Theories, Models, and Standard Systems of Measurement

Marion S. Aftanas
University of Manitoba

Measurement theories in psychology may be classified in terms of whether they begin from a general measurement framework or from a specific area of measurement. Points of contact between theories and different specific measurement areas have been limited by the choice of focus in discussions of general measurement theories and specific theories or models. This presentation outlines a metatheoretical framework that begins with the obvious common factor in all areas, the standard system of measurement. Just as a standard is a commonly accepted unit of measurement, a standard system is a commonly accepted mechanism of measurement for a given area. The concept of a standard system amplifies general definitions of measurement and clarifies metatheoretical statements concerning the requirements for measurement. Differences between measurement situations may be categorized by the type of standard system used and by features of the attribute measured. Identification of different standard systems and elements of the measurement process provides a focus for comparisons between measurement theories and models in different measurement situations.

Index terms: applied measurement models, comparison of measurement models, definition of measurement, measurement process, measurement theory, metatheoretical framework for applied measurement models, standard systems of measurement.

Measurement theories attempt to explain how and under what circumstances it is possible to use numbers to represent information about magnitudes of attributes. Two types of theories can be identified, depending on the generality of the discussion. Some measurement theorists have focused on the general case, such as Campbell (1928), Ellis (1966), Mitchell (1986), Stevens (1951), and Suppes and Zinnes (1963), while others have dealt with specialized areas or aspects of measurement, such as Johnston and Pennypacker (1980), Lord (1953), Torgerson (1958), and Wherry (1982). These latter developments concern specific topics such as behavioral observation, latent trait theory in testing, scaling, and personnel assessment, respectively.

Points of contact between general measurement theory and applied measurement areas have been limited by the restricted discussion in general measurement theory and by the highly specialized nature of the discussions in specific measurement situations. General measurement theories have focused on such issues as the formal requirements for achieving a particular scale type, whereas specialized areas of measurement have been concerned with models related to characteristics of the measurement instrument and/or evaluation of the measurement outcome. What is needed is a broader focus for measurement theory so that the concepts of applied measurement areas can be brought into the context of general measurement theory. This paper shows how focusing on the common factor in all measurement—a generic concept for the mechanism of measurement—coupled with an operationalized analysis of the requirements for mea-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 4, December 1988, pp. 325–338

0146-6216/88/040325-14$1.95

surement that incorporates the concept, can integrate a metatheory of measurement and discussions of applied measurement.

Focusing on the common factor serves several additional functions:

1. It expands current definitions of measurement by showing that an agent is specified in the act of measurement, which provides a more complete definition.
2. It helps establish that measurement should be conceptualized as a two-stage process. The initial stage involves an assessment made by a mechanism, and the second stage converts that assessment into numerical information. Sometimes these two stages are accomplished concurrently so they appear to be part of a single process.
3. It provides a general measurement theory basis for areas of measurement that do not currently have one. Discussions of psychological measurement involving human assessment of behavioral attributes (e.g., Johnston & Pennypacker, 1980; Landy & Farr, 1986; Mitchell, 1979) have not provided a foundation for considering human and test assessment as alternate approaches to deriving measurement outcomes. This presentation attempts to provide such a foundation.
4. It provides a basis for a systematic analysis of types of measurement situations. This serves a heuristic function in helping to identify points of similarity and difference between them, which may lead to an increased interchange of ideas and concepts between areas.

Issues in Measurement Theory

General measurement theories have focused on one or more of three issues: (1) definitions of the act of measurement; (2) statements of the logical or formal requirements in measurement; and (3) determination of the type of scale that may be derived from measurement. The first two refer to the issue of the necessary conditions required for the appropriate use of numbers in measurement and the third to the types of scalar information that may be derived from measurement.

Definitions of Measurement

Stevens (1946, 1951) defined measurement as the ''assignment of numerals to objects or events according to rules'' (1951, p. 1). He used the term *numerals* specifically because he wanted to allow for the possibility for one type of assignment that led to what was termed a nominal scale, in which numerals could be used to identify objects or events but did not imply the order, interval, or zero characteristics of the number system. Torgerson (1958) suggested that the nominal form of assignment was not really measurement at all but a type of ''naming.'' He also proposed that the properties of objects or events were being measured and not the objects themselves. As a consequence of these two disagreements, Torgerson (1958) modified Stevens' definition to ''Measurement of a property . . . involves the assignment of numbers to systems to represent the property'' (p. 14).

A third more elegant and general definition may be derived from the presentation of Suppes and Zinnes (1963). In their terms, measurement involves the mapping of an empirical relational system onto a numerical relational system. The empirical relational system refers to a set of magnitudes of the attribute of interest, and the numerical relational system designates the set of numbers that may be used to represent the magnitudes. Measurement involves the conversion of information about the empirical system into the numerical system.

