The Journal of Mind and Behavior
Summer 1996, Volume 17, Number 3
Pages 213–228
ISSN 0271–0137

Measurement Units and Theory Construction

Warren W. Tryon

Fordham University

The central thesis of this article is that measurement units are theoretical concepts because measurement presumes theoretical definition. New theoretical constructs can be defined in terms of algebraic combinations of previously defined measurement units. Physics has developed an impressive hierarchical knowledge structure on this basis. The unitless measures favored by psychology preclude the generation of such a knowledge hierarchy. It also leads to definitions of reliability and validity in correlational terms which can result in inaccurate measurement. Psychology has long used time as a fundamental unit. Adoption of a second measurement unit would enable construction of a psychological knowledge hierarchy to begin.

Measurement is central to science. Little that we call science would remain if all reference to measured quantities was removed from existing books and journals. The unit of measure is fundamental to measurement in at least two ways. First, and most pertinent to the present discussion, is that measurement units define the fundamental quanta or packets of what is being measured. They specify a standard amount of the phenomenon being assessed and in doing so make a theoretical statement about what is being measured and by implication what is not being measured. For example, Tryon and Williams (in press) define activity in units of acceleration. This physics perspective differs markedly from a biological understanding of activity as calories of heat energy expended, as well as a number of psychological perspectives based on subjective or value judgments associated with movement. Second, standard units of measure enable different investigators to replicate results in the sense that by using the same units, two investigators are more likely to obtain similar data than if no control is exerted over the

Requests for reprints should be sent to Warren W. Tryon, Ph.D., Department of Psychology, Fordham University, Bronx, New York 10458–5198. E-mail can be sent to Wtryon@Murray.Fordham.edu

size of the unit. Exact replication is only possible with equivalent units of measurement because an investigator using a larger unit will find fewer of them than will an investigator using a smaller unit when exactly the same quantity has been replicated. If investigators are unaware of their differing units, they may mistake the differing numbers in their results as evidence against replication. It is the number of units times the size of each unit that bears upon exact replicability. Otherwise, investigators must rely on ordinal replication where only duplication of rank order is evaluated.

Evidence of the theoretical nature of measurement units is plentiful in physics where a knowledge hierarchy has been created by defining new concepts in terms of previously defined units. After commenting on the absence of units in psychology, the knowledge hierarchy built from measurement units in physics is briefly reviewed. Additional evidence of the theoretical nature of measurement units is presented by noting that characteristic properties, that is, traits, of substances are defined in terms of measurement units. This also enables the concept of accuracy. It is shown how measures can be perfectly reliable and perfectly valid but be inaccurate in the absence of standard units.

Units in Psychology

Psychophysics is the field of psychology that has paid closest attention to measurement units. Mapping the psychological impact of physical stimuli caused investigators to retain measurement units for the physical stimuli. Scaling techniques were derived for estimating the magnitude of psychological response but standard units were not derived. Psychometricians developed many psychological tests but did not define any standard measurement units. Rather, they reported test results as T-scores which are derived from z-scores where the mean z-score is always zero and the standard deviation is always 1.0 (Hinkle, Wiersma, and Jurs, 1994). Multiplying z-scores by 10 and adding 50 produce the familiar T-score centered at 50 with a standard deviation of 10. MMPI T-scores are derived in this way. Patients' scores on depression can be directly compared to their scores on schizophrenia despite vast theoretical differences in what is being tapped. Equal T-scores on the two scales implies that the person is as depressed as he/she is schizophrenic. Multiplying z-scores by 100 and adding 500 result in the traditional College SAT scores centered at 500 with a standard deviation of 100. Intelligence quotients are derived by multiplying z-scores by 16 and adding 100 (Anastasi, 1988).

