Handbook of INDUSTRIAL, WORK AND ORGANIZATIONAL PSYCHOLOGY

Volume 1 Personnel Psychology

Edited by

NEIL ANDERSON
DENIZ S. ONES
HANDAN KEPIR SINANGIL
CHOCKALINGAM VISWESVARAN

2001

SAGE Publications
London • Thousand Oaks • New Delhi

2

Measurement in Work and Organizational Psychology

HERMAN AGUINIS,
CHRISTINE A. HENLE,
and CHERI OSTROFF

The goals of this chapter are to (a) provide an overview of measurement and the process of measure development, and (b) describe recent and future trends in the field of measurement in work and organizational psychology. First, we define measurement, discuss some of its benefits, and describe scales of measurement. Second, we describe the process of measure development. This section includes topics such as defining the purpose of measurement and the attribute(s) to be measured, the development of a measurement plan, creating items, conducting a pilot study and item analysis, selecting items, establishing norms, and assessing the reliability and validity of a measure. Finally, we address a selective set of recent and future trends in measurement including issues pertaining to levels of analysis, the impact of technology on measurement, cross-cultural measurement transferability, emerging legal and social issues in measurement, and the globalization of measurement.

Measurement is pervasive in our everyday lives. As we go through our daily activities we glance at our watches to check the time, step on the scale to assess our weight, and look at the speedometer to see how fast we are driving. In addition, schools grade our knowledge, employers test our intelligence and personality, and medical doctors evaluate our health. In sum, we are continually measuring and being measured by others. Not only does measurement influence our daily lives, but also the science and practice of work and organizational (W&O) psychology rely on good measurement. Without good measurement as a foundation, our field could not advance or provide a valuable service to the business community.

As W&O psychologists, we continuously make many decisions that rely on accurate measurement. In practice, we use our knowledge to make decisions, for example, about employee selection, classification, placement, and guidance. These decisions rely on solid measurement of employee attributes, skills, interests, and values. If we do not have reliable and valid measures of employee characteristics, the decisions we make are not justified and numerous lives may be affected negatively. Thus, as practitioners, we have a responsibility to our clients to ensure that we base our recommendations and decisions on sound measurement.

The decisions we make in practice also rely on research we have conducted in both laboratory and field settings. The accuracy of this research relies on sound measurement of the variables examined. Measurement is essential to our research because it allows us to describe, predict, explain, diagnose, and make decisions about the issues under investigation. If our research lacks good measurement, results will be meaningless and unable to inform the practice of W&O psychology.

Measurement is the cornerstone of both the science and practice of W&O psychology. Without solid measurement, our research is misleading and our practice is haphazard. We must focus on

measurement because it can provide accurate and relevant information that leads to informed decision-making.

This chapter is organized as follows. First, we define measurement, discuss some of its benefits, and describe scales of measurement. Then, we describe the process of measure development. This section includes topics such as defining the purpose of measurement and the attribute(s) to be measured, the development of a measurement plan, creating items, conducting a pilot study and item analysis, selecting items, establishing norms, and assessing the reliability and validity of a measure. The last section of the chapter addresses a selective set of recent and future trends in measurement including issues pertaining to levels of analysis, the impact of technology on measurement, cross-cultural measurement transferability, emerging legal and social issues in measurement, and the globalization of measurement.

DEFINITION OF MEASUREMENT

Measurement is the assignment of numbers to attributes or properties of people, objects, or events based on a set of rules (Stevens, 1968). From this definition we can derive several characteristics of measurement. First, measurement focuses on *attributes* of people, objects, or events not actual people, objects, or events. Second, measurement uses a set of rules to quantify these attributes. Rules must be standardized, clear, understandable, and easy and practical to apply. Third, measurement consists of two components, scaling and classification. Scaling is the assignment of numbers to attributes of people, objects, or events in order to quantify them (i.e., determine how much of a particular attribute is present). Classification refers to defining whether people, objects, or events fall into the same or different categories based on a given attribute.

The above definition alludes to a process of measurement. First, we need to determine the purpose of measurement (e.g., prediction, classification, decision-making). Second, we must identify and define the attribute we intend to measure. A definition must be agreed upon before the attribute is measured or different rules may be applied, resulting in varying numbers assigned to the attribute. The purpose of the measurement should guide the definition. Next, we determine a set of rules, based on the definition, to quantify the attribute. Finally, we apply the rules in order to translate the attribute into numerical terms.

BENEFITS OF MEASUREMENT

We asserted above that the science and practice of W&O psychology cannot exist without sound measurement. Science cannot progress any faster than the measurement of important variables in the field. By following the process of measurement outlined above, we can develop good measures which, in turn, reap several benefits (Nunnally, 1978). First, measurement contributes to objectivity. It minimizes subjective judgment from scientific observation and allows theories to be tested because attributes being examined can be adequately assessed and measured (Aguinis, 1993). Second, measurement leads to quantification. By quantifying the attributes we are exploring, more detail can be gathered than with personal observations and judgments. In addition, more subtle effects can be observed and more powerful methods of statistical analysis can be used, which enables us to make precise statements about the patterns of attributes and their relationships among each other (Pedhazur & Pedhazur Schmelkin, 1991). Third, standardized measures result in better communication because they create a common language and understanding of attributes, thus research can be compared. Fourth, sound measures save time and money by allowing researchers and practitioners to focus their energy elsewhere because less-trained individuals can administer and score standardized measures.

Arguably, the most important benefit of measurement is better decision-making about individuals and groups. Measurement provides relevant and accurate information that decision-makers can use to make sound and informed decisions. Thus, measurement provides an important set of tools for improving the information available to decision-makers regarding employee selection, placement, classification, guidance, training and development, compensation, and so forth.

SCALES OF MEASUREMENT

As mentioned earlier, measurement uses a set of rules to quantify attributes of people, objects, or events. The type of measurement scale places a limit on the statistical analyses that can be applied to the quantification of attributes. Stevens (1951) proposed four types of measurement scales: Nominal, ordinal, interval, and ratio. As the measurement of attributes progresses from nominal to ratio, more sophisticated quantitative analyses can be implemented.

Nominal Scales

A nominal scale is the most basic and it involves assigning numbers as labels to individual objects (e.g., telephone numbers) or categories of objects (e.g., sex, organizational unit). Nominal scales determine whether objects belong in the same or different categories (e.g., male or female) based on a given attribute (e.g., sex). Thus, nominal scales classify people or objects.

Data collected using nominal scales have a limited number of transformations and statistics available. First, each category may be assigned any number as long as it is different from other category numbers. For example, men may be labeled 1 and women 2 or men could be labeled 123 and women 654. The categories are not ordered; one is not more than the other, but they are different from each other. Second, the amount of difference between categories is unknown and the only permissible statistics for nominal scales are those based on counting the number of subjects in each category (i.e., frequencies) and proportions.

Ordinal Scales

Ordinal scales involve assigning numbers to people or objects so that their rank order can be determined. That is, ordinal scales help decide if one person is equal to, greater than, or less than another based on a given attribute. For example, a supervisor believes María is a better performer than Bob, thus María is given a 2 while Bob is assigned a 1 to show María has a higher performance ranking than Bob. However, this does not indicate the magnitude of the difference between María's and Bob's performance levels, we just know that María is better than Bob.

Monotonic transformations are permissible for ordinal scales. This means that the transformation must maintain the rank order of individuals or categories. Categories labeled 1, 2, and 3 can be transformed to any numbers as long as their order is preserved (e.g., 4, 5, 6 or 10, 20, 30 is permissible, but 6, 5, 4 is not). Permissible statistics for data collected using ordinal scales include the median and the mode; the mean cannot be calculated because a different mean will be obtained whenever the categories are recoded while the median and mode categories will stay the same. Percentile ranks, correlation coefficients based on ranks (e.g., Spearman's rho and Kendall's W), and rank-order analysis of variance can also be used.

Interval Scales

Interval, like ordinal scales, assign numbers to reflect whether individuals or objects are greater than, less than, or equal to each other. However, interval scales also indicate the difference between objects on a particular attribute. A common example of an interval scale is Celsius temperature. If one city has a temperature of 20° and another has a temperature of 40°, we not only know that the second city has a warmer temperature than the first, but that it is 20° warmer than the first. Thus, interval scales use constant units of measurement so that differences between objects on an attribute can be expressed and compared. However, the absolute magnitude of the attribute is not known because the zero point on an interval scale is arbitrarily determined (e.g., zero point on Celsius scale is set arbitrarily at the freezing point of water). Most measures used in W&O psychology include interval scales.

Linear transformations (e.g., $X' = a + b\ X$) are permissible with interval scales where X' is the transformed score, X is the score to be transformed, and a and b are constants. For example, Celsius temperature can be transformed to Fahrenheit using the following linear transformation: $F = 32 + 1.8C$. Arithmetic means, variance, and Pearson product-moment correlation are permissible on data collected using interval scales.

Ratio Scales

Ratio scales have a true zero point. The true zero point is the point at which no amount of the attribute is present. Because a zero point can be determined, the ratio between actual scores of an attribute can be examined. Weight and height are two good illustrations of ratio scales. Using length as an example, let's say three rulers have lengths of 10, 20, and 60 centimeters (i.e., approximately 3.94, 7.87, and 23.62 inches, respectively). We can state the second ruler is twice as long as the first and the third is three times as long as the second because length is measured on a ratio scale. Unfortunately, ratio scales are rare in W&O psychology, but they do exist. One example is reaction time on performance tests.

A transformation allowed with ratio scales is $X' = b\ X$. Scores may be multiplied by a constant b, which changes the units of measurement, but not the ratio between two objects because this transformation does not change the zero point. Permissible statistics include the geometric mean.

So far, we have defined measurement, discussed some of the benefits of measurement, and described the four types of measurement scales. Now, we turn to the process of developing measures.

MEASURE DEVELOPMENT

While data can often be gathered using previously developed measures, W&O psychologists are often faced with a situation in which a new measure needs to be developed (e.g., because a previously developed measure lacks strong psychometric properties or because there is no measure for a specific attribute). The careful construction of measures ensures that they are dependable and accurate assessments of the attributes examined. If precautions are taken during measure development, fewer revisions will

have to be made later to increase the measure's usefulness. There are many types of measures, some of which require special steps or processes during their development. However, we will limit our discussion to the general process of constructing a measure. This general process involves determining the purpose of measurement, defining the attribute to be measured, developing a measure plan, writing items, conducting a pilot study and item analysis, selecting items, establishing norms, and determining the reliability and validity of the measure.

Determining a Measure's Purpose

The first step in developing a measure is to determine its purpose. Measures may be designed to assess an attribute for research purposes (e.g., measure of perceived social power and its relationship with various outcomes; Nesler, Aguinis, Quigley, Lee & Tedeschi, 1999), predict future performance (e.g., measure of cognitive ability used to select applicants most likely to succeed on the job), evaluate performance adequacy (e.g., measure of reading ability to assess proficiency), diagnose individual strengths and weaknesses (e.g., measure of performance completed by supervisor), evaluate programs (e.g., measure of participant attitudes towards a training program), or give guidance or feedback (e.g., measure of vocational interests used for career development). The intended use of the measure will guide the development by dictating factors like thoroughness of attribute definition, types of items included, and length and complexity of the measure. Clearly stating the purpose before constructing the measure will help ensure that the measure does what it was intended to do.