These definitions are similar because they each mention an attribute to be measured, the numbers that result from the measurement, and the notion that magnitudes of the attribute are somehow assigned to, or mapped onto, different numbers. The definitions do not emphasize or refer specifically to a general mechanism for performing the assignment or mapping.

Statements on the Requirements for Measurement

The most pervasive and influential statement of the requirements for measurement is attributed to Campbell (1928). These requirements have been referred to in one form or another in subsequent discussions of general measurement theories (e.g., Ellis, 1966; Fraser, 1980; Stevens, 1951; Suppes & Zinnes, 1963). Campbell stipulated in the first two "laws" of measurement that the relations identified in the number system must be empirically demonstrable in the property being measured. The first law is referred to as that of order, which permits statements such as $A > B$ to be made. The second is the additivity requirement, so that statements such as $C + D = E$ can be made.

The stipulation of the additivity requirement meant that an empirical operation of addition had to be demonstrated in the magnitudes of the property before the additivity function in the number system could be invoked. This did not seem possible for many attributes. The matter went to court, in a sense, when a committee was set up by the British Association for the Advancement of Science to arbitrate whether Stevens' (1936) sone scale of loudness was measurement. One of the committee members, J. Guild, expressed the view that

> Any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is in fact misleading unless and until a meaning can be given to the concept of addition as applied to sensation. (Committee Final Report, 1940, p. 345)

As has been pointed out by others (e.g., Fraser, 1980), this particular requirement creates difficulty not only for psychological attributes but also for many physical properties, including density or temperature, where additivity cannot be demonstrated.

Types of Scales in Measurement

A variation on the requirements for measurement was introduced by Stevens (1946, 1951). In his 1951 statement he indicated that "A rule for the assignment of numerals (numbers) to aspects of objects or events creates a *scale*" (p. 23). The type of scale that was generated in a given measurement situation was directly related to the characteristics of the number system that could be used. Stevens outlined the basic scale types as being the nominal, ordinal, interval, and ratio. The nominal scale did not incorporate any of the characteristics of the number system, whereas the ratio involved all of them—namely, order, equality of interval, and existence of an origin. Subsequent writers have suggested the possibility of other scale types. For example, Coombs (1953) included the possibility of an ordered-metric scale, and Torgerson (1958) proposed the possibility of an ordinal scale with zero point.

The Concept of a Standard System of Measurement

None of the discussions of general measurement theory emphasizes or even specifically introduces a concept for the mechanism of measurement. At the same time, the applied measurement literature has referred to a number of different terms to identify these mechanisms. Specific concepts such as instrument, test, rater, questionnaire, mechanism, inventory, judge, in addition to the term "scale", have been used to identify measuring devices. Generally these terms have been identified with a specific measurement situation and have not been amenable to generalization across measurement situations. For example, the concept of a scale as used in personality assessment would need to be generalized beyond recognition to apply it to stimulus scaling or to observational assessment of behavior. The introduction of a generic concept would not only identify an important component in measurement theory but would also provide the basis for generalizations across applied measurement situations.

The concept being proposed as the generic term for the mechanism of measurement is a *standard system*. A standard system is defined as any device, mechanism, or discriminative process that may be used to denote and indicate extent of magnitude or differences between magnitudes of a property. *Magnitude* refers to any specific amount of a property or attribute being measured. *Discriminative process* is used in the same sense as when it was first introduced by Thurstone (1927) and with the same intent. It provides for the possibility that human judgment could be used as a basic standard system in the estimation of magnitudes or determination of differences between magnitudes. The term *denote* designates the act of encountering the magnitudes of the property by the standard system. For a measurement to be effected, the standard system must encounter, or somehow come into contact with, the property.

The term standard system was chosen to refer to the generalized measurement mechanism because as a new term it will not be associated with any given area of applied measurement and therefore does not connote or denote specific mechanisms. The term *system* is used in the same general sense suggested by Suppes and Zinnes (1963) in referring to the terms empirical and numerical relational systems. A standard system provides a designated mechanism for converting information about the empirical system into the numerical system. The term *standard* is used in a similar vein to that used in the context of units of measurement. Just as a standard is a commonly accepted unit of measurement, a standard system is a commonly accepted mechanism for the measurement of a given attribute.

An Analysis of Measurement Issues Based on Standard Systems

Definition of Measurement

The concept of a standard system makes possible a definition of measurement which specifies that an agent is involved. It also makes clear that an important element in the measurement enterprise is the identification and/or development of an appropriate standard system for a particular attribute. Measurement may then be defined as use of a standard system to map magnitudes of properties or attributes onto a formal system, such as that of the number system.

Requirements For Measurement

What seems evident from statements on measurement requirements is that unique situations have been used as prototypical examples; as a result, the theory has limited applicability. Campbell (1928) and others, for example, focused on particular properties, such as length and mass, when they specified the requirements for measurement. It is possible to perform the operations stipulated by Campbell logically and empirically for small magnitudes of length. In this case it may be perceptually evident to the human observer that relations between magnitudes could be identified so that the relationships $A > B$ or $C = D$ could be determined for the magnitudes. In the case of most other properties, however, this evidence is not available directly to the human observer and it is therefore necessary to rely on mechanisms to identify order, equality, or an absence of magnitudes. These mechanisms are standard systems for the measurement of the property of interest.