Effect sizes refer mainly to differences between means. Rather than report the difference in terms of the number of measurement units by which the two groups differ, psychologists have chosen to express effect sizes in terms of

z-scores by dividing the difference between the means of the experimental and control groups by the standard deviation of the control group or the pooled standard deviation. The z-score is a theoretical two-edged sword. Its benefits include the ability to quantitatively compare the relative standing of subjects across qualitatively disparate measurements. All measurements for all subjects are placed on the same –3 to +3 scale centered on zero with zero being the average. The primary negative consequence is the loss of the original measurement unit. The numerator of the z-score equals the difference between a person's score and the mean score and is therefore expressed in original measurement units. The denominator of the z-score equals the standard deviation expressed in the same units of measure. Hence, the original units of measure cancel, resulting in a unitless measurement. All theory construction effort devoted to the development of measurement units is discarded. Any unit of measure that may have been developed is replaced by the standard deviation. This result is extremely problematic because the size of the measurement unit is not constant but is dependent upon who was measured — it differs with every new group of subjects. The standard deviation changes as people are added or dropped from a given group and it changes from one testing to the other. Imagine the quandary that could be created in physics by making the length of a meter stick dependent upon who or what is measured. It is difficult to estimate the confusion psychologists have created by linking unit size to subjects studied; yet psychologists appear content to routinely allow measurement units to continuously change in unknown and unpredictable ways.

Null Hypothesis Testing

Rejection of the null hypothesis is a primary focus of experimental design and analysis. Cohen (1994) observed that an unfortunate consequence of null hypothesis testing has been to devalue units of measure. Cohen's views are shared by the prominent psychometrician Tukey (1969) who states "Being so uninterested in our variables that we do not care about their units can hardly be desirable" (p. 89). Cohen recommends that null hypothesis testing be replaced by confidence intervals and regression equations but observes that "To work constructively with 'raw' regression coefficients and confidence intervals, psychologists have to start respecting the units they work with . . ." (p. 1001). These comments underscore the importance of measurement units.

Reaction time measurement is an important exception to Cohen's criticism. Experimental psychology has studied reaction time since its inception in 1879 (see Brebner and Welford, 1980) and clinical psychology has studied response latencies since 1910 (Jung, 1910, 1918). The unit of measure is the

second and data are frequently measured to an accuracy of 1 millisecond. More recently, personal computers have set the occasion for presenting personality test items and measuring subjects' response times (Holden and Fekken, 1990; Holden, Fekken, and Cotton, 1990, 1991; Holden, and Kroner, 1992; Hsu, Santelli, and Hsu, 1989; Tryon and Mulloy, 1993). However, this literature quickly moved to dispose of measurement units through a double z-transformation to control for differences in item length and other variables.

There are at least two undesirable consequences associated with transforming data into a unitless metric. First, it precludes constructing hierarchical knowledge structures based on algebraic combinations of basic units to create derived concepts. This includes the important concept of characteristic properties which is defined in terms of measurement units. Second, it has led to the definition of reliability and validity in correlational terms which has created the previously mentioned, and yet to be explained, possibility that measures can be perfectly reliable and valid but inaccurate.

Hierarchical Knowledge Structure

This section makes two interrelated points. The first point is that measurement units are theoretical quantities and not unimportant methodological details. Second, one can begin constructing a knowledge hierarchy from measurement units once two fundamental units have been defined by combining these two units; usually through division and/or multiplication.

Theoretical Units

Conceptualization, definition, and measurement are related ideas. Conceptualization entails formulating *essential properties*, necessary and sufficient attributes which give a concept meaning. Prototypes are examples of empirically derived entities where common elements of various instances of a concept are abstracted into an idealized representation that has never before been experienced and may never be experienced exactly as represented. For example, having seen 100 different dogs of varying species, one abstracts features of an idealized dog that is characteristic, to some degree, of all dogs seen but does not exactly represent any specific dog seen in the past or perhaps any dog that will ever be seen. Definition specifies properties of the concept so that others can recognize *positive instances* of it. Having four legs, tail, fur, barking, etc. are aspects that define a dog. An amputee dog will be recognized as a dog despite only having three legs because it meets other criteria. Concept and definition are intimately connected. Thus, when definition is sufficiently clear to allow others to reliably identify positive and

negative instances of the concept, then the definition functions as a binary measurement device capable of discriminating 1 = presence from 0 = absence of the concept. Repeated application of the definition determines numerosity. For example, by knowing what constitutes a positive instance of the concept "chair," I can apply it to every object in the room to count the number of chairs and can reapply the same definition to determine if this number increases or decreases subsequent to other events. The concept in question is the basis of numerosity and functions as a measurement unit. Therefore, concepts *are* measurement units. Logical equivalence allows us to further conclude that measurement units are concepts.