Defining the Attribute

The second step is to define precisely the attribute to be measured. Without a clear definition, it will be difficult to be sure the measure is assessing the desired attribute. To clarify the attribute, it is necessary to state what concepts are included in the attribute as well as what is excluded. For example, a measure of perceived social power in dyadic relationships may include power bases such as expert power, coercive power, and legitimate power, but exclude trustworthiness (Nesler et al., 1999). Also, it is helpful to explain the psychological processes underlying the attribute. Continuing with the social power example, a process may be that the display of specific nonverbal behaviors leads to a supervisor being perceived as having high coercive power (Aguinis & Henle, forthcoming; Aguinis, Simonsen & Pierce, 1998), resulting in a dissatisfactory relationship with his or her subordinate which, in turn, may adversely affect subordinate performance (Aguinis, Nesler, Quigley, Lee & Tedeschi, 1996). Further, it is important to state a theory describing the properties of the attribute (e.g., overall or global social power may be broken down into various power bases; Aguinis & Adams, 1998). A thorough description of the attribute provides a domain of content for writing items for the measure. Without a precise and clear definition of the attribute in question, we do not know what is to be measured or if it has been measured well (Guion, 1998).

Developing a Measure Plan

After the purpose of the measure is specified, and the attribute is defined, the next step is to establish the measure plan. The measure plan is a blueprint of the content, format, items, and administrative conditions for the measure to ensure it will be well constructed. First, the measure plan must include an outline of content to be included in the measure, which is derived from the attribute definition and will enable adequate coverage of important aspects of the attribute. Next, a description of the target population, who will be responding to the measure, including their demographics and reading level, is needed. Then, based on the target population, a description of the types of items to be used (e.g., multiple choice, true/false, short answer, essay, verbal responses), number of items, and examples of the items is written. Further, administrative procedures like instructions, how long the measure will take to administer, how it will be administered and by whom, and how it will be scored and interpreted, is outlined. Once the measure plan is written, experts and potential users should review it. A well thought-out plan enables appropriate items to be written and indicates intentions to design a good measure.

Writing Items

Next, using the definition of the attribute and the measure plan as guidelines, items are written. The closer these guidelines are followed, the more likely it is that items will measure the intended attribute. At this stage, twice the number of items desired for the final measure should be written because items will be discarded or revised. Although it is hard to know ahead of time how many items will be needed, Nunnally (1978) advises that at least 30 items are needed for a measure to have high reliability and, thus, initially at least 60 items should be written (we will discuss reliability later in the chapter). Note, however, that many measures in W&O psychology include fewer than 30 items and, nevertheless, estimates of their reliability are acceptable. Thus, given other things equal, although the number of items improves reliability, the number of items needed to reliably

measure an attribute depends on the attribute in question.

There are many guidelines for writing good items (e.g., Berk, 1984; Flaugher, 1990; Thorndike, Cunningham, Thorndike & Hagen, 1991). In general, items should be written as simply and clearly as possible, should not be vague or ambiguous, never contain double negatives, have the appropriate level of complexity given the target population, avoid sexist or otherwise offensive language, and when using negatively phrased items, the negative word should be capitalized, bolded, or underlined.

Conducting a Pilot Study and Item Analysis

After the items are written, they need to be reviewed, with the attribute definition and target population in mind, for appropriateness, difficulty, and clarity (Nunnally, 1978). The measure is then administered, following the procedures outlined in the measure plan, to a sample that is representative of the target population in terms of age, gender, ability level, and so forth. Also, the sample must be large in order to sufficiently evaluate the measure (e.g., at least five times as many subjects as items; Nunnally, 1978). Respondent reactions are gathered to evaluate the clarity of items and administrative procedures as well as to determine if the time limit is adequate.

Gathering feedback from respondents will provide information about the clarity of items and procedures. In addition, to gather more in-depth information about the quality of the items, an item analysis can be conducted. Item analysis helps eliminate items that are poorly written as well as items that are not relevant to the targeted attribute. Thus, item analysis can explain why a measure has a certain level of reliability or validity (Murphy & Davidshofer, 1998). The following three types of indicators can be computed to better understand item functioning: (a) distractor analysis, (b) item difficulty, and (c) item discrimination. In addition, Item Response Theory can be used to conduct a comprehensive item analysis. We discuss these issues next.

Distractor Analysis

Distractor analysis evaluates multiple choice items that may appear on measures of achievement or ability. The frequency that respondents choose each response is calculated to determine the effectiveness of distractors (i.e., incorrect responses). The frequencies for the distractors should be about equal. If a distractor is chosen less frequently than the others, it may be too transparent and should be replaced. Alternatively, if a distractor is selected more often than the others, it may be tapping partial knowledge of the item or indicate that the item is misleading.

Item Difficulty

Item difficulty evaluates how difficult it is to answer an item correctly. An indicator of item difficulty, known as the p value, can be calculated to determine the percentage of respondents answering the item correctly. The p value is computed by dividing the number of individuals answering the item correctly by the total number responding to the item. A high p value indicates that most respondents answered the item correctly, and thus the item may be too easy. In contrast, a low p value indicates a difficult item since few were able to answer the item correctly. Ideally, the mean item p value should be about .5, which indicates a moderate difficulty level for the measure. Extreme p values do not discriminate among individuals, and items with such extreme values should be omitted or revised. However, an average p value of .5 may not be optimal for all measurement purposes (e.g., assessing the cognitive ability of applicants for an engineering position may require a measure with difficult items and thus a low mean p value).

Item Discrimination

Item discrimination analysis is appropriate for most measures and it evaluates whether the response to a particular item is related to responses on the other items. It determines which items are best measuring the attribute and whether the items are differentiating between those who do well on the measure and those who do not. That is, those who do well on a measure overall should answer an item correctly while those performing poorly on a measure should answer the item incorrectly. There are several statistics that serve this purpose, but we will limit our discussion to the discrimination index and the item–total score correlation.

The discrimination index d compares the number of respondents who answered an item correctly in the high scoring group with the number who answered it correctly in the low scoring group. If an item is discriminating adequately, more respondents with high scores should answer the item right as compared to respondents with low scores. To calculate d, the top and bottom scoring groups are selected (this can be done by taking the top and bottom quarters or thirds), and d is computed using Equation 2.1:

$$d = \frac{p_u}{n_u} - \frac{p_l}{n_l} \qquad (2.1)$$

where p_u and p_l are the number of individuals passing the item in the upper and lower scoring groups, and n_u and n_l are the size of the upper and lower groups, respectively. Items with large, positive d scores are good discriminators; that is, the item is harder for the lower scoring group and easier for the higher scoring group. An item with a negative d

score should be discarded because negative scores indicate the item is easier for those who do poorly on the measure overall.

The second and most popular method for determining the ability of an item to discriminate is the correlation between an item and the total score on a measure. Items with high, positive item–total score correlations are related to the attribute the measure is examining and, thus, contribute to the measure's reliability. These items also have more variance than items with low item–total score correlations, which allows the measure to discriminate between individuals who do well on the measure and those who do not. Any items with item–total score correlations that are low or near zero should be revised, omitted, or replaced. Item–total correlations above .30 are preferred (Nunnally, 1978).

Item Response Theory

In addition to the statistics described above, Item Response Theory (IRT) can be used to conduct a comprehensive item analysis. IRT explains and analyzes the relationship between responses to individual items and the attribute being measured (Hulin, Drasgow & Parsons, 1983; Lord, 1980; Thissen & Steinberg, 1988). Specifically, IRT explains how individual differences on a particular attribute affect the behavior of an individual when he/she is responding to an item. That is, individuals with a large amount of the attribute will be more likely to respond correctly to an item requiring more of that attribute. Thus, the amount of an attribute can be estimated based on how an individual responds to items on the measure.

IRT holds assumptions about the mathematical relationship between an individual's level of the attribute and the likelihood that he/she will answer an item in a certain way. These assumptions and responses to the measure combine to form an item-characteristic curve (ICC). The ICC is a graphical representation of the probability of selecting the correct answer on an item due to an individual's level of the attribute. If an item is assessing the attribute, the probability of choosing the correct answer should increase as the level of the attribute does (Drasgow & Hulin, 1991).

By examining the ICC, we can determine item difficulty, discrimination, and the probability of answering correctly by guessing. Item difficulty is evaluated by examining the position of the curve. If the item is difficult, which is defined as requiring a large amount of the attribute in order to answer the item correctly, the curve starts to rise on the right side of the ICC plot. Alternatively, for easy items the curve begins to rise on the left side of the plot. Item discrimination is assessed by the steepness of the ICC. The flatter the curve, the less the item discriminates among individuals. Finally, from the ICC the probability of guessing the correct answer when an individual is low on the attribute can be determined. The higher the lower asymptote of the curve is, the easier it is to guess correctly on that item. That is, the higher the curve begins on the *y*-axis, the higher the probability of guessing.

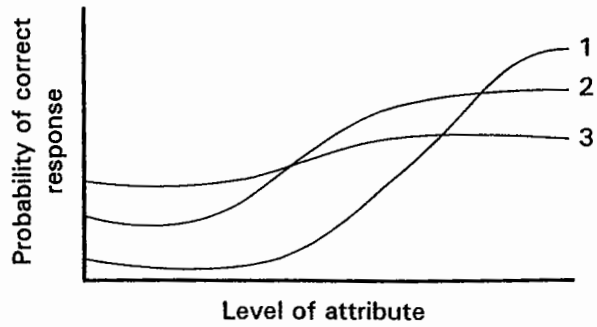

Figure 2.1 *Illustration of item-characteristic curve (ICC) for three hypothetical items*

Consider the ICC shown in Figure 2.1. Items 2 and 3 are easier than 1 because their curves begin to rise further to the left on the plot. Item 1 is the most discriminating while item 3 is the least because its curve is relatively flat. Finally, item 3 is the most susceptible to guessing because it begins higher on the *y*-axis.

Selecting Items

Based on the results of the pilot study and item analysis, items are selected and revised. A common method for selecting items uses the results of the item analysis to rank items based on their item–total score correlations from highest to lowest (Nunnally & Bernstein, 1994). A group of the top items is selected (e.g., 30 items) and the reliability of the items is calculated using coefficient alpha (reliability will be discussed more thoroughly in the next section). If the reliability, of those items is high (e.g., $\geq .80$), no more items are selected. If the reliability is not high enough, then five to ten more items are added, depending on the gap between current and desired reliability, and then reliability is re-computed for the new set of items. This iterative process is repeated until the desired level of reliability is reached. Note that items with low item–total score correlations (i.e., below .20) should not be added because they do not improve reliability. Also, if reliability is no longer increasing or it is decreasing, the process of adding items should stop.

Once the desired reliability level is reached, a frequency distribution of scores on the entire instrument is plotted. A normal distribution is ideal, but if the distribution is skewed, adjustments can be made. When the distribution is positively skewed (i.e., scores cluster at the lower end of the plot), the items are too hard. Thus, items with low p values should be replaced with ones that have higher

p values. Alternatively, when the distribution is negatively skewed (i.e., scores cluster at the high end of the plot), the items are too easy and items with high p values should be replaced with ones that have lower p values.

Establishing Norms

If the measure will be used to make decisions about individuals, norms should be established. Norms are used to provide standards for interpreting the scores of individuals, and are determined by gathering scores on the measure from a representative cross-section of individuals who are members of the target population (e.g., women and men, various levels of socioeconomic status; see Angoff, 1971 for more details). Norms are typically expressed in either standard scores (i.e., z) or percentiles. Standard scores are scores on a measure referenced to the normal distribution (i.e., $z = [X-M]/SD$, where X is an individual's score on the measure and M and SD are the measure's mean and standard deviation, respectively). Percentiles indicate the percentage of individuals in the sample who score below a particular score.

Determining Reliability

Reliability refers to the extent that a measure is dependable, stable, and consistent over time. If a measure is reliable, there is consistency between two sets of scores on a measure. For example, if a personality measure is administered to a job candidate and the candidate does not get the job, but applies for a similar position 6 months later and takes the measure again, the scores from the two administration periods should be similar if the measure is reliable. If the scores are considerably different, they may contain errors of measurement.