A completely generalizable statement suggests that when *any* standard system is used to establish order or equality among magnitudes, a numerical statement about these magnitudes is justified. In the example used by Campbell (1928), human observation acting as a standard system was used to identify equality of two magnitudes, such as the length of two rods. When a pan balance is used to establish equality of two weights, the balance in conjunction with human observation determines the equality of the weights.

The approach taken here generalizes the statement regarding the determination of empirical equalities and orders to include those determined by standard systems other than those based on direct human observation. Other standard systems can be used to identify order or equality among magnitudes. The generalized statement proposes that when any standard system can be shown to identify order among magnitudes or equality between magnitudes, a numerical statement about these relationships is possible. The critical evaluation of such statements rests in confirmation that the statements of order or equality can be verified. In the case of human observation, the evaluation involves confirmation by other observers.

In the general case, however, verifiability of the assessments consists of two parts: (1) whether other standard systems (of the same or different types) agree; and (2) if agreement is found, whether a relationship between the magnitudes and the indications of these magnitudes provided by the standard system can be established. In the case of direct observation, as in the observation of lengths, these two tests may be evaluated concurrently. If several individuals agreed that $X > Y$ for two magnitudes that were being observed directly, and none disagreed, the researcher would have confidence in assigning a larger number to X.

When direct observation of the attribute is not possible, sequential verification can be performed. First, determine whether different standard systems intended to measure an attribute in fact give the same indication of magnitude. Then, test for the relationship between the magnitude of the property and the indication of the magnitude provided by the standard system in a theoretical and indirect way; this may be determined by a validation process that establishes the standard system as a veridical and meaningful indicator of the attribute.

Scale Types

The term scale has not always been interpreted in a consistent manner. Although Stevens proposed that a scale ''be identified as the rule of assignment [in measurement] . . . the term 'scale' is sometimes used to refer to the measuring instrument, and sometimes even to the standard of measurement'' (Kaplan, 1964, p. 189). The term should convey the notion that the characteristic(s) implied in the use of the numbers reflect something about the information that was obtained concerning magnitudes of attributes. If this is the case, and given the ambiguity involved in using the term scale, it might be more useful to refer to the type of metric information that is conveyed in the numbers. The metric information may be derived directly or indirectly from the information provided by the standard system.

The Measurement Process

Statements on the requirements for measurement in general measurement theory are theoretical in nature and do not deal specifically with the practical aspects of arriving at a measurement. In contrast, discussions in applied measurement areas have been concerned with the construction of standard systems and determination of measurement accuracy. What is needed is an operationalization of measurement requirements that can be used as a metatheoretical guideline in any applied measurement situation.

The beginning point in any measurement situation is the construction or use of a standard system that may be used to differentiate between the magnitudes of attributes. The end point is the evaluation of whether the numerical values assigned to the magnitudes are justified. The intermediary steps specify how the standard system is used to make an assessment of the attribute and what use of the number system is permitted. These aspects of arriving at a measurement will be referred to as components of the measurement process.

Components of the Measurement Process

First, the process requires the identification of an appropriate standard system. For most physical attributes this implies the creation of some physical system. As Astin (1968) made clear, the development of physical systems is a highly technical matter. For psychological attributes the measurement process involves developing or using standard systems which differ in kind, but not in the complexity involved. As with physical standard systems, the construction of psychological systems such as ability or achievement tests, for example, is a highly technical matter. A categorization of standard systems used in psychology is presented below.

Second, the measurement process requires that the standard system denote the attribute of interest. The standard system must be capable of encountering, interacting with, or identifying the presence, similarity, magnitude of difference, or absence of the property. Different standard systems denote the attribute of interest in different ways. A statement about denotability in any measurement situation would involve protocol or descriptive statements on how the attribute may be dimensionalized, how the standard system encounters the attribute, and how assessments are made.

The term denotability was used by Allport (1955, p. 19) to designate the perceptual encounterability of an object, attribute, or property. In the present context, the term is used to identify the possibility that a property or attribute may be made manifest or evident by or through a standard system, but not be directly and immediately observable. For example, a particle of matter may be denotable only through an electron microscope. Or a distant planet may be denotable only by a radar telescope which translates the information to make it identifiable to the human senses. In psychology much measurement involves denotability through the interaction between the standard system and the attribute. In psychological testing, for example, it is just such interaction between the test and the person that makes possible the assessment of latent attributes.