Physicists recognize two fundamental theoretical quantities: length and time. The current standard unit of *length* is the meter, defined as 1,650,763.73 wavelengths of orange–red light emitted from krypton-86. Any fixed natural standard would do; but this one agrees with previous definitions of the meter and will not deteriorate over time as a fixed metal bar might through oxidation or other processes. The second fundamental quantity recognized by physicists is time. The current standard unit of *time* is the second, defined as 9,192,631,770 transitions between two energy levels of the cesium-133 atom. The concepts of length and time are fundamental to science in that they provide the basis upon which physicists have built their knowledge hierarchy.

Knowledge Hierarchy

New concepts are derived from the multiplication and division of two previously defined units of length and time. Our first example is the concept of *area* which derives from multiplying two length measurements taken in two physical dimensions, e.g., length times width, and its measurement unit is the square meter (m^2). Most likely the concept of area was initially applied to squares and rectangles and then extended to circles, triangles, and irregularly shaped polygons. Extending the concept of *squared* units of length to geometric shapes that were *not square* demonstrates the conceptual nature of the measurement unit we call area.

The concept of *volume* derives from extending the concept of area into a third dimension by multiplying three length measurements taken in three physical dimensions, e.g., length times width times height. The resulting unit is the cubic meter (m^3). Though probably the concept of volume was initially applicable to cubes and solid rectangles, its theoretical nature is demonstrated by extending it to spheres, cones, and irregularly shaped objects. The unit of measure remains *cubed* units of length.

The definition of *mass* was derived from the unit of length in conjunction with a common substance, water. One gram was originally conceptualized as

the weight of 1 cubic centimeter of water at its maximum density in a vacuum. Practical problems of evaporation, etc. resulted in the development of a platinum–iridium bar equivalent to 1,000 grams as kept by the Bureau of Weights and Measurements. The important point here is that the new concept of mass was formulated in terms of the *derived concept* of volume in conjunction with a common substance, water. Put otherwise, the concept of mass could not exist as we know it without the previously defined concept of length. The concept of mass is predicated upon length.

The concept of *velocity* is derived by dividing units of length (distance) by units of time. Interestingly, time is taken as the primary or constant unit whereas distance is taken as the secondary or variable unit. Position is examined at 1 second intervals and the number of meters traversed is measured. The resulting units of measure for velocity are meters/second.

An alternative conception of velocity could have been developed by taking units of length as primary (constant) with units of time as secondary (variable). Time could be measured to traverse 1 meter distances resulting in velocity units of seconds/meter. Theoretical physics might well have evolved differently had this latter definition of velocity been adopted as standard.

The concept of *acceleration* was derived by dividing units of velocity by units of time resulting in meters per second per second ($m/s/s = m\ s^{-2}$). *Deceleration* was derived as negative acceleration. Objects falling to earth traverse approximately 9.8 meters during the 1st second, 2 x 9.8 = 19.6 meters during the 2nd second; 3 x 9.8 = 29.4 meters during the 3rd second, etc. Newton observed that this mathematical series pertained to both heavy and light objects and concluded that all objects fall at the same rate. Generalizing this observation to both earthly and celestial objects, he proposed the universal law of gravitation. Newton would not have been able to conceptualize a law of gravitation in the absence of *units* of length and time because the law presupposes both.

Newton derived the concept of force as the product of mass times acceleration ($f = ma$). The term "equation" indicates that both sides are logically and quantitatively equivalent. Logical equivalence implies *is conceptually equal to*. The term "formula" implies that we have formulated a theoretical relationship between previously defined units. Multiplying 1 Kg times an acceleration of 1 meter per second per second equals one kilogram–meter per second squared ($Kg\ m/s^2$). The force required to produce this result was defined as the unit of force and is called the Newton (N), in Isaac Newton's honor (cf. Heimler, 1989, p. 586). Put otherwise, 1 Newton is the force required to accelerate a 1 Kg mass to an acceleration of 1 m/s/s.

