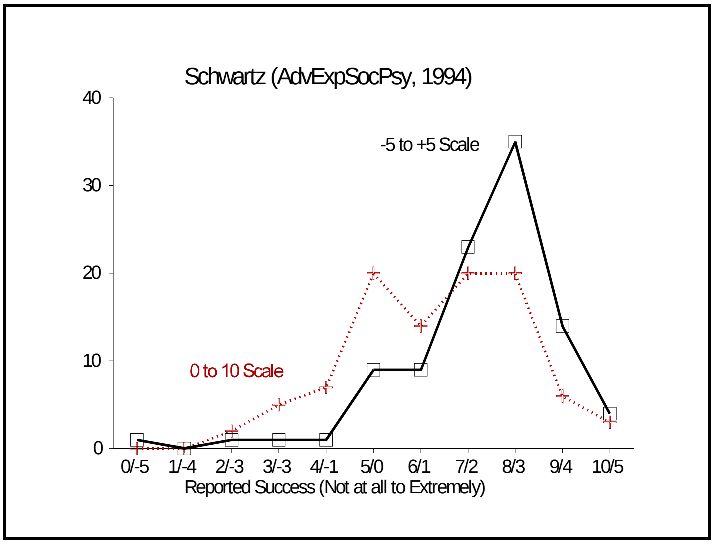

Figure 2. Effect of Bipolar versus Unipolar Response Scales.

The example in Figure 2 show that subjects are more likely to use the full range of an 11-point scale if the scale goes from 0 to 10 than if the scale goes from -5 to +5, even though the number of alternatives available are exactly the same. Subjects avoid the negative numbers, perhaps because they imply a negative conceptualization of the underlying continuum. That is, 0 to 10 might suggest less to more happy, without implicating sadness, whereas -5 to +5 may prime subjects to interpret the choice as being between sadness and happiness. These are clearly different psychological choices.

A second example from Schwarz concerns the frequency range of the alternatives offered. Subjects given numerous low frequency alternatives (up to ½, ½ to 1, 1 to 1½, 1½ to 2, 2 to 2½, more than 2½) reported fewer hours of watching television than subjects given high frequency alternatives (up to 2½, 2½ to 3, 3 to 3½, 3½ to 4, 4 to 4½, More than 4½). As shown in Figure 3, the proportion of people reporting that they watch more than 2½ hours of television per day differs markedly for the two ranges of alternatives. Schwarz suggested that subjects obtain social comparison information from their location along the scale and use this information in making judgments.

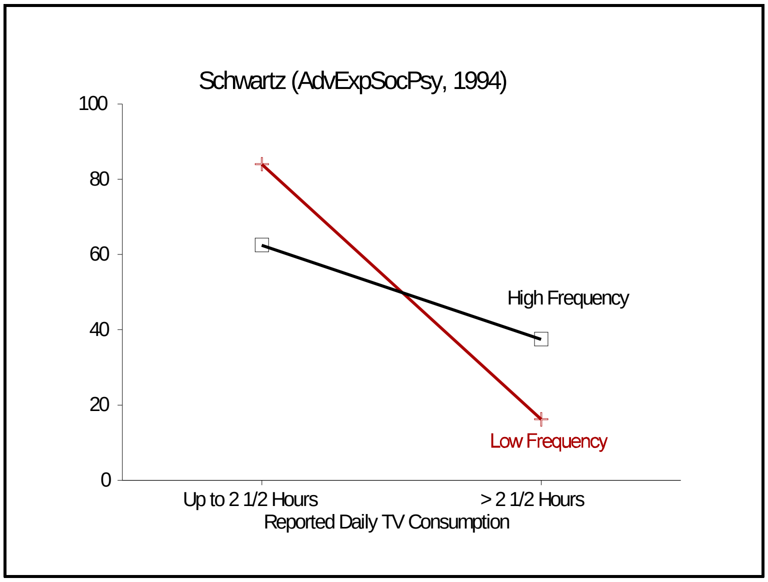

Figure 3. Effect of Range of Alternatives on Responding.