Denotability issues involve the determination of the extent of scientific agreement on whether a particular standard system is denoting a given property or attribute. This could range from complete agreement to complete disagreement. For cases in which the evidence is clear, as would be the case when a ruler is being used to measure length, agreement might be very high. This occurs because general agreement exists regarding what constitutes the property of length and regarding the appropriateness of using the ruler to measure that property. When self-report standard systems are used to measure personality characteristics, however, agreement tends to be somewhat lower. The term *scientific agreement* suggests that some degree of rigor is needed to identify and stipulate the degree of agreement; the way that it is determined may vary in different measurement situations.

The notion of a standard system denoting an attribute has implications for understandings of measurement that involve the term ''operational definitions.'' Rather than discussing the definition of a construct in terms of the operations necessary for its measurement, the act of measuring an attribute can be referred to in terms of the interaction that takes place between a particular standard system and the attribute. The attribute to be measured is in a sense defined and the assessment programmed into the standard system by the psychologist observing or conceptualizing the attribute. Such a specification is more precise in its implications than the specification of an operational definition because an investigator is not free to specify any definition. There must be scientific evaluation and agreement on the denotability of the attribute, and the accuracy of the assessment, by a particular standard system.

Third, a model must be specified for deriving metric information from the assessment provided by the standard system. The derivation of metric information may be conceptualized as a two-stage sequential process. The first stage consists of the assessment by the standard system. The second involves the conversion of this assessment to numerical information consistent with the appropriate usage of the number system. When the standard system is provided with a calibration in terms of a unit system, metric information is derived simultaneously with the assessment by the standard system. A quantity is obtained by comparing a specific magnitude of the characteristic with the number of units indicated by the standard system. An ability test containing calibrated items is an example of such a direct determination of metric information using a unit of the standard system.

Metric information can be derived indirectly as well by using a model describing the conversion of the information provided by the standard system. Thurstone's (1927) Law of Comparative Judgment and Coombs' (1950) Unidimensional Unfolding technique are examples of such conversions. In addition, comparisons provided by the standard system could be converted to a *z*- or *T*-score system in which scores are compared in terms of the standard normal curve.

The assessment provided by the standard system serves as a basis for representing the magnitudes by the number system. The specific characteristics of the number system that may be employed in that representation are dictated by the type of information provided by the standard system. Coombs (1953), Stevens (1946), and Torgerson (1958), among others, discussed some of the basic metric representations.

Fourth, criteria and procedures must be developed for determining the extent to which the measurement outcome accurately represents and reflects the attribute being measured. These criteria may be derived from the generalized statement of the requirements for measurement. They evaluate whether the indications of a given magnitude provided by the standard system are verifiable or replicable through independent standard systems, evaluating empirically or inferentially whether the different quantities indicated by the standard system accurately reflect different magnitudes of the attribute of interest, identifying the conditions that might modify an appropriate measurement outcome, and establishing the meaningfulness of the measurement outcome. The procedures designed to evaluate whether these criteria have been met may need to be more detailed and elaborate for some attributes and some types of standard systems. The procedures would differ somewhat, for example, depending on the type and degree of denotability of the attribute and the operating characteristics of the standard system.

Types of Standard Systems in Psychology

The importance of the concept of a standard system lies not only in the fact that it highlights an important element of the activity of measurement, but also in that it permits a useful categorization of applied measurement situations. These can potentially be differentiated in terms of the type of standard system that is being used and other distinguishing features of the measurement situation.

An analysis of applied measurement situations suggests that at least three differentiable types of standard systems exist in psychology. The differentiation is based on the descriptive features of the measurement mechanism used at the point of interaction between the standard system and the attribute being assessed. These types are the *elementary*, *devised*, and *dual-process* standard systems.

The Elementary Standard System

The elementary standard system is identified with the use of the human capacity to distinguish between magnitudes of attributes which may, through appropriate calibration and/or conversion processes, provide metric information. In psychology as in other disciplines, there are many attributes that can only be denoted, and magnitudes differentiated, in this way. The term *elementary* refers to this type because the assessment provided is more primitive from an historical or accuracy perspective. The measurement of most physical properties was undoubtedly preceded by human observation and assessment prior to the development of more precise devised standard systems.

The conceptualization of standard systems of measurement that includes the possibility of human assessment provides a basis for integrating these types of assessments within a general measurement theory. Previous discussions of personnel assessment (e.g., Landy & Farr, 1980) and behavioral assessment (e.g., Johnston & Pennypacker, 1980) have not provided a theoretical basis for conceptualizing the human as a measurement mechanism.

The Devised Standard System

Devised standard systems are constructed by psychologists to denote and determine magnitudes of certain characteristics of individuals just as devised standard systems are constructed by physicists to determine magnitudes of physical properties. Devised standard systems include ability tests, achievement tests, and the many physical instruments used in psychology to measure behavioral attributes.

The Dual-Process Standard System

The third type of measurement situation associated with self-report instruments involves two standard system processes. It includes, as one element, a constructed standard system consisting of a series of questions developed by the psychologist. This represents an independent standard system. In addition, the respondent's interaction with the standard system requires a self-assessment which can be considered as a separate standard system assessment. Both of these components of the measurement situation need to be analyzed to determine the meaningfulness of the measurement in these categories.