The concept of reliability assumes that scores obtained from a measure include a 'true' score or accurate representation of an individual's level of the attribute being measured. For example, if we give a typing test to job applicants, we assume that the test is assessing their true ability to type. However, in addition to the true component, measures in W&O psychology contain error. Errors of measurement are unsystematic or random and affect the obtained score on a measure, but are not related to the attribute being measured. Errors of measurement can be the result of changes in individuals responding to the measure (e.g., fatigue, anxiety) that affect their scores at one administration but not at another, or the result of changes in administrative conditions (e.g., noise, poor lighting). These errors prevent direct measurement of true scores and force us to rely on obtained scores as estimates of true scores. Thus, a score obtained from a measure has a true score component as well as an error component. Equation 2.2 demonstrates this relationship:

Table 2.1 *Sources of error in the different reliability estimates*

Method of estimating reliability	Source of error
Test–retest	Time sampling
Parallel forms (immediate)	Content sampling
Parallel forms (delayed equivalent)	Time and content sampling
Split-half	Content sampling
Cronbach's α	Content sampling
Kuder–Richardson 20	Content sampling
Interrater agreement	Interrater consensus
Interclass correlation	Interrater consistency
Intraclass correlation	Interrater consistency

$$X_{\text{obtained score}} = X_{\text{true score}} + X_{\text{error}} \qquad (2.2)$$

In order to increase the reliability of a measure, errors of measurement must be minimized. Ideally, they should be completely eliminated. By decreasing error and subsequently increasing reliability, it is more likely the measure will reflect an individual's true possession of the attribute measured. If the measure contains a substantial amount of error, we cannot be confident that it is measuring the attribute. However, what constitutes errors of measurement varies from one situation to another depending on the purpose of measurement. Different methods of estimating reliability treat some factors as error while others do not. In sum, what is classified as errors of measurement depends on the purpose of measurement and subsequently, the method used to estimate reliability.

Methods for Estimating Reliability

Methods for estimating the reliability of a measure use the correlation coefficient to assess the relationship or degree of consistency between two sets of scores. The reliability coefficient can range from 0 to 1, with numbers closer to one indicating high reliability and little measurement error, and values closer to zero indicating low reliability and a large amount of measurement error.

Next, we discuss the following four methods for estimating reliability: Test–retest, parallel forms, internal consistency, and interrater. Each method calculates a reliability coefficient, but they differ regarding what they define as error (see Table 2.1 for a summary). Thus, the choice for a method to estimate reliability depends on the purpose of the measure as well as what is considered to be an important source of error.

Test–retest reliability involves giving the measure to the same group of individuals at different points

in time. Scores are correlated from Time 1 and Time 2 to get a reliability coefficient referred to as coefficient of stability, which assesses the amount of error due to random fluctuations in scores over time. Thus, error is defined as changes in individuals (e.g., anxiety, fatigue, mood, health) and changes in measure administration (e.g., lighting, noise, distractions) that affect scores at one time but not at the other. The coefficient of stability can assess if a measure given now will be representative of the same individuals at a later time. In sum, this method should be used to estimate reliability when the attribute being measured is believed to be stable over time because this method can determine if the measure is free from error associated with the passage of time.

If the measure is reliable, scores should only change slightly from Time 1 to Time 2 and the rank order of individuals should stay the same. However, the reliability coefficient may differ depending on the length of time between administrations. If the time period is too short, the effects of memory may inflate the reliability coefficient because respondents may be able to recall how they answered the measure the first time. However, if the time period is too long, learning may affect the reliability coefficient. If individuals learn the answers to the items on the measure or if they learn information that changes how they respond to the measure, reliability may be underestimated because their scores will have changed from one administration to another. Although there is no magical number for the time interval between measure administrations, there should be at least 8 weeks between administrations (Nunnally, 1978), but not more than 6 months.

Parallel forms, also called alternate or equivalent forms, is a second method for estimating reliability. This method examines the consistency with which an attribute is measured across different versions of a measure. This is achieved by calculating the correlation between two forms to obtain a coefficient of equivalence. The two forms can be administered close together but, to prevent order effects, half of those taking the measure should be given form A first and the other half form B. Error using this method is defined as content sampling or samples of items that are nonequivalent. That is, high coefficients of equivalence indicate that the content sampled on the two versions of the measure are equivalent and, thus, measuring the same attribute. This method can be modified to assess error due to both content and time sampling. The modified version, labeled *delayed equivalent forms*, estimates reliability by increasing the amount of time between administrations (like test–retest) to get a coefficient of stability and equivalence by computing the correlation between one form given at Time 1 and the other form given at Time 2.

Unfortunately, it is hard to design equivalent measures. To be equivalent, measures must have the same number and type of items, same difficulty level, and the means and standard deviations of the scores obtained by respondents on both forms should be the same. Because it is hard to design equivalent forms of a measure, reliability coefficients determined by this method will be conservative estimates of reliability. Despite the difficulties associated with this method, parallel forms is useful for measures that are likely to be administered repeatedly (e.g., achievement measures).

The above discussion of measurement equivalence focused on parallel forms (Lord & Novick, 1968). Parallel measures have equal regressions of observed scores on true scores and equal error variances, and they can be used interchangeably. However, there are additional, less stringent, types of measurement equivalence. First, Tau-equivalent measures have equal regressions of observed scores on true score, but possibly different error variances (Jöreskog, 1971). Second, congeneric measures assess the same underlying construct (i.e., they are linearly related), but have different regressions of observed scores on true scores as well as different error variances (Jöreskog, 1971) (we refer readers to Vandenberg and Lance, 2000, for a more detailed discussion of measurement equivalence).

Internal consistency is a third method for estimating reliability. Internal consistency determines the degree to which various items of a measure correlate with each other. Error is defined as item heterogeneity; the more homogenous the items, the lower the error. This is important because items that are highly intercorrelated indicate they are measuring the same attribute. Three popular methods of determining internal consistency (i.e., split-half, Cronbach's coefficient alpha, and Kuder–Richardson 20) are discussed below.

The split-half method estimates internal consistency by administering a measure once and splitting it into two equivalent halves after it has been given to get two scores for each individual. This method is based on the premise that any item or group of items should be equivalent to any other item or group. The correlation between the two halves is a coefficient of equivalence that demonstrates the similarity of responses between the two halves. Thus, error is defined as inconsistency in content sampling between the halves for the attribute being measured. However, this reliability coefficient is based on a single administration of the test, so it does not take into account errors of measurement that occur over time (e.g., changes in individuals or administration) and, thus, it provides a liberal estimate of reliability.

Like parallel forms, equivalent halves need to be equal in terms of content, difficulty, and means and standard deviations of responses. The measure can be divided by placing the odd items in one half and even items in the other or, preferably, by random selection of items. The resulting coefficient of equivalence from the split-halves is the reliability of

a measure half the length of the original one, which underestimates reliability because reliability increases as number of items does. Therefore, the Spearman–Brown prophecy formula shown in Equation 2.3 is used to determine the reliability of the entire measure:

$$r_{nn} = \frac{nr_{11}}{1 + (n-1)r_{11}} \qquad (2.3)$$

where n is the factor by which a measure is increased (e.g., $n = 2$ indicates the measure is doubled in size), r_{11} is the obtained reliability coefficient, and r_{nn} is the estimated reliability of a measure n times as long. For example, a mathematical ability measure is divided into two halves with odd items in one and even in the other. The correlation between the two halves is .68, which represents the reliability for a measure half the length of the original. If we use these values in Equation 2.3, the estimated reliability for the entire measure is:

$$r_{nn} = \frac{2(.68)}{1 + (2-1)(.68)} = .81$$

The second method for estimating internal consistency is Cronbach's α (see Cortina, 1993, for a review). Like split-half, Cronbach's α indicates the degree that items on a measure are correlated with each other. However, this method recognizes that there are many ways to divide a measure, so it takes the average of all possible split-halves of a measure (Kuder & Richardson, 1937). Cronbach's α is computed when there is a range of responses to items on a measure (e.g., 'always,' 'sometimes,' 'occasionally,' 'never'). As noted above, this type of reliability coefficient is determined by taking the average of all the possible split-halves of the measure so that it can assess how similar items are to each other and, thus, whether they are measuring the same attribute. If reliability is low, the measure may be assessing more than one attribute. The equation for computing Cronbach's α is the following:

$$r_{tt} = \frac{k}{k-1}\left(\frac{\sigma_t^2 - \sum\sigma_i^2}{\sigma_t^2}\right) \qquad (2.4)$$

where k is number of items included in the measure, σ_t^2 is the variance of total scores on the measure, and $\sum\sigma_i^2$ is the sum of the variances of item scores.

A special case of Equation 2.4 occurs when responses to items are binary in nature (i.e., two responses such as true or false, and right or wrong). For this special case, Kuder and Richardson (1937) developed the following variation of Equation 2.4 (i.e., KR-20):

$$r_{tt} = \frac{k}{k-1}\left(\frac{\sigma_t^2 - \sum pq}{\sigma_t^2}\right) \qquad (2.5)$$

where k and σ_t^2 are defined in Equation 2.4, and $\sum pq$ is the sum of all the products of p and q for each item, with p representing the number of individuals who pass the item and q representing the number of individuals who fail the item.

Interrater reliability is a fourth method for estimating reliability. This method is useful when a measure is subjectively scored (e.g., observational data, ratings) and judgment is involved because raters' biases and inconsistencies (e.g., raters interpret rating standards differently or inconsistently) may influence ratings (Kraiger & Aguinis, 2001). Interrater reliability determines the consistency among raters and whether characteristics of the raters are determining the ratings instead of the attribute being measured.

In general, interrater reliability determines the degree of consistency across raters when rating objects or individuals. A distinction is made, however, between interrater consensus (i.e., absolute agreement between raters on some dimension), and interrater consistency (i.e., interrater reliability, or similarity in the ratings based on correlations or similarity in rank order) (Kozlowski & Hattrup, 1992). We discuss the following three ways to calculate interrater reliability: Interrater agreement, interclass correlation, and intraclass correlation.

Interrater agreement focuses on exact agreement between raters on their ratings of some dimension. The most commonly used statistics are (a) percentage of rater agreement, (b) Tinsley and Weiss's (1975) index of agreement T, (c) Kendall's (1948) coefficient of concordance W, and (d) Cohen's (1960) kappa (κ)[26]. When a group of judges rates a single attribute (e.g., organizational climate), the degree of rating similarity can be assessed by using James, Demaree and Wolf's (1984, 1993) r_{wg} index. All of these indices focus on the extent to which raters agree on the level of the rating or make essentially the same ratings.

Interclass and intraclass correlations are indices of consistency, are correlational in nature, and refer to proportional consistency of variance among raters (Kozlowski & Hattrup, 1992; Lahey, Downey & Saal, 1983; Lawlis & Lu, 1972; Shrout & Fleiss, 1979). Interclass correlation is used when two raters are rating multiple objects or individuals (e.g., performance ratings). Pearson product-moment correlation r and Cohen's (1960) weighted kappa (κ)[29] are the two most commonly used statistics. Intraclass correlation (ICC) is typically used when multiple raters are rating objects or individuals. This method determines how much of the differences among raters are due to differences in individuals on the attribute being measured and how much is due to errors of measurement.

There are six different forms of intraclass correlations, which allow for assessing situations including a group of raters and a single and/or multiple dimensions. Intraclass correlation is typically expressed as

the ratio of the variance associated with targets (e.g., objects or individuals being rated in performance evaluations) over the sum of the variance associated with targets plus error variance based on the results of an analysis of variance (see Lahey et al., 1983 or Shrout & Fleiss, 1979, for the formulae for computing each of the six forms of intraclass correlations). ICC(1,1) is used to evaluate the reliability of multiple raters making judgements about multiple targets on a single dimension; ICC(2,1) is appropriate when the judges are randomly sampled from the larger population of judges, but each judge rates each of the targets; ICC(3,1) is used when each target is rated by each of the same judges and there are no other possible judges of interest; ICC(2,1) differs from ICC(3,1) in that ICC(2,1) allows one to generalize reliability to other judges while ICC(3,1) is used when there is an interest in the reliability of only a single judge or a fixed set of judges. The remaining three forms of intraclass correlations are identical to the above but include cases when multiple dimensions are rated for each target.