The concept of *work* derives from units of force and length. It is formulated as the force necessary to move a specified mass a particular distance. Applying the force of 1 N until 1 Kg is moved 1 m expends 1 Newton–meter (N m) of

energy — which has been named 1 Joule. Importantly, one unit of work was defined in terms of one unit of mass (Kg) times one unit of distance (m).

The concept of *power* was derived by dividing the unit of energy by the unit of time. It was formulated as the rate with which energy is consumed. Expending 1 J of energy per second defines 1 watt (W) which can be thought of as a Newton–meter per second (N m/s). The concept of *pressure* was derived by dividing the unit of force by the square unit of length, called area. It was formulated as a constant force applied to a specific area. Applying the force of 1 N over an area of 1 square meter equals 1 Pascal ($Pa = N/m^2$) of pressure.

The above examples can be extended until all of the concepts based on the meter–kilogram–second (mks) unit system have been described. The idea is that physicists derived many higher order concepts from three basic ones (meter, kilogram, second) and even one of the basic units, the kilogram, was derived from cubic length, volume, and water. In the absence of units, the aforementioned algebraic manipulations would make no more sense than, for example, anxiety times depression, self-confidence divided by hypochrondriasis, or the cube of introversion. Units are theoretical constructs based either on primitive definitions or on algebraic combinations of other units. Physics has taught us that only two fundamental units are needed to begin the process of creating a hierarchical knowledge structure. Psychologists should seriously consider what concepts are fundamental to their field. If only two can be identified, then a third may be derived from the first two and the beginnings of a knowledge hierarchy may emerge.

Characteristic Properties

Physical substances have traits called characteristic properties. Our understanding of characteristic properties is entirely dependent upon previously derived units of measure. For example, dividing the weight of an object in grams by its volume in cubic centimeters formulates the characteristic property know as *density*. *Concentration* of a liquid is theoretically specified as grams of solute per cubic centimeters of solvent. The characteristic property known as *solubility* is formulated as the maximum concentration possible at a specified temperature; solubility increases with temperature. Another characteristic property is the temperature at which a liquid turns solid, its *freezing point*. *Boiling point* is another characteristic property and is formulated as the temperature at which a liquid becomes a gas. The last two characteristic properties are especially important because in combination with a common substance, water, they define the centigrade unit of temperature. One degree on the Celsius temperature scale is defined as 1/100 of the difference between the freezing and boiling point of water. *Specific heat* is defined as the number of calories needed to raise the temperature of 1 gram of a substance

1 degree C. The characteristic property known as *specific gravity* is the ratio of the mass of a given volume of the substance to the mass of the same volume of a comparison substance; water for solids and liquids, hydrogen for gases.

Points of invariance can serve as important references in a world of variation. Characteristic properties provide reference points for the natural sciences. For example, specific properties are useful in identifying unknown substances because all substances have a unique profile of characteristics. Sometimes it is possible to identify or separate substances on the basis of a single characteristic property. The density of pure gold is 19.32 g/cm^3 but the density of lead is 11.35 g/cm^3. Any attempt to pass off gold plated led as pure gold can be detected by evaluating its density. In other cases, two or more characteristic properties may be required for unique identification. Without the theoretical basis provided by the derived units of measure, characteristic properties as we know them today would probably not have been discovered.

Psychologists have long sought to find stable dimensions along which to characterize people. For example, personality is thought to develop from infant temperament (Ahadi and Rothbart, 1994; Eaton, 1994; Martin, Wisenbaker, and Huttunen, 1994; Wachs, 1994) into a stable five factor structure that remains rather constant over time (Costa and McCrae, 1980). Understanding that units of measure can be combined into new units, some of which constitute characteristic properties, provides additional motivation to consider measurement units in psychology.