Seven Categories of Applied Measurement

Although types of standard systems describe basic differentiating features for measurement mechanisms, other structural and/or functional features might usefully be highlighted between different measurement situations. The human standard system, for example, has been used to make assessments of stimuli in scaling studies, to make estimates of observable behavior in studies of infant movement, and to assess latent attributes such as effective performance. Each of these represents a slightly different task for the human standard system.

Devised standard systems can be further differentiated in terms of defining and operating characteristics. Multi-item tests of latent ability and achievement attributes may be contrasted with physical mechanisms used to measure behavioral movement. In addition, the dual-process assessment of personality characteristics may be distinguished from the dual-process assessment of preferences and interests. The former involves a self-assessment of the individual's typical behavior, whereas the latter focuses on a person's evaluation of a given object or situation.

Based on such considerations, seven categories of measurement representing different measurement situations have been identified. These are summarized in Figure 1. The columns to the right identify the different categories of measurement associated with each type of standard system.

Elementary Standard Systems

Category 1: Elementary assessment of stimuli. This category consists of situations in which the human observer estimates the magnitude of a stimulus property or a non-behavioral attribute of another organism. The human standard system can compare the stimuli to each other to determine similarity, difference, or degree of difference as well as identify the magnitude of the stimulus with some response system that has been learned. Several reviews (Dawes, 1972; Guilford, 1954; Torgerson, 1958) outlined these methods and indicated similarities and differences between them. By far the most commonly used methods are variations of the rating scale and the method of magnitude estimation introduced by Stevens (1956).

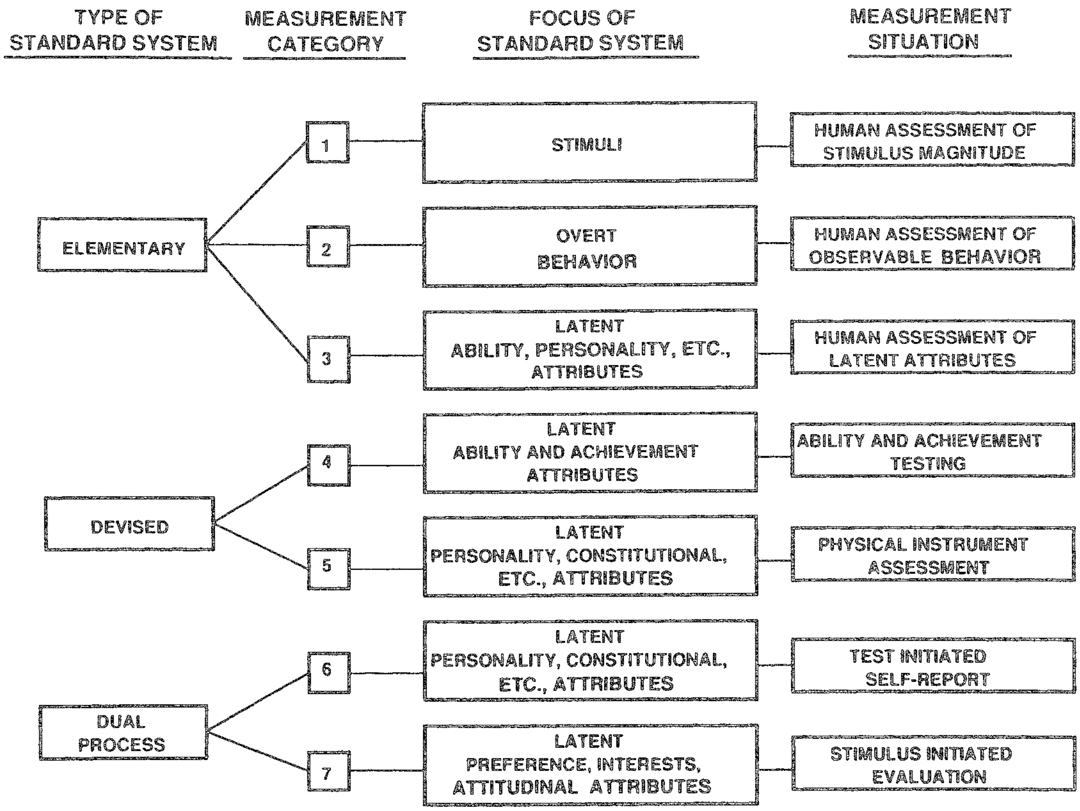
Category 2: Elementary assessment of overt behavior. This type of measurement involves using the elementary standard system to assess observable behavior. The category is associated with the type of assessment undertaken in developmental studies of infants' and children's behavior (e.g., Fassnacht, 1982; Hutt & Hutt, 1970) or in the area of behavioral observation (e.g., Johnston & Pennypacker, 1980). The focus is on directly observable behavior that may be counted or dimensionalized in some way.