Interpreting Reliability Coefficients

Reliability coefficients are the means to an end. The end is to produce scores that measure attributes consistently across time, forms of a measure, items within a measure, or raters. We compute a reliability coefficient to understand if our scores are consistent. But, what exactly do the reliability coefficients tell us? What constitutes an acceptable level of reliability for our measure?

A reliability coefficient can be translated as the percentage of score variance on a measure that results from 'true' differences in the attribute being measured. For example, if a measure of cognitive ability has a reliability coefficient of .92, this means that 92% of score variance can be accounted for by differences in cognitive ability among respondents, and 8% can be attributed to errors of measurement. The acceptable size of a reliability coefficient depends on the purpose of the measure. If the measure is used to compare individuals (e.g., selection measure), the reliability coefficient should be greater than .90 (Nunnally, 1967). But, .70 may be sufficient for most measures in W&O psychology and even lower coefficients may be acceptable for research purposes.

Standard Error of Measurement

Reliability estimates provide information about the consistency of most individuals' scores on a measure. However, they do not provide information about the consistency of a given individual's score on the measure (Aguinis, Cortina & Goldberg, 1998). Rather, reliability reflects the error associated with a particular measure. To gather information about how much error we can expect for an individual's score on a measure, we can calculate the standard error of measurement. Standard error of measurement provides an estimate of the standard deviation of a normal distribution of scores that an individual would obtain if he/she responded to the measure an infinite number of times. The standard error of measurement σ_{Meas} is computed as follows:

$$\sigma_{Meas} = \sigma_x \sqrt{1 - r_{xx}} \tag{2.6}$$

where σ_x is the standard deviation of the distribution of obtained scores, and r_{xx} is the reliability estimate for the measure. Using the standard error of measurement, we can derive confidence intervals that estimate the range of scores that will, at a certain probability level, include an individual's true score (cf. Equation 2.2). If the standard error of measurement for a reading measure is 2.21 and an individual obtained a score of 60 on the measure, by adding and subtracting the standard error from the obtained score (60 ± 2.21), a confidence interval of 57.79 to 62.21 is derived. This range of scores can be interpreted as if the individual was given the test 100 times, the reading scores would fall between 57.79 and 62.21 about 68 times (i.e., 68% confidence interval). Note that the level of confidence can be increased from 68% to 95% by adding and subtracting two standard errors from the obtained score (i.e., the interval would go from a low of $60 - 4.42 = 55.8$ to a high of $60 + 4.42 = 64.42$).

The standard error of measurement can aid decision-making about individuals in several ways. For example, if we are deciding whether to hire Sarah by comparing her score of 60 to a cutoff score of 65, the standard error of measurement can help with this decision. Sarah's score is only five points away from the cutoff, but when we examine the 68% confidence interval calculated earlier (i.e., 57.79 to 62.21), we estimate that it is not likely that she will meet this cutoff upon retesting. Further, the standard error can be used to assess whether two applicants' scores on the reading test are different from one another (cf. Aguinis, Cortina & Goldberg, 2000). For instance, Sarah scored a 60 and Rachel scored a 62. The standard error is 2.21 and the difference between the candidates is only 2 points; therefore, upon retesting, Sarah may score higher than Rachel. The standard error can also be used to evaluate scores between groups. For example, it can determine if scores for men and women differ significantly.

Improving Reliability Coefficients

We want the reliability of our measures to be as large as possible to ensure that our measures are dependable, consistent, and stable over time. However, the size of reliability coefficients may be limited by several factors and if we are not aware of these factors and do not take them into consideration, we may over or underestimate reliability. First, the method for estimating reliability can affect the size of the obtained coefficient. As described above, the various methods for estimating reliability define error differently and, consequently, the reliability

coefficient for a measure differs depending on the method used. Some methods are more liberal (e.g., split-half), which may overestimate reliability, while others are more conservative (e.g., parallel forms), which may underestimate reliability.

Second, variability in scores can influence the size of reliability coefficients. If we administer a measure of perceived social power (i.e., ability to influence) in a flat organization where all employees have the ability to influence each other, there will be no variance in scores because everyone will score very high. Variability among measure scores allows for differentiation among the individuals taking the measure. If all respondents score a 20, we cannot differentiate among them based on social power. However, if there is a wide range of scores (e.g., 20, 17, 13, 12, 9, 7, 6, 1), we are able to make many differentiations among pairs or groups of individuals. In addition, variability can be affected by individual differences. As individual differences (i.e., variability) among scores increase, so does the correlation between them, which makes it easier for the measure to differentiate among individuals. Thus, other things equal, the greater the variability, the greater the reliability.

Third, as the length of a measure increases, so does its reliability. If the number of items relevant to measuring a particular attribute increases, we are able to obtain a more accurate picture of an individual's true score on that attribute. We can use Spearman–Brown's prophecy formula (i.e., Equation 2.3) to demonstrate the relationship between measure length and reliability. Assume we are using a measure of extroversion, which contains 15 items and has a reliability of .80, and we double the size of the measure. Entering these values in Equation 2.3 yields:

$$r_{nn} = \frac{2(.80)}{1 + (2 - 1)(.80)} = .89$$

By doubling the number of items on the extraversion measure to 30, we increased its reliability from .80 to .89. However, a caveat must be made. Indiscriminately adding items will not increase the reliability, especially internal consistency. Of course, additional items must be similar to previous ones and be relevant to the attribute being measured.

Fourth, the characteristics of the sample used also affect reliability. Sample size will influence the magnitude of the reliability coefficients because larger samples will have less sampling error than small samples (Aguinis, forthcoming), thus providing a better estimate of reliability. In addition, the sample must be representative of the population the measure is going to be used for or reliability will be over or underestimated.

Generalizability Theory

Our discussion of reliability thus far has followed a classical approach. However, an alternative approach to reliability is generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972), which is a process of determining the limits of the generalizability of inferences derived from measures. That is, it assesses the situations (e.g., different people, places, and times) to which inferences made from a measure can be applied and, thus, evaluates how well a measure is assessing an attribute.

The classical approach to reliability that we have discussed also explores issues of generalizability, but only in a limited way. For instance, the coefficient of stability assesses generalizability over time while internal consistency determines the extent to which inferences generalize across items on a measure. However, generalizability theory takes into account many sources of error simultaneously instead of examining one at a time and shows how much total variance is a result of each source of error.

Generalizability theory uses experimental studies to determine how much variance is attributable to different sources of error. A *generalizability study* is designed to evaluate the extent to which results obtained using a measure are consistent despite different administrative conditions. Information is collected from individuals responding to the measure under different circumstances to determine a coefficient of generalizability. A *decision study* evaluates decisions made based on a measure's scores. Thus, it tells us how sound are the conclusions made using a measure. For a more detailed discussion of generalizability theory, we refer readers to Cronbach et al. (1972) and Brennan (1992).

In sum, scores gathered using a measure are affected by numerous sources of error. As shown in Equation 2.2, observed scores have a true score as well as an error component. A reliability analysis allows us to estimate the extent to which observed scores are influenced by a random error component. A large reliability coefficient (i.e., small standard error of measurement, cf. Equation 2.6) suggests that scores are consistent. However, consistency does not ensure accuracy. For example, a scale may be consistently off by 20 pounds. The scale lacks random measurement error and, thus, scores are very consistent. However, scores do not represent true weight, and therefore decisions made based on these scores (e.g., change patterns of eating behavior) may be incorrect. The issues of whether scores, and decisions made based on scores, are accurate are issues of validity. We discuss this topic next.

Gathering Evidence of Validity

Validity refers to the utility of the inferences made from a measure's scores. Inferences made from measures can involve measurement issues (e.g., Is this measure of leadership effectiveness really assessing who is an effective leader?) or decisions (e.g., Can the measure of leadership effectiveness

help predict who will be successful as a manager?). Thus, the process of validation evaluates whether a measure is assessing the attribute it is supposed to and if a measure can be used to make accurate decisions. The measure itself is not validated, rather the inferences about what the measure is assessing and decisions made from the scores are. Empirical investigations are conducted to gather evidence to support these inferences. Evidence is continually gathered to evaluate a measure and to revise it if it is not fulfilling its intended purposes. Therefore, validation is an ongoing process. In sum, validity provides evidence attesting to what attribute a measure is assessing, how well it measures that attribute, and what decisions can be made from a measure's scores.

Originally it was posited that there was a particular type of validity that was appropriate for a given type of measure. The specific measurement purpose dictated which type of validity was used to establish validity. However, validity is now viewed as a unitarian concept. There are not different types of validity, rather different types of evidence for determining the validity of a measure (Binning & Barrett, 1989; Cronbach, 1988; Landy, 1986). Thus, many types of evidence should be gathered to support the inferences and decisions that are made based on a measure's scores. Next, we discuss the following three types of validity evidence: Content, criterion, and construct. Although they are discussed separately, they are interrelated and a combination of them is necessary to determine what inferences can be made from a measure's scores.

Content-Related Evidence

Content-related validity evidence examines the adequacy of domain sampling; that is, whether a measure is assessing the attribute it is intended to measure. This is demonstrated when the content of a measure (i.e., items) is judged to be a representative sample of the content of the attribute under consideration. Thus, this method of gathering evidence of validity relies on judgments of potential users and experts.

Establishing content-related evidence begins during the construction of a new measure. Developing a well thought-out plan for measure construction (as described earlier in the chapter) and adhering to that plan provides evidence of content validity. When potential users of a measure and experts of the attribute being measured agree the plan was well developed and implemented, the measure is most likely to be a representative sample of the content of the attribute. Thus, following the steps outlined previously for developing a measure will help establish content-related evidence.

The content validation process starts with a description of the content domain. The content domain is the total set of items that could be used to measure an attribute (Guion, 1977), and there are three parts to the content domain. First, a definition of the domain or attribute to be measured must be clarified. For example, if we are developing a measure of job satisfaction, a definition may be 'An individual's affective reaction to his/her job.' Next, the different areas or categories of the attribute to be included in the measure must be specified. For our job satisfaction measure, we may include the following categories of satisfaction: Pay, supervision, coworkers, and the work itself. Finally, the relative importance of the categories must be established. For example, if we believe that satisfaction with pay and the work itself are more important than satisfaction with supervision and coworkers, we must weigh them more heavily so that they will comprise a larger portion of the measure (e.g., pay = 30%, work itself = 30%, coworkers = 20%, supervision = 20%).

After the content domain has been described, and items have been written following this description, we can compare the content of the measure to the content domain to provide evidence of content validity. Each item on the measure is evaluated against the definition and classified into a category to determine if it falls within the domain of the attribute. The measure as a whole is also compared to the content domain to evaluate if the measure samples all the areas of the attribute and if there are more items representing the areas that were ranked as more important. The closer the measure matches the content domain, the stronger the evidence regarding content validity.

The extent to which experts agree on the content validity of a measure can be calculated using Lawshe's (1975) Content Validity Ratio (CVR). To compute CVR, experts who are familiar with the attribute measured (e.g., recognized researchers in the field of job satisfaction) rate whether each item is essential, useful but not essential, or not necessary for measuring the attribute. Their ratings are used in Equation 2.7:

$$\text{CVR} = \frac{n_e - N/2}{N/2} \qquad (2.7)$$

where n_e is the number of experts that rated the item as essential, and N is the total number of experts. The resulting CVR represents the overlap between the content of the attribute and the content of the measure. For example, if 10 experts rate an item of a measure and eight of them believe the item is essential, CVR is:

$$\text{CVR} = \frac{8 - 10/2}{10/2} = .6$$

CVR can range from -1 to $+1$ with values closer to $+1$ indicating that more experts agree the item is essential. In the above example CVR is .6, which is close to 1, so most experts believed that there was overlap between the content of the item and the

content of the attribute. Further, CVRs for all the items on a measure can be averaged to determine the extent that experts believe the entire measure overlaps with the attribute content.