Reliability, Validity, and Accuracy

The purpose of this section is to demonstrate that our concepts of reliability and validity are distorted by the absence of measurement units such that it is possible to develop a measurement device that is completely reliable and valid but inaccurate. This problem is eliminated once a measurement unit that allows calibration is introduced.

Consider how one might evaluate the reliability of a postal scale where a unit of measure, e.g., the gram, exists. To begin, a particular object could be weighed repeatedly and the result recorded. The data will vary due to random measurement error. The standard deviation of a large number of measurements might be defined as the standard error of measurement. Accuracy of the scale could be reevaluated with several different masses to determine if the standard error of measurement is directly or inversely proportional to the absolute weight of the test object.

Psychologists do not approach reliability in this way. If they employ repeated measurement it is to test and then retest the same group of people. The correlation between the two measures is accepted as an index of temporal stability. Split-half correlations or Chronbach's alpha, which equals the

mean of all possible split-half correlations, is also calculated to determine internal consistency. Our definition of standard error is presented in terms of the reliability coefficient. Anastasi (1988, p. 133) reports that the standard error of a measurement equals the standard deviation of the single measurements made on individual subjects times the square root of one minus the reliability coefficient. All of these correlation-based approaches to reliability substitute pairs of measurements over subjects for repeated measurements of the same quantity. These two approaches to reliability are fundamentally different. Here again, the role of measurement units is especially important. Measurement units force one to specify the amount of error in terms of the theoretical concepts used to define the units. This provides a more absolute interpretation of inconsistency because measurement units provide an interpretative reference point. Validation of the postal scale would be done by placing standard weights on it and comparing the obtained reading with the theoretically expected reading. The scale could be adjusted until measurement discrepancies became smaller than an acceptable maximum. Stated otherwise, we can now calibrate instruments by comparing obtained readings with an official standard thereby permitting the concept of accuracy. The instrument is altered until repeated measurements are within tolerable limits from the standard. This calibration process certifies that the measuring device is valid, reliable, and *accurate* to within the stated tolerance. Accuracy computations can be calculated when the unit of measure of the device and the standard reference are the same. Only then can one subtract the measured quantity from the reference value to compute a deviation value that calibration procedures can minimize.

Psychologists often lack a criterion expressed in the same units of measure as their test or measuring device and therefore cannot determine accuracy or calibrate their instruments. Psychologists are thus reduced to demonstrating validity by correlating test performance with an external criterion measured in some other unit. Instead of minimizing the difference between observed and expected values expressed in the same unit of measure, psychologists correlate two disparate measurements. Defining reliability and validity in terms of the unitless correlation coefficient gives rise to at least the following two fundamental problems.

(1) Winer (1971) describes how to evaluate reliability using intraclass correlation coefficients. His Table 4.5–1 (pp. 283–289) calls for obtaining k repeated measurements on N persons. The estimate of true variance (TV) is obtained by subtracting the mean square within subjects, error variance (EV), from the mean square between subjects and dividing by k. This approach breaks down when people are replaced by instruments. High quality calibrated instruments respond so similarly that between instrument variability approximates within instrument variability and reliability approaches

zero despite the high degree of consistency, validity, and accuracy of the instruments! This example reveals that our measure of instrument (test) reliability depends critically on between subject variability in the sample chosen for study; a feature that should be independent of the quality of the measuring device.

Tryon (1991) suggested an alternative index of reliability based on the coefficient of variation (CV) which equals the standard deviation of a set of repeated measurements divided by the mean of the measurement set times 100 to express the result as a percentage. Repeated measurements must form a scatter plot whose diameter is an acceptably small fraction of the mean value in order to be considered reliable. Since CV is an error proportion as is $1\text{-}r^2$, the two can be equated (i.e., $CV = 1\text{-}r^2$). Solving for r yields $r = [1\text{-}CV]^{0.5}$.