DEVELOPING PSYCHOLOGICAL MEASURES

Researchers generally have two options when selecting measures for a study. They can either find an existing measure that suits their purposes or develop a new measure that demonstrates adequate reliability and validity, as described in preceding sections of this chapter.

Despite the multitude of measures that are available, many researchers conclude that they must develop their own measure because the psychometric properties of available instruments are inadequate, the test is not available, or the the construct has not been operationalized in a way consistent with the to-be-tested theory. Although time-consuming, developing new psychological measures can increase the fit between the theoretical constructs and the measures.

Several psychologists have described models for test development that incorporate many of the psychometric properties discussed in this chapter. Following these models will minimize problems with the resulting measures.

Douglas Jackson, who has developed many personality, vocational, and ability scales, has described a procedure that he calls rational test construction. Some core features of the procedure are: conceptualization of the construct (i.e., definition), generating potential items, administering the test to subjects, evaluating the quality of the current version of the test (internal consistency, discriminant validity, bias), revising the test (keeping good items and revising or replacing poor items), and repeating the administration - evaluation - revision steps until the test is of adequate quality. Broughton (1984) describes a prototype strategy for the construction of personality scales.

Even if one decides to develop a new measure, it would be important to review existing measures during the development process. An effective search might even uncover a suitable test, which could save valuable resources.

FINDING EXISTING PSYCHOLOGICAL MEASURES

Finding an appropriate measure involves general literature search techniques similar to those reviewed in earlier Chapters. In addition, there are many specific sources of information about existing measures in psychology, both in print and in computer databases.

Literature Related to Testing

The importance of tests and measures in psychological research has led to the development of a number of valuable resources.

Test reviews. Excellent sources for measurement procedures in many areas of investigation are the Buros Mental Measurement Yearbooks and Tests in Print (Buros, various years; Mitchell, 1983). These resources describe tests, provide critical reviews, and list research studies that have used the instruments.

Similar resources to Buros include: Chun, Cobb, and French's (1975) description of 3,000 original psychological measures and their applications; Grommon, Braddock, and others (1976) review of selected published tests in English; Hammill, Brown, and Bryant's (1992) consumer guide to tests in print; and Sweetland and Keyser's (1991) comprehensive reference for assessments in psychology, education, and business. Hepner (1988) provides a cumulative index to psychological tests in microfiche form.

Miller's (1991) handbook of research design and social measurement provides substantial information about various aspects of measurement. Part 6 summarizes briefly selected sociometric scales, including a number relevant to psychology (e.g., personality tests). Section 6.L.5 describes several compilations of scale sources such as Lake et al. and the Buros volumes.

Testing books. General books on testing provide additional sources of information about psychological measures (e.g., Anastasi, 1988). In addition to discussing reliability, validity, and other measurement topics, such books often include overviews of illustrative tests and measures. Other books cover assessment in particular domains, such as neuropsychology (Berg, Franzen, & Wedding, 1994) or education (Salvia & Ysseldyke, 1988).

Testing journals. A number of psychological and educational journals specifically address assessment issues, including those listed in Box 6. In addition, specialized journals in particular areas will often include relevant measurement instruments. The journal Intelligence and such neuropsychological journals as Brain and Cognition and Neuropsychology, for example, include numerous articles that deal with assessment in the respective domains.

Applied Psychological Measurement
Behavior Research Methods, Instrumentation, and Computers
Behavioral Assessment
British Journal of Mathematics and Statistical Psychology
Educational Measurement: Issues and Practice
Educational and Psychological Measurement
Journal of Applied Psychology
Journal of Behavioral Assessment
Journal of Educational Measurement
Journal of Personality Assessment
Journal of Psychoeducational Assessment
Measurement and Evaluation in Guidance
Multivariate Behavioral Research
Psychological Assessment
Psychological Bulletin
Psychometrika

Box 7. Assessment-Related Journals.