Criterion-Related Evidence

As mentioned throughout this chapter, measurement is used to make important decisions about individuals. The second type of evidence, criterion-related, is particularly suited to determine if a measure can be used to make predictions and/or decisions. Thus, a measure demonstrates criterion-related evidence of validity if it can be used to make accurate decisions. Criterion-related evidence involves correlating scores on a predictor (i.e., measure of an attribute) with some criterion (e.g., measure of decision outcome or level of success) to determine if accurate decisions can be made from scores. There are two types of studies, predictive and concurrent, that can be designed to test the relationship between a predictor and a criterion.

Predictive validation studies focus on the prediction of future behavior. Predictive studies begin with obtaining scores from a random sample of the population in which decisions will be made, thus ensuring study results are generalizable. Next, decisions are made without using scores from the measure. After the decision is made, scores on a criterion are gathered and the correlation between the measure and criterion is calculated. An example of a predictive validation study is when job applicants are given a measure of integrity and selected for the job without considering their scores. After applicants have been hired and on the job for a period of time, information on absenteeism, theft, and other counterproductive behaviors is gathered and correlated with the integrity measure to determine its predictive ability.

Unfortunately, predictive studies are not as practical as concurrent studies because they require not using the measure to make decisions and a time delay before the criterion data are collected. Thus, *concurrent validation studies* are more commonly implemented to determine whether using a measure leads to accurate decisions. Concurrent evidence evaluates if an individual's level of an attribute is adequate to achieve the criterion at the present time. Concurrent validation studies gather scores on the predictor and criterion at about the same time from a preselected population. Then, the correlation between predictor and criterion scores is obtained. For example, current employees could complete the integrity measure and their employment files could be checked at the same time to determine how often they are absent, if they have been disciplined for theft, and any other information regarding counterproductive behaviors. The predictor and the criteria are then correlated to determine the predictive value of the integrity measure. Although concurrent studies are more practical than predictive, they may not be generalizable to the broad population because these studies rely on a preselected sample instead of randomly selecting from the target population.

Note that in both types of criterion-related validation studies (i.e., predictive and concurrent), there is an artificial reduction in the variance in one or more of the variables under consideration. This artificial reduction in variance, often labeled range restriction or censorship, deserves attention because it might have an impact on correlation coefficients, regression coefficients, and means. For example, a reduction in variance decreases the size of validity coefficients so that results obtained using restricted samples may underestimate actual validity coefficients. There are three types of range restriction (see Thorndike, 1949: 169–180 for a more detailed discussion regarding range restriction). Cases I and II are often labeled 'direct or explicit restriction', and Case III is often labeled 'indirect or implicit restriction'. Case I is a situation in which we are interested in the relationship (e.g., correlation) between predictor variable X and criterion variable Y, variable X's range is restricted, and we have information regarding variable Y's variance in both the restricted (sample) and unrestricted (population) groups, and information regarding variable X's variance in the restricted group only. This situation is not likely to be encountered by most W&O psychologists. Case II involves a situation in which we are also interested in the correlation between X and Y, variable X's range is restricted, and we have information regarding X's variance in both the restricted (sample) and unrestricted (population) groups, and information regarding variable Y's variance in the restricted group only. This situation is more frequently encountered by W&O psychologists, and it is particularly common in the personnel selection literature. Finally, Case III involves a situation in which we are also interested in the correlation between X and Y, but restriction of range has taken place on a third, or more, of often unspecified variables which are correlated with X and Y. Because of the correlations between X and the unspecified variable(s), and Y and the unspecified variable(s), we have variance information regarding both X and Y for the restricted groups only (Aguinis & Whitehead, 1997). Case III is the most pervasive type of range restriction in the personnel selection literature (Aguinis & Whitehead, 1997; Thorndike, 1949).

Construct-Related Evidence

Construct-related evidence is the third type of evidence that can be used to determine if inferences made from a measure's scores are valid. Construct, like content-related evidence, is the process of accumulating evidence to establish whether the measure is assessing the attribute it is intended to assess. However, instead of evaluating the measure plan and determining whether the measure includes a representative sample of the content of the attribute,

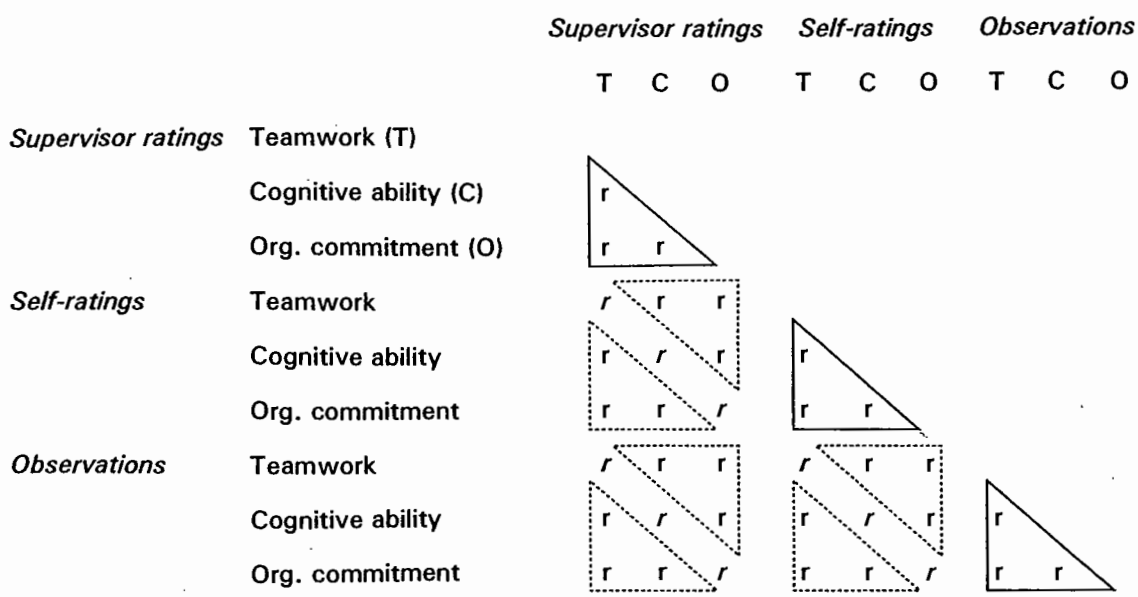

Figure 2.2 *Illustration of the multitrait–multimethod matrix for hypothetical study on the relationship among measures of teamwork, cognitive ability, and organizational commitment*

construct-related evidence investigates hypothesized relationships between a construct and other constructs to assess if actual relationships are similar to predicted ones. A construct is an abstract characteristic or attribute that a measure is believed to be assessing. Conscientiousness, extraversion, social power, job satisfaction, and intelligence are examples of constructs. Because we cannot observe these constructs, we need measures to be concrete and operational indicators of them. Thus, construct-related evidence involves conducting studies to support that a measure is indeed assessing the proposed construct by relating a measure to measures of other constructs.

The process of gathering construct-related evidence begins with defining the construct and identifying observable behaviors that operationally define the construct. Upon determination of observable behaviors of the construct, relationships among the different behaviors are investigated. If the observable behaviors are good indicators of the construct, they should be highly intercorrelated indicating that they are measuring the same concept. Once the internal consistency of the behaviors has been established, a nomological network is constructed (Cronbach & Meehl, 1955). A nomological network is a pattern of proposed relationships between the construct, its observable behaviors, and other constructs and observable behaviors. This network specifies variables to which the construct should and should not be related. Studies are then conducted to determine the degree that actual relationships match the expected ones delineated in the nomological network. The closer the match between the hypothesized nomological network and the actual relationships, the stronger the evidence of construct validity.

There are different types of studies that can be designed to support the hypothesized relationships between the construct and other variables. The more evidence accumulated from different sources, the more confident we can be that the measure is assessing the construct. One type of study that examines several types of evidence of construct validity is the multitrait–multimethod approach developed by Campbell and Fiske (1959). To conduct a study using this approach, data must be gathered on at least two constructs each measured by at least two different methods (e.g., supervisor ratings, observations, self-reports). Correlations among the different constructs measured by different methods are calculated to form a multitrait–multimethod matrix. The matrix shown in Figure 2.2 includes hypothetical correlations among measures of teamwork, cognitive ability, and organizational commitment using supervisor ratings, self-ratings, and observer ratings.

The first type of evidence provided by the matrix is convergent validity, which examines whether different methods of assessing the construct produce similar results. If results obtained using different methods are highly correlated, we can be more confident that our measures are assessing the intended construct. Convergent validity is determined by examining the italicized correlations in the matrix.

Next, we can assess divergent validity; that is, whether measures hypothesized not to be related are not related. Examining the correlations within the dashed triangles provides evidence regarding divergent validity.

Then, we can evaluate method bias, which is the inflation of correlations due to a common method of measurement. This is determined by investigating the correlations between different constructs using

the same method, which are contained in the solid-lined triangles. If these correlations are higher than correlations between different constructs measured by different methods, method bias exists.

Finally, we should note that structural equation modeling (SEM) can be used to gather construct-related evidence. SEM can be used to assess convergent and discriminant validity simultaneously (e.g., Pierce, Aguinis & Adams, 2000).

Improving the Size of Validity Coefficients

Similar to the reliability coefficient, there are several factors that affect the magnitude of the validity coefficient. First, to obtain high validity coefficients there must be variability among scores on both the predictor and criterion. If respondents have approximately the same scores, it will be hard for the measure to differentiate among individuals based on the criterion. Also, as described above, in many situations in W&O psychology, the variability in a sample is artificially smaller than that in the population (e.g., personnel selection research; Aguinis & Stone-Romero, 1997; Aguinis & Whitehead, 1997). Range restriction can occur in the predictor when criterion data are available only for those who are hired. Low scorers on the predictor are not hired and thus are not represented in the sample. Likewise, restriction in the criterion may occur as a consequence of terminations, turnover, or transfers that occur before data on the criterion are gathered. Note that when a sample is affected by range restriction in the predictor, the criterion, or both, there are formulae and computer programs available to determine what the validity coefficient would be in the absence of range restriction (Johnson & Ree, 1994).

Second, validity can be enhanced if the influence of factors unrelated to scores on the criterion is minimized. Criterion contamination occurs when factors that are unrelated to the criterion affect scores on the criterion and, consequently, lower validity. For example, an organization uses a general cognitive ability measure to predict job performance. However, if factors such as availability of resources, quality of equipment, or supervisory liking unduly influence supervisory ratings of performance, the validity of the cognitive ability measure will decrease. We are no longer just measuring cognitive ability but, in addition to cognitive ability, we are assessing differences in resources, equipment, and likeability.

Third, validity estimated using the correlation coefficient depends on the relationship between the measure and a criterion being linear. When the relationship is linear, the predictor can accurately predict both high and low scores. If this statistical assumption is violated (e.g., the relationship between the predictor and criterion is curvilinear), the validity coefficient is underestimated.

Finally, if the relationship between the measure and a criterion differs for various groups (e.g., men vs. women), the measure is not similarly valid for these groups (Aguinis, 1995; Aguinis & Stone-Romero, 1997). Thus, prediction of outcomes based on a measure's scores will differ depending on group membership (Aguinis & Pierce, 1998b; Aguinis, Petersen & Pierce, 1999). Consequently, the overall predictive accuracy of a measure will be diminished (Aguinis & Pierce, 1998a; Aguinis, Pierce & Stone-Romero, 1994).