(2) The absence of units of measure allows the following paradox to result (cf. Tryon, 1991, pp. 5–6). Imagine that five different investigators construct their own thermometer. They each independently establish the reliability of their device by demonstrating near perfect test–retest correlation coefficients across a variety of substances. Imagine that they each perform validity studies where they compare readings taken from water heated by a constant flame for varying periods of time obtaining progressively greater readings and consequently perfect rank order validity coefficients. If our investigators share their raw data, they would find that all $5(4)/2 = 10$ intercorrelations approach unity thereby constituting overwhelming cross validational evidence. Each investigator would therefore be confident that he or she had a nearly perfectly reliable and valid instrument by current psychometric standards. Finally, imagine that these investigators meet at a convention where they have gathered to reveal their perfectly reliable, perfectly valid thermometers to each other. When presented with a beaker of water, say at room temperature, they would all eagerly insert their thermometers and to everyone's shock find the beaker to simultaneously have five different temperatures! How can this be if each device is perfectly reliable and perfectly valid and if all devices intercorrelated almost perfectly? The answer is that none of them defined a unit of measure — they were unknowingly working with five different measurement units. By choosing glass tubes of different diameters and lengths and inserting different amounts of mercury, or other substances, into different size reservoirs at the bottom of each thermometer, the absolute heights of each column of substance in each thermometer differ even when placed in a medium having a single temperature. Some investigators may have used bimetalic strips connected to pointers thereby further augmenting discrepancies among devices.

Establishing a unit of measure completely resolves all confusion. First, place all five devices in ice and mark each dial. Then place the same five devices in boiling water and mark each dial. Finally, divide the five intervening dis-

tances into 100 equal parts thereby creating five centigrade thermometers. When replaced into the previous room temperature beaker of water, all five devices will now indicate the same, or very similar, temperature.

An alternative explanation of the above problem is that correlations are invariant over linear transformation. Consistently adding, subtracting, multiplying, and/or dividing one variable leaves its correlation with the other variable unchanged but can cause any amount of discrepancy between the means and/or variances of the two variables thereby producing varying degrees of inaccuracy.

If investigators working with perfectly reliable and valid instruments can encounter such empirical conundrums as single beakers of water with five simultaneously different temperatures, then even greater confusion can be expected when working with tests of moderate reliability and validity. Standard efforts to increase reliability and validity will not solve the problem because the intercorrelations among the five devices already approaches unity. Suggesting that the room temperature of a beaker of water is a multidimensional construct, requiring multiple convergent measurement, transforms a simple concept into an unnecessarily complex one. That the empirical dilemma of the multi-temperature beaker resolves rapidly and completely given a common unit of measure suggests that similar clarity may result where units of measure can be derived.

Choosing ice and boiling water as reference points and dividing the intervening distance on the thermometer into 100 equal parts are arbitrary decisions. It is necessary that investigators agree on a standard and derive a common measurement unit. Agreeing on ice and boiling water as two reference points along a thermal continuum enables a unit of measure to be derived. Given the success of this approach in the natural sciences, psychologists also might find it useful.

Psychological Units: Previous Attempts and Possibilities

A knowledge hierarchy, like the one illustrated above, can begin with the identification of two measurement units and expand with the addition of each new unit. Hence, it is important to consider efforts by psychologists to identify units of measure.

Johnston and Pennypacker (1980) stress the theoretical importance of units for psychology. They articulate a behavioral approach to seeking natural units and recommend the following six measurement aspects: (1) latency measured in time units, (2) duration measured in time units, (3) countability measured in cycles, (4) frequency measured in cycles per unit time, (5) celeration measured in cycles per unit time per unit time, and (6) inter-response time measured in time per cycle (cf. p. 128). Time is mentioned in five of the

six suggested units. The concept of behavioral cycle is the new unit. A behavioral cycle is defined as persisting until a point is reached where the cycle can start over (cf. p. 126). Rather than define a unit of measure in reference to two points on a continuum, as with the thermometer, this approach is based on the concept of periodicity. Like the sine or cosine trigonometric functions, different values obtain throughout the first rotation around the unit cycle but repeat once the start point is encountered a second time. While any position can be chosen as the starting point in the trigonometric example, some behavioral reference points may be better than others. The time to return to a start point would then quantify the duration of one cycle. The inverse of this time would equal the frequency of this event.