Perhaps the most general of these journals is Behavior Research Methods, Instruments, and Computers. This journal publishes articles specifically on research methods and the articles often concern measurement issues.

General Guides to searching for tests. Reed and Baxter (1983, 1992) devote several sections of their handbook for library use in psychology to finding out information about tests and other assessment procedures. Many university libraries also have prepared handouts on the topic of psychological assessment.

As well as the general resources discussed here, researchers gradually acquire a repertoire of standard measurement procedures used in their area. Such knowledge is part of becoming an expert in the area. Novices would do well to focus on the identification of the central constructs in their new area and the primary methods used to measure those constructs.

Computer Resources Related to Testing

In addition to examining relevant books, researchers increasingly use computers to search for relevant measures. The following brief description of these researchers may help students to get started with some of these resources. A number of the following links can be found at www.uwinnipeg.ca/~clark by following the Research Tools link.

APA Website. The website of the American Psychological Association (APA) is a good place to start. A section devoted to Tests and Measures can be found at the following location: www.apa.org/science/testing.html. This site includes introductions to many of the following resources, as well as many others, and provides much useful material on ethical codes for use of tests. To find information on particular tests, check out the following link, which appears on APA's main testing page: FAQ: Finding Information About Psychological Tests.

Some Other Useful Sites (Perhaps!). One important source in testing historically has been the Buros Institute of Mental Measures, at http://www.unl.edu/buros/. This site provided an index to reviews of different tests and other useful information. The Institute publishes the Mental Measurements Yearbook, in its 16th edition as of 2005. ERIC is the primary bibliographic source in education, and includes (or did include) much information on psychological tests. ERIC can be accessed at: http://ericae.net/. To find material specific to testing, choose Advance Search and limit to Test Questionnaires. Because of recent changes in these sites, some links and information may no longer be available, or may require some exploration before you find the relevant sites. For links to these and other sites, see www.uwinnipeg.ca/~clark/research/meas.html.

PsycINFO. PsycINFO is the main bibliographic resource in psychology. There are several helpful aspects of PsycINFO that facilitate searching for tests and measures in psychology. One of the Indexes available in PsycINFO is a Tests and Measures index. Select this Index and then Browse for tests relevant to your research question. Select the tests of interest, add them to the Search field with an "or" operator, and then search for references in PsycINFO.

A second approach is to use the Classification index, Browse for Tests/Testing, and then add the "Tests & Testing" classification code to the Search field. "And" this to the most appropriate content words for your search from the Thesaurus. The relevant Classification Code can be used to limit the results of your search to articles related to measurement.

Searches limited by some classification (e.g., Tests & Testing) can also be conducted using the Advanced Search option. In one field, limit the search to the relevant domain and in another field specify concepts to be searched for.

Search engines. Finally, do not overlook the use of the powerful search engines that are now available on the internet, including Google at www.google.ca. Entering phrases like "psychological tests" will identify numerous general sites on testing. Entering such phrases along with content descriptors may provide some access to relevant measures.

Getting the Test

Once one or more measures have been selected, researchers must locate an actual copy of the test. Some psychological measures are published in journal articles and can be used freely for research purposes. In the case of commercial tests, however, researchers must contact the test publisher or author. Some tests require specific training before they can be used and will not be sold except to qualified people (Standards, 1985). Another source of tests may be the test library in your Psychology department. Ask for a list of available tests and the procedures for using specific measures.

CONCLUSIONS

Measurement is fundamental to good research and, as illustrated here, more complex than it might appear. Weak, irrelevant, or ambiguous measures compromise any findings based on those measures. Although some useful guidelines have been presented, researchers should not blindly follow some "cookbook" view of measurement. Poor measures of the right construct are far more desirable than technically good measures of the wrong construct.

References

Allport, G. (1942). *The use of personal documents in psychological science*. New York: Social Science Research Council.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: MacMillan.