In sum, this section reviewed the process of measure development. We discussed the determination of the purpose of measurement, the definition of the attribute to be measured, the measure development plan, writing items, conducting a pilot study and item analysis, selecting items, establishing norms, and the assessment of reliability and validity. Next, we discuss our views on recent and future trends in the field of measurement in W&O psychology.

RECENT AND FUTURE TRENDS IN MEASUREMENT IN W&O PSYCHOLOGY

This section of the chapter is devoted to a selective set of issues that constitute what in our view are recent and future trends in the field of measurement in W&O psychology. Admittedly, due to space limitations, the following is only a subset of issues that we could describe. However, we hope that discussing these issues will provide an appreciation for what we believe are some important changes affecting the field. First, we discuss issues pertaining to levels of analysis. Specifically, we describe basic concepts regarding levels of analysis, different relationships at different levels, and measurement issues and aggregation. Second, we discuss the impact of technology on measurement. Third, we provide a brief overview of issues regarding cross-cultural measurement transferability. Fourth, we discuss legal and social issues in measurement. Finally, we describe the proliferation of measurement worldwide.

Levels of Analysis and Measurement

As noted above, the first step in the measurement process is to determine the purpose of our measurement. For example, is our purpose to draw conclusions about individuals in a particular organizational setting, individuals in general, groups, or organizations as a whole? Consideration of these different hierarchical 'levels' is important for developing appropriate measures and drawing appropriate conclusions.

For many years, researchers in W&O psychology have been conducting research and developing techniques for recruitment, selection, training and compensation of employees, for dealing effectively with unions, for enhancing productivity and job

satisfaction, for reducing turnover, and so forth. Hundreds of studies have been directed at examining and refining these practices resulting in numerous recommendations for the most effective means of dealing with human resources in organizations. That is, based on the results of these studies, researchers and authors have assumed that organizations will be more effective if we follow practices that are deemed technically superior regarding the management of individuals. Yet, this may not be the case (Ostroff & Bowen, 2000). It is inappropriate to assume that what applies when we study individual differences in organizations also applies to entire groups, divisions, organizational systems, industries or even countries (Klein, Dansereau & Hall, 1994; Ostroff, 1993). Thus, in designing measurement systems, we must attend to 'levels of analysis issues'.

Traditionally, W&O psychologists have focused primarily on the individual level of analysis. Much of our research has been conducted by gathering data from individuals, typically within a single organization, and examining relationships with individual-level performance, behaviors, and attitudes. This focus on individual differences is important and useful provided we only draw conclusions about individuals and do not assume that these same results would apply to all individuals across organizations, or to groups or organizations as a whole.

Basic Levels Concepts

It has long been recognized that multiple, interdependent levels in organizations exist and that understanding the interrelations within and between levels is critical to understanding organizations and organizational behavior (e.g., House, Rousseau & Thomas-Hunt, 1995; Roberts, Hulin & Rousseau, 1978). Individuals comprise groups, groups comprise organizations, organizations comprise industries or markets, and so forth. Interdependencies exist among these levels as, for example, individuals interact with others in their group, groups within the organization interact with other groups, and organizations interact with other organizations. For the purposes here, we will focus primarily on individuals, groups and organizations to illustrate our points, but these issues are also relevant to dyads (e.g., supervisor and subordinate pairs), industries, markets, countries, and other relevant groupings.

Single-level studies are common in organizational research. For example, an individual-level study might be conducted to examine the relationship between employees' perceived job autonomy and their job performance, and an organizational-level study might be conducted to examine the relationship between technology and productivity of organizations. In cross-level studies, a higher-level and a lower-level construct are examined simultaneously. For example, a cross-level study might investigate the impact of organizational climate (an organizational construct) on individual-level satisfaction and behavior. In multi-level studies, two or more levels are examined simultaneously. Cross-level and multi-level examinations fall under the rubric of the meso paradigm (House et al., 1995) because they pertain to the study of at least two levels simultaneously.

Problems are encountered when the level of theory, measurement, statistical analysis, and interpretation are not consistent (Dansereau, Alluto & Yammarino, 1984; Rousseau, 1985). The *level of theory* is the 'target' level or the level that we aim to explain (e.g., individuals, groups, or organizations). The *level of measurement* refers to the source of the data we gather (e.g., survey of individuals, supervisory measure of group performance). The *level of statistical analysis* pertains to the level at which we analyze our data during statistical analyses (e.g., if we gather data from individuals, but then average those data to form aggregate scores for each group during our analyses, then our level of statistical analysis is the group). Finally, the *level of interpretation* refers to the level at which we draw conclusions. For example, if we measure group morale and analyze our data at the group level by using aggregate scores for the group, then we can only draw conclusions about groups. Drawing inferences from our data about any other level, such as how individuals or organizations respond, is inappropriate. Attributing results to a different level than the one from which theory and corresponding analytical techniques were drawn results in the fallacy of the wrong level (Roberts et al., 1978).

Different Relationships at Different Levels

We noted above that different relationships among variables exist at different levels of analysis (individual, group, or organization). How can this be explained? First, consider Figure 2.3. Each oval represents an organization, each small dot represents an individual, and each square represents the average score among individuals within the organization. The solid line represents the correlation across all individuals, the dashed line represents the correlation among the aggregated (average) scores, and the solid line within an oval represents the correlation among individuals within an organization.

In the left panel of the figure we see that if we measured individuals within an organization we would find very little relationship between the two constructs. However, if we measured individuals across organizations we would find a positive relationship, and if we measured aggregated (averaged) scores at the organizational level we would find a strong positive relationship between the two constructs. How could this happen? Suppose, for example, that we are examining the relationship between satisfaction and performance. It may be difficult to

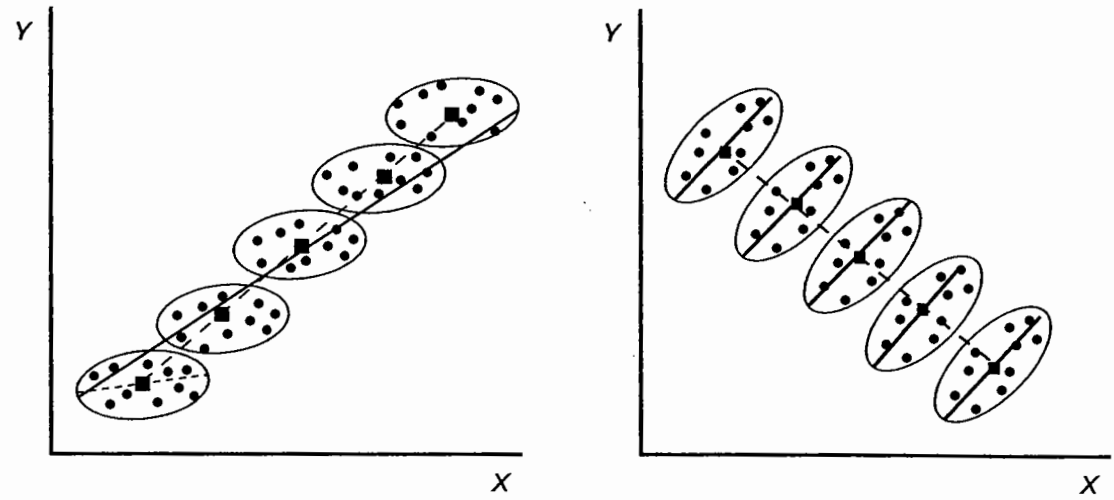

Figure 2.3 *Correlations of different magnitudes at different levels (left panel) and correlations in opposite directions at different levels (right panel)*

predict the relationship between any one individual's satisfaction and his/her performance within an organization. Lower performance is only one possible response to dissatisfaction. A dissatisfied employee may, for example, file a grievance, sabotage the workplace, ask for a transfer and so forth. In contrast, a satisfied employee could work harder, engage in more citizenship behaviors, improve skills or make suggestions for improvements in work processes. While the relationship between individual-level satisfaction and performance may be weak, collective organizational effects can be much stronger due to the cumulative interactions among employees and the cumulative impacts of the behaviors and responses of satisfied or dissatisfied employees overall (Ostroff, 1992).

A different scenario is presented in the right panel of Figure 2.3. Here, the relationship when we study individuals within a single organization is negative, but the relationship at the organizational level (among aggregated mean scores) is positive. How can a case like this be explained? Suppose we are examining the relationship between cognitive ability (X) and performance (Y). It may be that those organizations who rely primarily on cognitive ability tests have more intelligent employees, but they may be missing other critical employee attributes such as interpersonal skills, conscientiousness, or citizenship behaviors, thereby causing lower productivity for the organization.

There are many different configurations that can emerge when examining relationships across levels. The examples here serve to illustrate how different relationships can occur at different levels and the importance of considering one's theoretical interests and aligning them with one's measurement. These notions are explained in more detail in the following section.

Measurement Issues and Aggregation

The level of measurement refers to the level at which data are collected. Individual-level constructs should be collected at the individual level. Higher-level constructs (e.g., group, organization) may be assessed by gathering individual level data and aggregating individual scores to represent the higher level or by collecting a global measure for that level. For many constructs (e.g., group or organizational performance) a global index for the higher-level construct is preferable (e.g., objective measures, expert sources to provide a rating for each group). Frequently, researchers do not have a global index of the higher-level constructs of interest. Further, for some constructs such as organizational climate or group norms, which are based on shared perceptions of members, it is appropriate to gather data from the individuals within the unit and aggregate them to represent the higher level.

One question that arises when relying on aggregated (or averaged) data from individuals to represent the higher-level construct pertains to the extent of 'agreement' among individuals in the focal unit on the construct. While the focus of some debate, it has been generally assumed that individuals within the same focal unit should have relatively similar scores. Individual-level measures should not be aggregated to represent a higher-level construct unless some degree of within-unit agreement can be demonstrated. This is particularly critical for constructs that rely on notions of shared perceptions such as organizational climate, group norms, and cohesiveness (Kozlowski & Klein, 2000).

It is highly unlikely that the variation in responses among individuals within a unit will be close to zero. One issue that arises is how to treat the variance within a unit. One argument is that the aggregate (mean) response per group or organization

is a more accurate representation of the organizational characteristics (Glick, 1985; James, 1982). Intragroup or intraorganizational variance in responses is viewed as a source of inaccuracy or random error. Individual deviations from the mean of their focal unit are not of substantive interest, except in terms of measurement accuracy. Aggregation results in a more stable assessment of the constructs. For example, if one is interested only in examining whether there is a relationship between training of employees and organizational productivity at the organizational level, then issues of individual variation may be irrelevant.

Alternatively, a researcher could view individual variation within a group as partly a result of random error, but also as a reflection of systematic variance or real individual differences. Here, a researcher might examine whether there is a relationship between the constructs among individuals and also among organizations, and whether the relationship at the organizational level is stronger or weaker than the one at the individual level (Ostroff, 1993).

It is likely that some, but not all, of the individual deviation from the mean score of their group is random error. The individual's score is the group true score (mean score for the group) plus some systematic individual variation from the mean plus some random measurement error (cf. Equation 2.2). Given these assumptions, the correlation among the mean scores will differ from the correlation among individuals. This is because measurement error attenuates or reduces the correlation among individuals. However, the correlation among the mean scores is not affected by individual-level random measurement error because the random errors essentially 'average out' in the aggregated score. Hence, before comparing a correlation at the individual level to a higher-level correlation among aggregated scores, it is important to first correct the individual-level score for random measurement error (Ostroff, 1993). Once this has been accomplished, the magnitude of the individual correlation and group or organizational level can be compared. If they are similar, the same processes and relationships exist among individuals as they do for groups (or organizations); if they differ, then different processes are operating at the different levels and more investigation is needed to determine the cause of the differences (Ostroff, 1993).