The above approach assumes that psychology requires new fundamental units of measure. While it may be true that some new useful measures will be derived, little systematic consideration has been given to the adequacy of existing units for describing behavior. Tryon (1991, pp. 1–22) recommends applying the meter–kilogram–second (mks) system of measurement used in physics to measure behavior. He calls this extension of behavioral assessment *behavioral physics*. Since behavior entails movement in space over time it can be described in physical terms. For example, pedometers are usually calibrated in terms of miles or kilometers walked. Accelerometers are calibrated in terms of g where g = 9.8 m/s/s: the rate with which objects freely fall to the ground. While some devices attach to the wrist or ankle, the waist is an especially important site of attachment because it corresponds to the body's center of gravity which allows one to ascribe the person's body mass in kilograms to this point. Integrating vertical acceleration over time yields velocity; which integrated over time gives distance moved; which multiplied by weight yields work energy in Joules. Dividing work energy either by total seconds of wearing-time yields power in watts. Tryon (1991, pp. 138–141) describes how calories of energy expended by activity can be calculated from measures of vertical acceleration about the waist. These calculations employ the knowledge hierarchy described above.

This approach to measuring behavior has at least two benefits. First, it solves the unit problem by adopting well defined and accepted measurement units. It seems efficient to pursue this course simultaneously with exploring new psychological units rather than to hold psychological research in abeyance until progress has been made on the unit issue. Second, it may be that substantial progress can be made by reconsidering psychology from the perspective of mks units. Psychology is assumed to be a natural science, at least by some investigators. We can then reason as follows: natural sciences developed theoretically and empirically using mks measurement units, psychology is a natural science, therefore psychology will develop both theoretically and empirically using mks measurement units.

In general, physical instruments that quantify behavioral and psychological states make use of the mks measurement system. Reaction time is the classic example. Response rates, latencies, and forces placed on operant manipulanda are measured in terms of the mks system. Computer administered personality tests and neuropsychological exams and rehabilitation programs employ mks units. All physiological measures such as EEG, skin conductance, temperature, blood pressure, etc. are defined in mks units. Biofeedback modalities are also definable in mks units. Some activity monitors report distance walked while others integrate forces of acceleration detected at the site of attachment. Brain imaging of resting or psychologically active subjects also reports in these units.

Our challenge lies in deriving units for measurements for constructs currently tapped by psychological tests. Units of measure could be introduced into the assessment of intelligence in the following way. Beginning from an information processing theory of intelligence, one might define the characters per second that can be processed on a digit-symbol (or similar) task as a measure of intelligence. For example, a computer-administered digit-symbol task could be formulated so that the first key press activates a count-down timer which would terminate the task when time expires. The number of correct and incorrect responses plus total digit symbols processed would be the dependent variables. The unit of measure for these data would be digits/second. Time, the denominator, is the primary or constant metric while the number of digits processed within each unit would vary and thus would constitute the secondary or variable unit. One would explain variation in total numbers of responses as a function of experimental conditions or subject selection.

Alternately, the time taken to process each correct and/or incorrect digit symbol could be tallied. The constant unit of measure now is the response. The number of milliseconds necessary to complete each response varies. One would explain variation in time as a function of experimental conditions or subject selection.

Conclusions

The present article is critical of the unitless nature of much of contemporary psychology because such an approach (a) precludes evaluating the accuracy of our measurements, and (b) precludes the development of a knowledge hierarchy. Cohen (1994) has detailed problems with null hypothesis testing and indicates that units of measure must be attended to if psychology is to advance as a science. Physics possesses defined units of measure and consequently derived a hierarchical set of concepts by combining previously defined units through the algebraic operations of division and multiplication.

Measurement units have enabled physicists to define fundamental characteristic properties of matter. The absence of measurement units has led to psychometric definitions of reliability that break down when applied to instruments such as activity monitors — further illustrated by the thermometer example — and which may be responsible for other instances where inconsistent findings are reported.