Category 3: Elementary assessment of latent attributes. This category of measurement is exemplified by the performance rating and personnel assessment literature (e.g., Landy & Farr, 1980; Saal, Downey, & Lahey, 1980). The focus is on the human standard system, which makes estimates of characteristics by inferences from the behavior of other humans. These standard system estimates of human characteristics usually are retrospective ratings of the behavior, and it is generally assumed that a high rating is associated with ''good'' to ''excellent'' performance and a low rating is associated with ''poor'' performance. The basic information that enters into the evaluation of a characteristic or overall performance is the remembered or described behavior inferred or indicated to be relevant to that characteristic or overall performance.

Devised Standard Systems

Category 4: Devised assessment of ability and achievement. This refers to the use of the familiar multi-item test as a measure of ability and achievement. The test consists of individual items which are intended to measure a given latent attribute. Ability and achievement tests have sometimes been differentiated for measurement-related reasons. Glaser (1963), for example, identified the former as norm-referenced and the latter as criterion-referenced measurement.

Figure 1
Types of Standard Systems and Categories of Measurement in Psychology


TYPE OF STANDARD SYSTEM
MEASUREMENT CATEGORY
FOCUS OF STANDARD SYSTEM
MEASUREMENT SITUATION
ELEMENTARY
DEVISED
DUAL PROCESS
1
2
3
4
5
6
7
STIMULI
OVERT BEHAVIOR
LATENT ABILITY, PERSONALITY, ETC., ATTRIBUTES
LATENT ABILITY AND ACHIEVEMENT ATTRIBUTES
LATENT PERSONALITY, CONSTITUTIONAL, ETC., ATTRIBUTES
LATENT PERSONALITY, CONSTITUTIONAL, ETC., ATTRIBUTES
LATENT PREFERENCE, INTERESTS, ATTITUDINAL ATTRIBUTES
HUMAN ASSESSMENT OF STIMULUS MAGNITUDE
HUMAN ASSESSMENT OF OBSERVABLE BEHAVIOR
HUMAN ASSESSMENT OF LATENT ATTRIBUTES
ABILITY AND ACHIEVEMENT TESTING
PHYSICAL INSTRUMENT ASSESSMENT
TEST INITIATED SELF-REPORT
STIMULUS INITIATED EVALUATION

Category 5: Devised assessment of behavior. This is the measurement situation in which the standard system is a physical instrument which reflects or records the magnitude of some behavior. An actometer used to assess activity level in children (e.g., Eaton, 1983) would be an example of such an instrument, as would the electromyograph and the pupilometer. Instruments of this type are extensively used in experimental studies of behavior.

Dual-Process Standard Systems

Category 6: Dual-process assessment of personality. The prototype of this measurement situation is the questionnaire or self-report personality test. The standard system consists of a set of questions or statements with which the respondent must agree, in a true-false format, or indicate extent of agreement in a Likert-type response format. From the perspective of a standard system analysis, this measurement situation involves two types of estimation procedures. One is associated with the independent standard system consisting of test items developed by the test constructor, and the other is associated with the respondent's self-assessment of the characteristic under consideration in the test item. The ultimate magnitude of direct interest to the psychologist is the extent to which respondents report that they possess the characteristic of interest.

Category 7: Dual-process assessment of pref-

erences and interests. This category of measurement is similar to the previous one in that both involve a form of self-report. It is differentiated because the response is not so much a self-evaluation, as it was in the previous category, as a personal evaluation of the object, situation, or activity being portrayed in the items of the standard system. The responses must be unfolded (Coombs, 1950) or analyzed in other ways (Dawes, 1972; Torgerson, 1958) to uncover the respondent's position on the continuum defined by the items of the standard system.

Categories of Measurement and Models of the Measurement Process

The designation of different categories of measurement and elements of the measurement process identifies a basis for a metatheoretical framework for organizing and evaluating specific measurement models. Figure 2 outlines such a framework in schematic form. The categories of measurement are outlined by rows, and elements of the measurement process form the columns. Examination of established theories and models of measurement in the context of the framework identifies whether uniquenesses of the category of measurement are considered, whether all elements of the measurement process are included in the model, and the extent to which the model is generalizable over categories.

A number of theories and models have been formulated in the different applied measurement literatures. An exhaustive analysis of all measurement theories and models is not possible here because of the considerable numbers involved, but some examples should suggest the unique perspective that this metatheoretical framework provides.

Standard System Issues

The single most important aspect of measurement situations that leads to differentiations between models relates to the operating or construction characteristics of the different standard systems. These impose requirements and/or restrictions on the model. For example, a model involving the elementary standard system should take into consideration the perceptual and cognitive aspects of the assessment, as Feldman's (1981) performance appraisal model in Category 3 has done.