Finally, there are additional issues that must be considered in constructing items from a multi-level perspective. Clearly, if one is interested in individuals, then the referent for the items should be the individual. However, if one is interested in groups for example, the referent in the items might be better focused on the group. For example, rather than phrase an item as 'I think...' or 'My work...', the item could be phrased as 'Members of this group think...' or 'Our work...'. This referent-shift (Chan, 1998) may result in greater within-group agreement on the construct (Kozlowksi & Klein, 2000) and is preferable if the unit of theory is the group or a higher level.

IMPACT OF TECHNOLOGY ON MEASUREMENT

Improvements in technology are leading to many advances in the area of measurement. Specifically, computers and the Internet have produced new methods for assessing attributes. Measures can be administered online and computers can be used to instantly score measures, store the results, and interpret the meaning of scores through computer-generated reports. Although there are costs associated with hardware and software, the benefits of computerized measurement may outweigh the initial investment. Computers provide standardized and easy administration, quick scoring procedures, efficient storage of results, and less error and chance of cheating. Further, computer unfamiliarity and anxiety are decreasing as the technology becomes more available (Nunnally & Bernstein, 1994).

Computers can administer attribute measures in the exact format as paper and pencil versions. Thus, the computer acts as an electronic page turner. However, an additional advantage of computer technology is that measures of cognitive ability and knowledge can be administered as computerized adaptive tests (CATs). Unlike conventional paper and pencil measures, which require high-ability respondents to waste time answering a number of easy items and low-ability respondents to become frustrated answering many difficult items, CATs are tailored to the ability level of each respondent. CATs use item response theory (discussed earlier in the chapter) to determine the difficulty and discriminability of items. Using the information derived from item response theory, CATs begin with an item of moderate difficulty and the next item administered is determined by a respondent's answer. If a respondent answers correctly, a more difficult item is selected while an incorrect response results in the selection of a less difficult item. The test continuously estimates each respondent's ability level, chooses items appropriate for that level, and uses responses to items to revise the ability estimate (Weiss & Davinson, 1981). This process continues until a certain number of items is administered or the estimate of ability stops changing. In sum, computerized adaptive testing is an efficient method because it measures ability with fewer items (Drasgow & Hulin, 1991) and, thus, can be administered in less time.

Measures can also be administrated via the Internet. Attribute measures can be e-mailed to a targeted population (e.g., attitude measures e-mailed to current employees) or they can be posted on a web page with either open access to the page

(e.g., measure of customer satisfaction) or restricted access using a password (e.g., cognitive ability measure administered to job applicants passing initial screening). Posting measures on a web site and collecting responses through the Internet can be less expensive and time consuming than conventional paper and pencil measures (Schmidt, 1997). Few, if any, proctors are needed; data entry and errors associated with it are eliminated; many individuals from various locations can respond at their leisure; and, unlike e-mailed measures, confidentiality is guaranteed. Unfortunately, when open access to a measure administered through a web page exists, the Internet may not attract a representative sample. The Internet may lead to samples overrepresenting males and professionals, and higher educated and more computer literate individuals (Nicholson, White & Duncan, 1998; Stanton, 1998).

In spite of the advantages of using computers to administer measures, the following caveat is in order. Although converting a conventional paper and pencil measure to a computerized version may allow for quicker and more efficient collection of information, it may also change the meaning of scores. Stated differently, respondents may not obtain similar scores on the paper and pencil and computerized versions of a measure due to the format in which the measure is presented. Mazzeo and Harvey (1988) asserted that differences between the two formats may depend on whether the measure is speeded, contains graphics, has items requiring passages to be read that do not fit on the same screen as the items, and whether respondents can omit or return to items. However, in a meta-analysis comparing paper and pencil to computerized versions of cognitive ability measures, Mead and Drasgow (1993) found that the two methods were equivalent for power or computerized adaptive tests of cognitive ability, but not for speeded measures. Further, King and Miles (1995) found that computerized and paper and pencil versions of attitude and personality measures were equivalent. Although these results are promising, others have found that some personality measures may not be measuring the intended attribute (e.g., Davis & Cowles, 1989) because respondents may have a stronger tendency to fake good (i.e., present themselves in a favorable light) on computerized versions of the measures. More recently, Richman, Kiesler, Weisband & Drasgow (1999) conducted a meta-analytic review and ascertained that, overall, computer-administered measures are not more adversely affected by social desirability distortion than paper and pencil measures.

While technology has advanced the field of measurement, there are still several unresolved issues. First, measures that are translated from a paper and pencil to a computerized version may not be equivalent, that is, they may not be measuring the same attribute. Second, measures that are administered through the Internet with open access may not result in representative samples. Finally, although the availability of computers and the Internet has increased and individuals have become more competent at using these technologies, their use is still not widespread and there are marked differences across countries. Administering measures through these technologies may exclude certain populations (e.g., lower social-economic statuses) and induce anxiety in those who are not familiar with this medium. However, computerized measurement is growing worldwide as evidenced by recent international publications on the topic (e.g., *Applied Psychology: An International Review*, vol. 36, issues 3–4, as cited in Murphy and Davidshofer, 1998).

CROSS-CULTURAL MEASUREMENT TRANSFERABILITY

As interest in measurement increases worldwide, the questions of the transferability of a measure designed in one culture to another and the feasibility of comparing different cultures on the same attribute become important (Cheung & Rensvold, 1999). Concern has been expressed about transferring measures from one culture to another without modifying them to account for cultural differences (Hofstede, 1993). Thus, a measure developed in one type of culture (e.g., individualistic) may not be applicable to a different culture (e.g., collectivistic). Many factors can affect the validity of a measure used in different cultures. Different cultural beliefs, political structures, languages, economies, technologies, and acceptability of and familiarity with measures, may influence the effectiveness of measures. Thus, it is important to cross-validate measures developed in one culture before using them in another culture to ensure that decisions based on measurement are sound. Further, additional explanation and instructions, practice items, and proctor training may be required for cultures not accustomed to particular types of measurement.

The first step in transferring a measure to another culture is establishing translation equivalence. Blind back-translation assesses the equivalence of the wording of a measure that has been translated into a different language (Brislin, Lonner & Thorndike, 1973). The process begins with an individual translating the measure from the original language to another. Next, a second individual, who has not seen the original measure, translates it back to the original language. Finally, the second version of the measure is compared with the original and discrepancies are discussed and resolved.

Unfortunately, translation equivalence does not ensure transferability of the measure to another culture. Stated differently, the measure must also have conceptual equivalence, which is when the

attribute being measured has similar meaning across cultures (Brett, Tinsley, Janssens, Barsness & Lytle, 1997). Measures must produce the same conceptual frame of reference in different cultures, which means different cultures are defining the attribute in the same way (Riordan & Vandenberg, 1994). Some items on a measure or even the attribute in general may have different meanings in different cultures. For instance, a measure of initiative asking questions about individual contributions may be interpreted as boastful or arrogant in a collectivistic culture instead of as a sign of initiative. Further, respondents must interpret response options on the measure similarly (Riordan & Vandenberg, 1994). For example, the response option of 'neither disagree nor agree' may be interpreted as indifference in one culture and as slight agreement in another. In sum, before measures developed in one culture can be used in another, translation and conceptual equivalence must be established or the measure cannot be used to make accurate decisions.

EMERGING LEGAL AND SOCIAL ISSUES IN MEASUREMENT

In the United States (US) measurement is strongly influenced by employment law (for more detail, see AERA, APA & NCME, 1999; SIOP, 1987). Various laws require that measures used in work settings do not discriminate against applicants or current employees on the basis of, for example, race, color, sex, religion, national origin, age, and disability status. If measures do discriminate against a protected group, it must be demonstrated that the measure is related to job performance and that decisions made based on the scores are valid. Thus, measurement in the US focuses on establishing measures that are nondiscriminatory against protected groups and valid for making decisions about individuals.

The influence of the legal system on measurement is increasingly becoming a global phenomenon, as evidenced by the proposal or enactment of similar laws in other countries. For instance, South Africa recently implemented the Employment Equity Act (EEA) of 1998. EEA provides equal opportunity and fair treatment in employment by eliminating unfair discrimination. More importantly, EEA mandates, among other things, that psychological measures used in employment settings be prohibited unless they are reliable and valid. Of course, the passage of equal opportunity laws does not necessarily mean that these laws are strictly enforced. In the US, a government office (i.e., Equal Employment Opportunity Commission) is responsible for such enforcement; however, similar government offices are not common in other countries. Nevertheless, as equal employment opportunity laws are proposed and passed in other countries, measurement adhering to the principles and processes discussed in this chapter will become essential.

GLOBALIZATION OF MEASUREMENT

To support the aforementioned claim regarding the increasing importance of measurement worldwide, we reviewed all articles written by authors with affiliations outside of the US in *Educational and Psychological Measurement* and *Applied Psychological Measurement* from January 1995 to December 1999. This is an admittedly selective review, particularly in light of the fact that these journals are published in English. Consequently, non-English speaking W&O psychologists may not be able, or even wish, to submit their work for publication consideration in these outlets. Nevertheless, results of this selective 5-year review suggest that measure development is increasing in importance. More specifically, many studies described the construction and validation of a variety of measures (e.g., Bessant, 1997; Chang, 1996; Koustelios & Bagiatis, 1997). Moreover, there were numerous studies examining the reliability and validity of existing measures (e.g., Byrne, Baron & Balev, 1998; Cheung, 1996; Mateo & Fernandez, 1995; Tharenou & Terry, 1998). In addition, the goal of many of these studies was to validate a measure cross-nationally.

An additional finding of our selective review is that many of the topics discussed in this chapter are currently being studied in several countries. Computerized adaptive testing is a popular topic, especially in the Netherlands (e.g., Eggen, 1999; Meijer & Nering, 1999; van der Linden, 1998). Another popular topic is Item Response Theory with researchers in the Netherlands, Belgium, Canada, Spain, and Australia exploring this issue (e.g., Andrich, 1995; Janssen & De Boeck, 1997; Maranon, Barbero-Garcia & Costas, 1997; Sijtsma & Verweij, 1999; Zumbo, Pope, Watson & Hubley, 1997). Other topics investigated outside of the US include reliability and validity (e.g., Raykov, 1997), measurement equivalence (e.g., Rensvold & Cheung, 1998; Sukigara, 1996), item analysis (e.g., El-Korashy, 1995), generalizability theory (e.g., Chang, 1997), and the multitrait–multimethod matrix (e.g., Massey, 1997), among others.

In sum, this review is encouraging because it demonstrates that measurement and the issues we have discussed in this chapter are growing in importance worldwide.

In closing, we set two ambitious goals for this chapter. First, our goal was to discuss basic issues in measurement. Second, we also wanted to go beyond basic concepts and discuss a selective set of present and future trends in the field of measurement in W&O psychology. As noted throughout the chapter, sound measurement is a *sine qua non* condition for the science and practice of W&O

psychology. Changes in technology and the legal environment worldwide suggest several challenges as well as the globalization of the field of measurement in W&O psychology. We certainly hope W&O psychologists around the world will continue to appreciate the criticality of sound measurement in their science as well as their practice.

ACKNOWLEDGEMENTS

We thank Charles A. Pierce (Montana State University) for helpful comments on previous drafts. Portions of the research reported herein were conducted while Herman Aguinis was on sabbatical leave from the University of Colorado at Denver and holding visiting appointments at China Agricultural University-International College of Beijing (People's Republic of China), City University of Hong Kong (People's Republic of China), Nanyang Technological University (Singapore), University of Science Malaysia (Penang, Malaysia), and University of Santiago de Compostela (Spain). This research was supported, in part, by grants from the Graduate School of Business Administration (University of Colorado at Denver) and the Institute for International Business (University of Colorado at Denver) to Herman Aguinis.

REFERENCES

Aguinis, H. (1993). Action research and scientific method: Presumed discrepancies and actual similarities. *Journal of Applied Behavioral Science, 29*, 416–431.

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management, 21*, 1141–1158.