Berg, R. A., Franzen, M., & Wedding, D. (1994). *Screening for brain impairment: A manual for mental health practice* (2nd ed.). New York: Springer.

Broughton, R. (1984). A prototype strategy for construction of personality scales. *Journal of Personality and Social Psychology, 47*, 1334-1346.

Buros, O. K. (Ed.) (1978). *Eighth mental measurements yearbook*. Highland Park, NJ: Gryphon Press. \UOW REF Z 5814 P8B932 v.3-11, v.9-10 SUPPL. SUPPL.

Buros, O. K. (Ed.) (1970-75). *Personality tests and reviews; including an index to The mental measurements yearbooks*. Highland Park, NJ: Gryphon Press. \UOW REF Z 5814 P8B95

Buros, O. K. (Ed.) (1961, 1974). *Tests in print* (Vol. 1 and 2). Highland Park, NJ: Gryphon Press. \UOW REF Z 5814 E9B8 v.1-3

Chun, K., Cobb, S., French, J. R. P. (1975). *Measures for psychological assessment: A guide to 3,000 original sources and their applications*. Ann Arbor, MI: Survey Research Center, Institute for Social Research. \UOW REF BF 698.5 C45

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally.

Ericsson, K. A., & Simon, H. A. (19840. *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press. \UOW BF 455 E68

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. H. Freeman.

Hammill, D. D., Brown, L., & Bryant, B. R. (1992). *A consumer's guide to tests in print* (2nd ed.). Austin, TX: Pro-Ed. \UofManitoba EDUC LB 3060.32 N67 H36 1992

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10*, 103-116.

Hepner, J. C. (1988?). *Cumulative index to tests in microfiche*. Princeton, NJ: Educational

Testing Service. \UOW MICROFICHE LB 3051 T45 Index 1975\87

Hersen, M. (Ed. In Chief) (2004) *Comprehensive handbook of psychological assessment* (Vol. 1 to 4). Hoboken, NJ: J. Wiley. \BF 176 C654 2004

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Hopkins, B. L., & Hermann, J. A. (1977). Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis*, *10*, 121-126.

Johnson, C. J., & Clark, J. M. (1988). Children's picture naming difficulty and errors: Effects of age of acquisition, uncertainty, and name generality. *Applied Psycholinguistics, 9*, 351-365.

Katz, A. N. (1979). Demonstrating semantic memory in the teaching laboratory with a paper-and-pencil task. *Teaching of Psychology*, *6*, 173-175.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 1,* 141-150.

Kent, R. N., O'Leary, K. D., Dietz, A., & Diament, C. (1979). Comparison of observational recordings via in vivo, mirror, and via television. *Journal of Applied Behavior Analysis, 12*, 517-522.

Loftus, E. F., Fienberg, S. E., Tanur, J. M. (1985). Cognitive psychology meets the national survey. *American Psychologist*, *40*, 175-180.

Loftus, E. F., Klinger, M. R., Smith, K. D., Fiedler, J. (1990). A tale of two questions: Benefits of asking more than one question. *Public Opinion Quarterly*, *54*, 330-345.

Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*, *11*, 114-120.

Maddox, T. (Ed.) (2003). Tests: A comprehensive reference for assessments in psychology, education, and business (5th ed.). / edited by Taddy Maddox. Austin, TX: Pro-Ed. \Z5814P8T43 2003

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Review, 89*, 575-588.

Miller, D. C. (1991). *Handbook of research design and social measurement* (5th ed.). London: Sage.

Mitchell, J. V. (Ed.) (1983). *Tests in print* (Vol. 3). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska.

Reed, J. G., & Baxter, P. M. (1983). *Library use: A handbook for psychology*. Washington, DC: American Psychological Association. \UofW BF 76.8 R43 1992 2nd ed.

Salvia, J., & Ysseldyke, J. E. (1988). *Assessment in special and remedial education* (4th ed.). Boston, MS: Houghton Mifflin.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology, 26*, 123-162.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychologist, 36*, 1181-1189.