In addition, the model should consider between- and within-standard system variability in assessment. A number of models dealing with this issue have been described in the first three categories. Some examples are Teghtsoonian's (1973) dynamic range theory and Poulton's (1979) discussion of bias in sensory judgments in Category 1; Kazdin's (1977) description of artifact, bias, and complexity in Category 2; and Cronbach's (1955) model of judgmental accuracy in Category 3. Although these three categories focus on similar is-

Figure 2
Schematic Framework for Identifying and Organizing Specific Measurement Models

TYPE OF STANDARD SYSTEM	CATEGORY OF MEASUREMENT	STANDARD SYSTEM ISSUES	DENOTABILITY ISSUES	METRIC INFORMATION ISSUES	ACCURACY ISSUES
ELEMENTARY	1				
	2				
	3				
DEVISED	4				
	5				
DUAL-PROCESS	6				
	7				

sues of human variability in judgment, they differ with regard to the unique types of attributes that are being assessed.

In other categories (e.g., Category 4), the measurement outcome depends on the construction characteristics of the standard system, and issues related to item and test difficulty and whether a time limit is imposed become important. The construction characteristics are quite different in Category 5 (e.g., Martin & Venables, 1980). The operation of physical instruments depends on the particular design and materials used.

Category 6 models of dual-process assessment must consider the dual aspects of the assessment involved in the measurement. The construction characteristics of the standard system and the characteristics of the self-assessment need to be considered. Wiggins (1973), for example, indicated that sources of variance in response to self-report personality tests can be identified with "(*a*) by-products of the strategy under which a scale is constructed, (*b*) item characteristics which introduce method variance, (*c*) organized response styles which exist in the individual, and (*d*) manifestations of a particular construct domain" (p. 426). The sources of variance identified in *a*, *b*, and *d* are associated with the test, whereas that identified in *c* is associated with the individual.

Denotability

Denotability issues are addressed in a model when statements are made regarding issues such as how the attribute is dimensionalized, how the standard system encounters the attribute, and how the assessments are made. Very few measurement models deal with this issue in any detail.

Different standard systems present quite different assessment situations. Elementary standard system assessment involves observation and cognitive evaluation, devised standard system assessment involves a direct interaction between the standard system and the attribute, and dual-process situations involve elements of both. Some models that have discussed denotability issues are Thurstone's (1927) model of discriminal processes in Category 1, Johnston and Pennypacker's (1980) model of dimensional quantities in Category 2, Feldman's model of cognitive assessment in Category 3, and Lord's (1953) model of latent traits in Category 4. Each of these provided some detail that is more or less descriptive and/or speculative concerning the way in which an assessment of the attribute is effected.

Metric Information

Some quite elaborate models have been presented for translating standard system assessments into metric information. In other cases, however, a fully developed model is not specified for a measurement situation. Users of a rating scale, for example, may assume that a metric is built into the creation of a five- or seven-point line segment.

An example of a fully developed model in Category 1 is Thurstone's (1927) Law of Comparative Judgment which permits the transformation of a comparative judgment to an equal interval metric. The model specifies in some detail how proportion judgments may be converted into distance information in the standard normal curve.

An example of a reasonably well-developed model in Category 2 has been outlined by Johnston and Pennypacker (1980). They defined the basic descriptive elements of observable behavior in terms of what they refer to as dimensional quantities of that behavior: latency, duration, and countability. Latency and duration are time-related descriptive elements of the behavior, whereas countability refers to the number of discrete behaviors that occur per unit time.

In Category 3, much of the measurement relies on rating scale assessments. Models in this area generally assume that metric information may be derived directly from the ratings assigned by the elementary standard system. In Category 4, the latent trait model provides for an estimation of the magnitude of the latent ability in terms of item difficulty. Coombs' (1950) model of Unidimensional Unfolding is an elegant model associated with Category 7 measurement.

Models of Measurement Accuracy

The term *accuracy* is used here in a generic sense and is not tied to a particular area or model. In theory, accurate measurement occurs when a faithful representation of magnitudes or differences between magnitudes is captured by the standard system. In practice, accurate measurement may be approximated by adopting appropriate strategies at each step in the measurement process. However, the determination of measurement accuracy usually focuses on the opposite case, that is, establishing whether the measurement outcome includes some type and amount of error. The existence of error may be indicated when standard systems of the same or different type disagree on magnitudes and/or when a relationship between the magnitudes and the indications of the magnitude provided by the standard system cannot be firmly established. A measurement model dealing with accuracy speaks to the issue of how accurate assessment is established or how error is conceptualized or identified.

A number of measurement models in the applied psychological measurement literature may be viewed as models of measurement accuracy or inaccuracy. Models of true score (e.g., Lord & Novick, 1968), generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and person reliability (Lumsden, 1977) all discussed how (in)accuracy should be conceptualized. Although some models have been developed with Category 4 as a focus, variations of the model have been reported in other areas, such as the version of true-score theory outlined for Category 3 measurement (Wherry, 1982).