Units of measure are fundamental theoretical quantities and should be respected as such. This is not a simple-minded call for a return to operationalism but rather a request that we think fundamentally about elementary aspects of behavior and cognition and what common reference points can be used to create standard units of measure. Psychology has a long history of using time as a fundamental unit of measure. The introduction of a second measurement unit will allow the possibility of combining it algebraically with the first one, to define a new concept, and thereby to develop a knowledge hierarchy like the one described above.

References

Ahadi, S.A., and Rothbart, M.K. (1994). Temperament, development, and the big five. In C.F. Halverson, Jr., G.A. Kohnstamm, and R.P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 189–207). Hillsdale, New Jersey: Lawrence Erlbaum.

Anastasi, A. (1988). *Psychological testing* (sixth edition). New York: Macmillan.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (third edition, revised). Washington D.C.: Author.

Brebner, J.M., and Welford, A.T. (1980). Introduction: A historical background sketch. In A.T. Welford (Ed.), *Reaction time* (pp. 1–23). New York: Academic Press.

Costa, Jr. P.T., and McCrae, R.R. (1980). Still stable after all these years: Personality as a key to some issues in adulthood and old age. In. P.B. Baltes and O.G. Brim (Eds.), *Life-span development and behavior* (Volume 3, pp. 65–102). New York: Academic Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.

Eaton, W.O. (1994). Temperament, development, and the Five-Factor Model: Lessons from activity level. In C.F. Halverson, Jr., G.A. Kohnstamm, and R.P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 173–187). Hillsdale, New Jersey: Lawrence Erlbaum.

Heimler, C.H. (1989). *Focus on life science*. Columbus, Ohio: Merrill Publishing Co.

Hinkle, D.E., Wiersma, W., and Jurs, S.G. (1994). *Applied statistics for the behavioral sciences* (third edition). Boston: Houghton Mifflin Co.

Holden, R.R., and Fekken, G.C. (1990). Structured psychopathological test item characteristics and validity. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*, 35–40.

Holden, R.R., Fekken, G.C., and Cotton, D.H.G. (1990). Clinical reliabilities and validities of the microcomputerized Basic Personality Inventory. *Journal of Clinical Psychology, 46*, 845–849.

Holden, R.R., Fekken, G.C., and Cotton, D.H.G. (1991). Assessing psychopathology using structured test–item response latencies. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*, 111–118.

Holden, R.R., and Kroner, D.G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment, 4*, 170–173.

Hsu, L.M., Santelli, J., and Hsu, J.R. (1989). Faking detection validity and incremental validity of response latencies to MMPI subtle and obvious items. *Journal of Personality Assessment, 53*, 278–295.

Johnston, M.M., and Pennypacker, H.S. (1980). *Strategies and tactics of human behavioral research.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Jung, C.G. (1910). The association method. *American Journal of Psychology, 21*, 219–269.

Jung, C.G. (1918). *Studies in word association.* New York: Moffat, Yards, and Co.

Martin, R.P., Wisenbaker, J., and Huttunen, M. (1994). Review of factor analytic studies of temperament measures based on the Thomas–Chess structural model: Implications for the big five. In C.F. Halverson, Jr., G.A. Kohnstamm, and R.P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 157–172). Hillsdale, New Jersey: Lawrence Erlbaum.

Tryon, W.W. (1991). *Activity measurement in psychology and medicine*. New York: Plenum Press.

Tryon, W.W., and Mulloy, J.M. (1993). Further validation of computer-assessed response time to emotionally evocative stimuli. *Journal of Personality Assessment, 61*, 231–236.

Tryon, W.W., and Williams, R. (in press). Fully proportional actigraphy: A new instrument. *Behavior Research Methods Instruments & Computers*.

Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83–91

Wachs, T.D. (1994). Fit, context, and the transition between temperament and personality. In C.F. Halverson, Jr., G.A. Kohnstamm, and R.P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 209–220). Hillsdale, New Jersey: Lawrence Erlbaum.

Winer, B.J. (1971). *Statistical principles in experimental design* (second edition). New York: Wiley.