Aguinis, H. (forthcoming). Estimation of sampling variance of correlations in meta-analysis. *Personnel Psychology*.

Aguinis, H., & Adams, S.K.R. (1998). Social-role versus structural models of gender and influence use in organizations: A strong inference approach. *Group and Organization Management, 23*, 414–446.

Aguinis, H., Cortina, J.M., & Goldberg, E. (1998). A new procedure for computing equivalence bands in personnel selection. *Human Performance, 11*, 351–365.

Aguinis, H., Cortina, J.M., & Goldberg, E. (2000). A clarifying note on differences between the W.F. Cascio, J. Outzz, S. Zedeck, & I.L. Goldstein (1991) and H. Aguinis, J.M. Cortina, and E. Goldberg (1998) banding procedures. *Human Performance, 13*, 199–204.

Aguinis, H., & Henle, C.A. (forthcoming). Effects of nonverbal behavior on perceptions of a female employee's power bases. *Journal of Social Psychology*.

Aguinis, H., Nesler, M.S., Quigley, B.M., Lee, S., & Tedeschi, J.T. (1996). Power bases of faculty supervisors and educational outcomes for graduate students. *Journal of Higher Education, 67*, 267–297.

Aguinis, H., Petersen, S.A., & Pierce, C.A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods, 2*, 315–339.

Aguinis, H., & Pierce, C.A. (1998a). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods, 1*, 296–314.

Aguinis, H., & Pierce, C.A. (1998b). Statistical power computations for detecting dichotomous moderator variables with moderated multiple regression. *Educational and Psychological Measurement, 58*, 668–676.

Aguinis, H., Pierce, C.A., & Stone-Romero, E.F. (1994). Estimating the power to detect dichotomous moderators with moderated multiple regression. *Educational and Psychological Measurement, 54*, 690–692.

Aguinis, H., Simonsen, M.M., & Pierce, C.A. (1998). Effects of nonverbal behavior on perceptions of power bases. *Journal of Social Psychology, 138*, 455–469.

Aguinis, H., & Stone-Romero, E.F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192–206.

Aguinis, H., & Whitehead, R. (1997). Sampling variance in the correlation coefficient under indirect range restriction: Implications for validity generalization. *Journal of Applied Psychology, 82*, 528–538.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19*, 101–119.

Angoff, W.H. (1971). Norms, scales, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Berk, R.A. (1984). *A Guide to Criterion-referenced Test Construction*. Baltimore: Johns Hopkins University Press.

Bessant, K.C. (1997). The development and validation of scores on the mathematics information processing scale (MIPS). *Educational and Psychological Measurement, 57*, 841–857.

Binning, J.F., & Barrett, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential base. *Journal of Applied Psychology, 74*, 478–494.

Brennan, R.L. (1992). *Elements of Generalizability Theory* (2nd ed.). Iowa City, IA: American College Testing Program.

Brett, J.M., Tinsley, C.H., Janssens, M., Barsness, Z.I., & Lytle, A.L. (1997). New approaches to the study of culture in industrial/organizational psychology. In P.C. Earley and M. Erez (Eds.), *New Perspectives on International Industrial/Organizational Psychology*. (pp. 75–129) San Francisco: New Lexington Press.

Brislin, R.W., Lonner, W., & Thorndike, R.M. (1973). *Cross-cultural Research Methods*. New York: Wiley.

Byrne, B.M., Baron, P., & Balev, J. (1998). The Beck Depression Inventory: A cross-validated test of second-order factorial structure for Bulgarian adolescents. *Educational and Psychological Measurement, 58,* 241–251.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83,* 234–246.

Chang, L. (1996). Quantitative Attitudes Questionnaire: Instrument development and validation. *Educational and Psychological Measurement, 56,* 1037–1042.

Chang, L. (1997). Dependability of anchoring labels of Likert-type scales. *Educational and Psychological Measurement, 57,* 800–807.

Cheung, S. (1996). Reliability and factor structure of the Chinese version of the Depression Self-Rating Scale. *Educational and Psychological Measurement, 56,* 142–154.

Cheung, G.W., & Rensvold, R.B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1–27.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 10,* 37–46.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H.I. Braun (Eds.), *Test Validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Dansereau, F., Alluto, J., & Yammarino, F.J. (1984). *Theory Testing in Organizational Behavior: The Variant Approach.* Englewood Cliffs, NJ: Prentice Hall.

Davis, C., & Cowles, M. (1989). Automated psychological testing: Method of administration, need for approval, and measures of anxiety, *Educational and Psychological Measurement, 49,* 311–337.

Drasgow, F., & Hulin, C.L. (1991). Item response theory. In M. Dunnette & L. Hough (Eds.), *Handbook of Industrial and Organizational Psychology.* Vol. 1, 2nd ed. (pp. 577–636) Palo Alto, CA: Consulting Psychologists Press.

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249–261.

El-Korashy, A. (1995). Applying the Rasch model to the selection of items for a mental ability test. *Educational and Psychological Measurement, 55,* 753–763.

Flaugher, R. (1990). Item pools. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (pp. 41–63). Hillsdale, NJ: Lawrence Erlbaum.

Glick, W.H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review, 10,* 601–616.

Guion, R.M. (1977). Content validity – the source of my discontent. *Applied Psychological Measurement, 1,* 1–10.

Guion, R.M. (1998). *Assessment, Measurement, and Prediction for Personnel Decisions.* Mahwah, NJ: Lawrence Erlbaum.

Hofstede, G. (1993). Cultural constraints in management theories. *Academy of Management Executive, 7,* 81–94.

House, R., Rousseau, D.M., & Thomas-Hunt, M. (1995). The meso paradigm: A framework for the integration of micro and macro organizational behavior. *Research in Organizational Behavior, 17,* 71–114.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory: Application to Psychological Measurement.* Homewood, IL: Dow Jones-Irwin.

James, L.R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 57,* 219–229.

James, L.R., Demaree, R.G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 1–98.

James, L.R., Demaree, R.G., & Wolf, G. (1993). r_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78,* 306–309.

Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21,* 37–50.

Johnson, J.T., & Ree, M.J. (1994). RANGEJ: A Pascal program to compute the multivariate correction for range restriction. *Educational and Psychological Measurement, 54,* 693–695.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Kendall, M.G. (1948). *Rank Correlation Methods.* London: Griffin.

King, W.C., & Miles, E.W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Psychological Science, 6,* 203–211.

Klein, K.J., Dansereau, F., & Hall, R.J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19,* 195–229.

Koustelios, A.D., & Bagiatis, K. (1997). The Employee Satisfaction Inventory (ESI): Development of a scale to measure satisfaction of Greek employees. *Educational and Psychological Measurement, 57,* 469–476.

Kozlowksi, S.W.J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77,* 161–167.

Kozlowski, S.W.J., & Klein, K.J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In

K.J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel Theory, Research, and Methods in Organizations* (pp. 512–553). San Francisco: Jossey-Bass.

Kraiger, K., & Aguinis, H. (2001). Training effectiveness: Assessing training needs, motivation, and accomplishments. In M. London (Ed.), *How People Evaluate Others in Organizations* (pp. 203–220). Mahwah, NJ: Lawrence Erlbaum.

Kuder, G., & Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lahey, M., Downey, R.G., & Saal, F.E. (1983). Intraclass correlations: There's more there than meets the eye. *Psychological Bulletin, 93*, 586–595.

Landy, F. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192.

Lawlis, G.F. & Lu, E. (1972). Judgement of counseling process: Reliability, agreement, and error. *Psychological Bulletin, 78*, 17–20.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Lord, F.M., & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

Maranon, P.P., Barbero-Garcia, M.I., & Costas, C.L. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement, 57*, 559–568.

Massey, A.J. (1997). Multitrait–multimethod/multiform evidence for the validity of reporting units in national assessments in science at age 14 in England and Wales. *Educational and Psychological Measurement, 57*, 108–117.

Mateo, M.A., & Fernandez, J. (1995). Evaluation of the setting in which university faculties carry out their teaching and research functions: The ASEQ. *Educational and Psychological Measurement, 55*, 329–334.

Mazzeo, J., & Harvey, A.L. (1988). *The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests: A Review of the Literature* (College Board Report No. 88–8). Princeton, NJ: Educational Testing Service.

Mead, A.D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458.

Meijer, R.R., & Nering, M.L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187–194.

Murphy, K.R., & Davidshofer, C.O. (1998). *Psychological Testing* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Nesler, M.S., Aguinis, H., Quigley, B.M., Lee, S., & Tedeschi, J.T. (1999). The development and validation of a scale measuring global social power based on French and Raven's (1959) power taxonomy. *Journal of Applied Social Psychology, 29*, 750–771.

Nicholson, T., White, J., & Duncan, D. (1998). Drugnet: A pilot study of adult recreational drug use via the WWW. *Substance Abuse, 19*, 109–121.

Nunnally, J.C. (1967). *Psychometric Theory*. New York: McGraw-Hill.

Nunnally, J.C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory*. 3rd edn. New York: McGraw-Hill.

Ostroff, C. (1992). The relationship between satisfaction, attitudes, and performance: An organizational level analysis. *Journal of Applied Psychology, 77*, 963–974.

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology, 78*, 569–582.

Ostroff, C., & Bowen, D.E. (2000). Moving HR to a higher level: HR practices and organizational performance. In K.J. Klein and S.W.J. Kozlowski (Eds.), *Multilevel Theory, Research and Methods in Organizations* (pp. 211–266). San Francisco: Jossey-Bass.

Pedhazur, E.J., & Pedhazur Schmelkin, L. (1991). *Measurement, Design, and Analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Pierce, C.A., Aguinis, H., & Adams, S.K.R. (2000). Effects of a dissolved workplace romance and rater characteristics on responses to a sexual harassment accusation. *Academy of Management Journal, 43*, 869–880.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184.

Rensvold, R.B., & Cheung, G.W. (1998). Testing measurement model for factorial invariance: A systematic approach. *Educational and Psychological Measurement, 58*, 1017–1034.

Richman, W.L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*, 754–775.

Riordan, C.M., & Vandenberg, R.J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20*, 643–671.

Roberts, K.H., Hulin, C.L., & Rousseau, D.M. (1978). *Developing an Interdisciplinary Science of Organizations*. San Francisco: Jossey-Bass.

Rousseau, D.M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior, 7*, 1–38.

Schmidt, W.C. (1997). World-wide web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers, 29*, 274–279.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Sijtsma, K., & Verweij, A.C. (1999). Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement, 23*, 55–68.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the Validation and Use of Personnel Selection Procedures* (3rd ed.). College Park, MD: Author.

Stanton, J.M. (1998). An empirical assessment of data collection using the Internet. *Personnel Psychology, 51*, 709–725.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 1–49). New York: Wiley.

Stevens, S.S. (1968). Measurement, statistics, and the schemapiric view. *Science, 161*, 849–856.

Sukigara, M. (1996). Equivalence between computer and booklet administrations of the new Japanese version of the MMPI. *Educational and Psychological Measurement, 56*, 570–584.

Tharenou, P., & Terry, D.J. (1998). Reliability and validity of scores on scales to measure managerial aspirations. *Educational and Psychological Measurement, 58*, 475–492.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385–395.

Thorndike, R.L. (1949). *Personnel Selection*. New York: John Wiley & Sons.

Thorndike, R.M., Cunningham, G.K., Thorndike, R.L., & Hagen, E. (1991). *Measurement and Evaluation in Psychology and Education* (5th ed.). New York: Macmillan – now Palgrove.

Tinsley, H.E., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22*, 358–376.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*, 4–69.

van der Linden, W.J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195–211.

Weiss, D.J., & Davinson, N.L. (1981). Test theory and methods. *Annual Review of Psychology, 32*, 629–658.

Zumbo, B.D., Pope, G.A., Watson, J.E., & Hubley, A.M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement, 57*, 963–969.