Sweetland, R. C., & Keyser, D. J. (Eds.) (1991). *Tests: A comprehensive reference for assessments in psychology, education, and business* (3rd ed.). Austin, TX: Pro-Ed. \UOW REF Z 5814 P8T43 1991

Webb, E. T., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston: Houghton Mifflin.

APPENDIX: MEASUREMENT AND FACTOR ANALYSIS

As psychological theories and practices become increasingly sophisticated, researchers are faced with studies that involve multiple variables, some of which may measure related constructs (e.g., different measures of prejudice). It can be useful to reduce the multiple variables to fewer factors. For example, several measures of ability might be so highly correlated that they can be combined and treated as a single factor instead of being analyzed separately. The combined factor scores can be used as independent or dependent variables in subsequent analyses. One technique that achieves this kind of reduction is factor analysis. An SPSS program was used to generate 9 scores for each of 50 cases, such that variables x1 to x3 were correlated with one another; variables y1 to y3 were correlated with one another; and variables z1 to z3 were correlated with one another. Correlations between the three clusters were correlated only at chance levels.

The correlations among the 9 variables are shown in Box 7. The correlations between variables within each cluster tend to be positive and somewhat higher than correlations between clusters.

Even when we know what pattern to look for, however, finding clusters of related variables can be difficult. The SPSS Factor command in Box 7, which produced the correlation matrix, requests SPSS to perform a factor analysis on the variables. This is the simplest form of the FACTOR command and lets SPSS use the default values for the various steps involved.

```
FACTOR VAR = x1 TO z3  /PRINT = DEFAULT CORR
```

	X1	X2	X3	Y1	Y2	Y3	Z1	Z2
X2	**.22**							
X3	**.23**	**.29**						
Y1	-.24	.02	-.05					
Y2	.07	.11	-.17	**.33**				
Y3	-.01	.13	.07	**.22**	**.47**			
Z1	-.18	-.11	.06	-.20	-.09	-.11		
Z2	-.14	-.22	-.08	-.14	-.10	-.16	**.39**	

Box 8. Correlation Matrix.

VARIMAX ROTATION 1 FOR EXTRACTION 1 IN ANALYSIS 1 - KAISER NORMALIZATION. VARIMAX CONVERGED IN 5 ITERATIONS.

ROTATED FACTOR MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
X1	-.14227	-.11192	**.64904**
X2	-.08743	.22237	**.68713**
X3	.00129	-.08630	**.71505**
Y1	-.17149	**.60917**	-.29790
Y2	-.01607	**.83195**	.00616
Y3	-.09848	**.75489**	.16232
Z1	**.79582**	-.06966	.03707
Z2	**.73984**	-.12496	-.17264
Z3	**.70554**	-.05321	-.08894

Box 9. Rotated Factor Matrix.

Box 8 shows the final result of the factor analysis. SPSS Factor decided on statistical grounds that three factors were needed to "explain" the correlation matrix in Box 7. The numbers in Box 8 are called factor loadings; factor loadings are essentially correlations between each measure and imaginary new scores called factors. That is, the loading of .24082 for x1 on factor 1 is a correlation between x1 scores and factor 1 scores (like a sub-scale score).

The factor loadings in Box 8 reveal a systematic pattern, and indeed correspond very closely to the pattern we would expect given what we know about how the data were generated. Factor 1 has high loadings on the z variables, factor 2 has high loadings on the y variable, and factor 3 has high loadings on the x variable. Factor analysis was able to "find" this pattern of loadings from the correlation matrix among the 9 scores and could do the same for any number of variables.

The interpretation of a factor analysis involves much conceptual work, both before collecting the data (e.g., deciding what variables to measure and in what ways) and during and following the analysis. Researchers can examine the factor loadings and attempt to label the factors in a meaningful way based on their knowledge of the individual scales making up the factor. In our example, Factor 1 might be called Factor Z, Factor 2 might be called Factor Y, and Factor 3 might be called Factor X.