Other models have identified the basic sources of inaccuracy or have outlined procedures for determining the extent of inaccuracy. For example, Cronbach's (1955) model of judgmental accuracy in Category 3 measurement identified the components of the judgment situation that affect the assessment outcome. Tinsley and Weiss (1975) provided a differentiation between accuracy and reliability for elementary standard system rating assessments and an extensive review of indices that could be used to identify accuracy and reliability in this type of assessment. Saal et al. (1980) recently reviewed conceptualizations of inaccuracy in Category 3 and outlined some statistical procedures for identifying the sources of inaccuracy. Wiggins (1973), in examining accuracy issues in Category 6 measurement, identified two components of variance in measurement. The first component was identified with item characteristics and referred to such characteristics as item social desirability and the direction of item keying. The second component was associated with response style in the self-assessments and included individual tendencies such as lying or role-playing in response to test items.

Conclusions

The major importance of the concept of a standard system lies in the fact that it can be used as a basis for integrating different applied measurement areas with each other and with general measurement theory. Standard systems of measurement are an integral part of applied measurement situations and measurement theory. A standard system makes it possible to convert information about magnitudes of the empirical system into the numerical system. Although different standard systems have been used in different areas of psychology to assess magnitudes of attributes, they can be categorized in a systematic way.

All applied areas of psychology share a concern with measurement issues that can be identified as operationalized requirements for measurement. These components of the measurement process can be generalized as concerns related to the operating characteristics of the standard system, type and degree of denotability of the attribute, modes of determining metric information, and development of criteria and procedures for evaluating measurement accuracy. The statement on the operationalized requirements for measurement and a review of the different measurement mechanisms that are used in psychology also provide a heuristic focus for evaluation of measurement models. A complete measurement model for a given type of standard system would include statements regarding all aspects of the measurement process. These statements may need to be formulated in slightly different form for different standard systems because of the inherent differences in the nature of these standard systems.

The generic concept also serves several additional functions. The definition of measurement that includes the concept of a standard system is more complete when compared to existing definitions in that an agent is specified for the conversion of information about magnitudes into the number system. The concept also highlights the separation of the assessment function of the measurement mechanism from the conversion of that information into numerical form, which serves to describe the measurement procedure more adequately. Finally, and perhaps most importantly, the conceptualization of human judgment as a type of standard system of assessment provides a measurement theory basis for measurement situations involving such judgment.

References

Allport, F. H. (1955). *Theories of perception and the concept of structure*. New York: Wiley.

Astin, A. V. (1968). Standards of measurement. *Scientific American, 218*, 50–62.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London: Longmans, Green.

Committee Final Report. (1940). Quantitative estimates of sensory events. *Advancement of Science, 1*, 331–349.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57*, 145–158.

Coombs, C. H. (1953). Theory and methods of social measurement. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences*. New York: Dryden Press.

Cronbach, L. J. (1955). Processes affecting scores on ''understanding of others'' and ''assumed similarity''. *Psychological Bulletin, 52*, 177–193.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Dawes, R. M. (1972). *Fundamentals of attitude measurement*. New York: Wiley.

Eaton, W. O. (1983). Measuring activity level with actometers: Reliability, validity, and arm length. *Child Development, 54*, 720–726.

Ellis, B. (1966). *Basic concepts of measurement*. Cambridge: Cambridge University Press.

Fassnacht, G. (1982). *Theory and practice of observing behavior*. (C. Bryant, Trans.). London: Academic Press. (Original work published 1979)

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127–148.

Fraser, C. O. (1980). Measurement in psychology. *British Journal of Psychology, 71*, 23–34.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed). New York: McGraw-Hill.

Hutt, S. J., & Hutt, C. (1970). *Direct observations and measurement of behavior*. Springfield IL: C. C. Thomas.

Johnston, J. M., & Pennypacker, H. S. (1980). *Strategies and tactics of human behavioral research*. Hillsdale NJ: Erlbaum.

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco: Chandler.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABC's of reliability. *Journal of Applied Behavior Analysis, 10*, 141–150.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107.

Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517–548.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1*, 477–482.

Martin, I., & Venables, P. H. (Eds.). (1980). *Techniques in psychophysiology*. Chichester: Wiley.

Mitchell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*, 398–407.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86*, 376–390.

Poulton, E. C. (1979). Models for biases in judging sensory magnitudes. *Psychological Bulletin, 86*, 777–803.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413–428.

Stevens, S. S. (1936). A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review, 43*, 405–416.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *American Journal of Psychology, 69,* 1–25.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. New York: Wiley.

Teghtsoonian, R. (1973). Range effects in psychophysical scaling and a revision of Stevens' Law. *American Journal of Psychology, 86,* 3–27.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.

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

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Wherry, R. J. (1982). The control of bias in ratings: A theory of ratings. (Edited and Comments by C. J. Bartlett). *Personnel Psychology, 35,* 521–551.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading MA: Addison-Wesley.

Acknowledgments

The author thanks the anonymous reviewers and the Editor for helpful comments on earlier versions of this paper.

Author's Address

Send requests for reprints or further information to Marion S. Aftanas, Department of Psychology, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada.