DOI: 10.1111/j.1475-6773.2005.00443.x

Validity of Measures Is No Simple Matter

Lee Sechrest

Purpose and Method. This article aims to promote a better understanding of the nature of measurement, the special problems posed by measurement in the social sciences, and the inevitable limitations on inferences in science (so that results are not overinterpreted), by using the measurement of blood pressure as an example. As it is necessary to raise questions about the meaning and extent of the validity of something as common as measured blood pressure, even more serious questions are unavoidable in relation to other commonly used measures in social science. The central issue is the validity of the inferences about the construct rather than the validity of the measure per se.

Conclusion. It is important to consider the definition and validity of the construct at issue as well as the adequacy of its representation in the measurement instrument. By considering a particular construct within the context of a conceptual model, researchers and clinicians will improve their understanding of the construct's validity as measured.

Key Words. Validity, measurement, blood pressure, constructs

It is common in published articles in the social and behavioral sciences to encounter a statement to the effect that a measure used in research can be considered "valid." In reviewing research proposals for scientific review committees, statements such as "This measure has been shown to be valid," or "This measure has demonstrated construct validly," abound. The idea of "validity" of measures is often taken to be straightforward, and, indeed, it may be if one confines one's interests to *empirical* or *predictive* validity (utility). A measure can be considered to have empirical validity to the extent that it correlates with some other phenomenon in which one is interested. From a conceptual/theoretical perspective this is a nearly trivial case, even though empirical validity may have considerable practical importance.

More often than not, however, when the idea of the validity of a measure is at stake, the interest is in *construct validity,* a term introduced in 1955 by Cronbach and Meehl, in one of the most important articles on measurement ever published in the social and behavioral sciences. Construct validity refers to the extent to which a measure reflects accurately the variability among

objects as they are arrayed on the underlying (latent) continuum to which the construct refers. Since an underlying or latent variable cannot be directly observed, there is no direct way to determine just how well a measured variable, the one we can observe, maps onto the underlying variable. Thus, there is no way of attaching any numerical quantitative estimate to the idea of construct validity, although we might have a strong sense that construct validity is greater in one case than another. As Cronbach and Meehl (1955) made clear, construct validity of a measure is established by demonstrating its place in a *nomological net* of consistent, related empirical findings. An *impression* of construct validity emerges from examination of a variety of empirical results that, together, make a compelling case for the assertion of construct validity for a given measure. Construct validity cannot be assumed simply because a measure correlates with some other measure or because in a factor analysis it seems to have an appropriate factor structure.

Those sorts of things are merely prerequisites, the beginning places for the search for construct validity. The complexity of the construct validity problem is made clear by Messick's (1989, 1995) delineation of six facets of construct validity: content, substantive, structural, generalizability, external, and consequential.[1] Construct validity is not simply a property of a measure but is a reflection of and resides in the conditions of its use.

The issues involved in validity are far too complex to fit well into any simple scheme, and they are remarkably difficult to translate into practice. Even standards such as the Joint Standards (APA, AERA, NCME 1999) for tests proposed by a consortium of professional organizations are difficult to apply and do not necessarily make much sense once one gets beyond the realm of commercially marketed instruments. Achievement of, or even understanding of, construct validity cannot be guaranteed by any template of requirements by which to judge the adequacy of measures, whether those measures are extant, in the developmental process, or merely being contemplated. Rather what is needed is a deeper understanding than seems to be prevalent of what is meant by validity and harder thought about how the measures we work with fit our conceptions of what validity is or ought to be.

The crux of the matter lies in Messick's assertion that "Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores" (Messick 1995, p. 741). That is exactly the problem: it is very difficult

---

Address correspondence to Lee Sechrest, Ph.D., Department of Psychology, University of Arizona, Tucson, AZ 85721.

to know what the test scores mean. It is not measures that are valid, but the scores that they yield and the interpretations we make of them.

It is important here to note the apparent, but perhaps not so real, difference between the position being taken here and that of Borsboom, Mellenbergh, and van Heerden (2004), who insist that the issues surrounding validity of tests can be made quite simple once one realizes that validity simply means that a construct exists and that the construct *causes* scores on the test. Their position is that mere correlation between two variables, even if they can be ordered with respect to time or conceptual priority, does not mean that one is a measure of the other. I would grant them their position and commend the clarity of their thought. The complexity with which I am concerned here, however, has a great deal to do with the correct specification or identification of the construct in the first place. That problem is at the core of our difficulties with validity and is, I insist, no simple matter.

The Measurement of Blood Pressure

Take blood pressure measurement as an example of a measurement task considered pretty well solved by the existence of good instruments. I choose this example because it is not in the realm of social science, perhaps making the issues to be dealt with somewhat starker but less effectively arousing than would an example from psychology or another social science. In fact, however, even more than 100 years after the invention of the sphygmomanometer, problems in measuring blood pressure persist, along with uncertainties about the meaning of the values obtained (Parati 2004). The medical literature on blood pressure measurement and meaning grows by dozens of articles each year.

The central problem of validity for psychosocial measures is, to reiterate, almost always *construct validity*, the extent to which we can legitimately claim that a measure reflects variability in the construct it purports to measure. Constructs are hypothetical factors that underlie (cause) behaviors, including those reflected in measuring devices or processes, and they cannot be measured directly, but only estimated. A major problem arises from the possibility that constructs of interest may be conceived at several different levels of meaning, and construct validity at one level may not apply to validity at another level. Another major problem, however, arises from the fact that what constitutes a "measure" may be conceived at different levels, and construct validity may obtain across those levels. To take one simple example,

self-reported physical functioning may be taken as a "measure" of what the subject wishes a clinician or researcher to know about his physical functioning, and the construct validity of the measure at that level would probably be quite good. Or, the same self-report could be taken as a "measure" of self-perceived physical functioning, and construct validity at that level would usually be good as long as one could assume that the subject did not have any strong reason to mislead the clinician or researcher. Or, the same self-report could be taken as a "measure" of "actual" physical functioning, as might be done in a study of treatment outcomes. Construct validity at that level might be quite variable, depending on the conditions of measurement, e.g., whether the subject wanted to flatter the clinician or whether the subject wanted to preserve secondary gain from disability.

To anticipate just a bit, the construct "blood pressure" has different meanings; our interest in it is, probably most of the time, in relation to its meaning as a measure of fitness/illness rather than blood pressure per se, e.g., we make allowance for anything that might have had a temporary effect of raising blood pressure. And "the validity" of an instrument is not the same as the validity of the measure (data) it produces.

What Blood Pressure Are We Interested in?

Blood pressure actually involves two elements (concepts), not just one: diastolic and systolic blood pressures. They are measured in the same general way and during the same process, but they are quite different concepts and phenomena and may not be measured equally well (Kay 1998). Measuring diastolic blood pressure requires a judgment about just when the artery is no longer constricted, and errors may be more frequent or larger in measurement of diastolic pressure. Potentially complicating matters is that blood pressure may be measured, with different results, when subjects are in a standing, sitting, or supine position. Some recommendations for accurate measurement of blood pressure call for measures to be taken in all three positions (e.g., Goldman 2002). So, that provides for the possibility of six different blood pressures. If the issue were simply one of the dependability (stability) of the measurement, it would do as well to measure blood pressure three times in, say, the sitting position. But the values from standing and supine position measures are thought to carry additional information, i.e., information about an at least somewhat different construct.

Blood pressure is conventionally measured in the upper arm, but measurement at the ankle may also be desirable. Clinicians sometimes measure the

blood pressure *gradient* between arm and ankle sites (brachial-ankle) in order to assess arterial occlusion in the lower extremities (Baccellis et al. 1997). And simply to extend the picture, there is some reason to believe that *pulse pressure*, the difference between diastolic and systolic pressures, may be critical information (Engvall et al. 1995; DeStefano et al. 2004). So, what is a blood pressure device, e.g., the sphygmomanometer, a "valid" measure of? Apparently it is to be taken as a measure of several different constructs, or, perhaps more accurately, as part of a measurement process for different constructs.

Diastolic and systolic blood pressures, although different concepts, are to some extent related to each other (but try to find data on the correlation between them!). The relationship, however, may be conceived as conceptual, or as empirical, or both. That is, diastolic and systolic blood pressures both represent pressure on the walls of arteries but at opposite points in the pumping cycle of the heart. They are probably correlated to some degree, but is that simply an empirical fact or is it inherent in the definitions of the two concepts? It could be that the cardiovascular system is constructed in such a way that, in "normal" people, the stronger the contraction forcing the blood through the arteries, the stronger the residual pressure when the ventricular muscle relaxes. What should the diastolic/systolic relationship be, then? High? Moderate? Low? But maybe the system is not constructed in that way at all, and there is no particular reason that there should be any relationship at all, and it is simply an empirical fact (if it is so) that the two pressures should correlate, and we use the data we gather to determine what the correlation is.

The Instrument: Sphygmomanometer

Sphygmomanometer is a strain gauge rigged to turn mechanical pressure into a rise or fall of mercury in a tube. That was an arbitrary but convenient choice. The tube could have contained any other visible liquid. The point is that the pressure of blood on the arterial wall (actually on the tissue within which the artery is embedded) has to be transduced in some way so it can be converted into a useful metric. Blood pressure per se has nothing to do with millimeters of mercury. In fact, because of risks of toxic exposure, mercury-tube sphygmomanometers are disappearing from use in this country today. Yet, the metric for registering blood pressure is still related to the height of a mercury column. Blood pressure devices being sold today depend on one or another of several different mechanisms for detecting physical changes associated with arterial pressure and for transducing the signal to produce number calibrated in terms of millimeters of mercury, e.g., even if the output is in the form of a pointer on a

dial or a digital readout. These various devices yield values that are highly related to one another in correlational terms, but they are not all equally good "measures" of blood pressure. That is because, although the values from two devices may be highly correlated, the absolute differences between estimates of blood pressure may be sufficiently large as to imperil correct interpretation of the findings.

The foregoing are details of the measuring device (instrument), somewhat like variations in the ways in which personality scales might be presented or responded to. The results from scales with the same name presumably are intended to be equivalent—actually *equal* in the case of blood pressure since the metric is fixed. The equivalent for most social science measures is to fix the metric by converting scores into percentiles or some such. That metric depends for its interpretability, although, on the equivalence of the populations on which norms are established or on whether it makes sense to compare scale values of a person with those of other persons . . .

Blood Pressure Is Estimated

It is important to remember that under most circumstances blood pressure is not measured directly. As directions for one set of devices note, "When you take a patient's blood pressure, you're measuring the pressure in the cuff—only indirectly are you measuring the pressure in the blood vessel." That same principle applies even more cogently to most measures in psychology and other social sciences. Characteristics of persons in which social scientists are interested are rarely measured directly but must be inferred from indicators, most of which are much more tenuously linked to underlying psychosocial constructs than cuff pressure is to blood pressure.

The relationship between the way the measure of blood pressure is structured and the underlying variable is reasonably transparent, but it is to some extent arbitrary, and a standard measure of blood pressure could have taken some other form. The average of blood pressures for standing, sitting, and supine positions could have become standard. Alternatively, an average of morning and evening blood pressures or from the two arms, or blood pressure taken after moderate exercise could have become standard. With modern equipment and computers it is possible to have a measure of average blood pressure over time, a measure of maximum systolic blood pressure, of resting blood pressure, fasting blood pressure, stress-induced blood pressure. In fact, devices and arrangements are available to make such measurements. Convenience and expense, however, dictated the directions taken over many

decades in developing the standard way of measuring blood pressure. It is worth keeping in mind that blood pressure is *not* merely a screener, the results of which are necessarily to be checked by a series of other, more precise measures. People are put on antihypertensive medications *solely* on the basis of blood pressure readings taken in a doctor's office, maybe during only a single visit.

In fact, the sphygmomanometer, used by a properly trained person is often considered the *gold standard* for blood pressure, e.g., in evaluating home blood pressure devices.[2] The provision that it should be used by a properly trained person is an essential codicil in the definition of the proper measurement of blood pressure. Validity is not invariably and simply a property of an "instrument." Rather, validity must be considered to inhere in a system or process of which the instrument itself is only a feature. An enlightening view of reliability argues persuasively that reliability is a characteristic of *data* not measures (Thompson and Vacha-Haase 2000). The same is necessarily true of validity, if for no other reason than that validity of data is limited by its reliability. Is a tape measure a valid measure of length of pieces of lumber? Only if the person using the tape measure understands its use and follows the usual conventions of its application. Is the Hamilton Rating Scale for Depression (Hamilton 1967) a valid measure of depression? It can be considered so if it is properly used, but not otherwise.

One of my friends, a diabetologist, will measure blood pressures of his patients only after they have sat quietly in the waiting room for 20 minutes. He has noted that blood pressure readings are affected by the ambient temperature outside, by whether and how far patients have walked across parking lots and campus to get to his office, and by other activities. He questions them about such matters as how much coffee or other stimulant they may have taken and when they took it before coming to his office. His use of the instrument is quite unlike that of many other clinicians. Maybe (or maybe not) he produces more reliable, and hence more valid, blood pressure readings than those of other clinicians. So what is *the* validity of the measuring instrument?

Blood Pressure as a Latent Variable

Blood pressure as measured is widely recognized by physicians and other medical personnel—but by no means by all of either group—as a latent variable, even though they have not much idea at all of what a latent variable is. That is, they recognize that measured blood pressure is one fallible indicator of

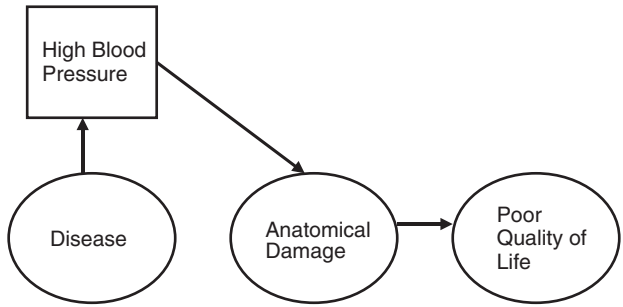
a construct that is not easily, if at all, accessible. Physicians know, for example, that blood pressure measured in the doctor's office may not be the same as blood pressure measured elsewhere—the "white coat" phenomenon (e.g., Godwin et al. 2004)—let alone being the same as "walking around" blood pressure. Probably many physicians seeing a patient will measure blood pressure more than once, but very few will adhere to the recommended standard that blood pressure should be measured in standing, sitting, and supine positions—too much trouble, one supposes. These observations do suggest, however, that medical personnel realize that the blood pressure readings they get are merely indicators of some hypothetical, unknowable blood pressures characterizing any given patient. That is, we assume that each person is characterized by some "real," underlying blood pressure, e.g., perhaps conceived as the 24 hours average of all possible systolic pressures (and the same for diastolic pressures or pulse pressures). What we get from measurements is one or more fallible indicators of that underlying reality.

The constructs of diastolic and systolic blood pressures are not as well delineated as they might be, at least in terms of their implications. That is in part because their implications are not completely understood. After more than 100 years of measuring blood pressures, their construct validity is still open to doubt, in large part because the focal construct is often not explicit. In recent decades it has gradually come to be accepted that systolic blood pressure may have implications for health problems different from those of diastolic blood pressure (e.g., Pocock et al. 2001). It may be that the level of the maximum pressure on the artery walls is more important than the level of the minimum pressure. Apparently, it still is not known whether sharp spikes of blood pressure may be more important than levels averaged over time. Recent articles have altered interpretations of blood pressures once considered in the "high normal" range so that they are no longer regarded as healthy.

Blood Pressure Is an Indicator of an Even More Latent Variable

Blood pressure, even high blood pressure, is just that. It is of interest only as it relates to some underlying disease process and, hence, to some implications for the patient and his or her health and welfare. That is why wise clinicians take into account such factors as exercise, temperature, time of day, coffee consumption, and so on. The "model" that is assumed is (more or less) as shown in Figure 1.

Figure 1: Model for Disease, Blood Pressure, and Outcomes



Disease causes high blood pressure—that may be the only detectable indicator of the disease—and high blood pressure produces anatomical damage in the circulatory system that results in poor health and poor quality of life (QoL). That is the basic *casual* (not necessarily causal) model (Rogosa 1987). Every one-point increase in diastolic blood pressure is accompanied by a 2 percent increase in risk of stroke or heart attack (Pocock et al. 2001). But is blood pressure the "cause" of disease or is blood pressure, along with other things, a cosymptom of an underlying disease? Or either, or both? High blood pressure may damage kidneys, whatever the reason may be for the high blood pressure. On the other hand, various diseases, e.g., atherosclerosis, may give rise to high blood pressure. So-called "essential hypertension" may (or may not) be a disease in and of itself.

The relationship between measured blood pressure and disease is, in any case, far from perfect, unless one simply defines high blood pressure itself as a disease, and the relationship between disease and anatomical damage is certainly quite modest. Moreover, the relationship between anatomical damage and reduction in QoL is probably only modest. This long chain of relationships means, then, that the "validity" of blood pressure as a predictor of anything very far downstream is quite modest. Most people with high blood pressure do not experience any obvious bad consequences. After all, a 100 percent increase in risk of heart attack may only raise the chances from 1 in 100 to 2 in 100 (a "100 percent increase").

"The Validity" of Sphygmomanometers

The validation of new blood pressure devices is almost invariably in terms of their agreement with accepted versions of sphygmomanometers. As noted, the sphygmomanometer has become, if not the gold standard, at least the

default option for blood pressure measurement, even though that method cannot in any sense be regarded as having final validity, that is in some ultimate sense of measuring "real" blood pressure. Medically it would be preferable to have a direct measure of arterial blood pressure, but that requires an invasive and somewhat risky procedure that cannot be justified except on fairly critical medical grounds, e.g., the need for direct measurement of blood gases. Does a finger cuff gain construct validity by being shown to produce blood pressure readings that correlate with those made by an arm cuff sphygmomanometer? Probably so, but extending that "gold standard" notion to validation of ability tests or personality measures is questionable. Should we believe that a measure "has" construct validity simply because it is correlated with some better-established measure? We might accept the proposition for a finger cuff because it operates by the same general principles as the arm cuff. But suppose some clever person devised a self-report measure of "blood pressure sensations" that correlated 0.60 with systolic sphygmomanometer readings? Such a measure would be "valid" (for up to 36 percent of the variance), but would we want to grant it "validity" as a measure of the systolic blood pressure construct? The mechanisms might seem just too different and the shared variance too small. As Borsboom, Mellenbergh, and van Heerden (2004) insist, we may want to avoid extending the concept of "measure" to variables that are merely correlates of a construct.

It is true that a measure has predictive utility (Borsboom, Mellenbergh, and van Heerden 2004) to the extent that it correlates with some other variable of interest, e.g., a criterion. So, in a sense, age has predictive utility for blood pressure, although the relationship is not large; i.e., the variance between is small in comparison with the variance within age groups, but it is "valid" nonetheless. A predictor, however, need not share in any essential way in the meaning of the construct underlying the predicted variable. So, self-reported "blood pressure sensations," even if correlated with actual blood pressure, might not in any direct sense reflect the construct "blood pressure." Scores on the "blood pressure sensations scale" might arise from, for example, knowledge that subjects might have about the relationship between weight or age and blood pressure, from family history of blood pressure, from a sense of being excited, and so on.

Validity of Blood Pressure Estimates

In short, it is a very difficult matter to know quite what is meant by "the validity" of a sphygmomanometer, let alone be able to attach a numerical

value to whatever we mean by validity. In all probability, the correlation between measured systolic or diastolic blood pressure and the momentary actual pressure of blood on the arterial walls is high. On the other hand, the correlation between measured blood pressure and what we must think of as a latent variable, "real" blood pressure, is not known and cannot be known. We can only assess the utility of measured blood pressure in predicting conditions or events that we need to know about, such as cardiovascular accidents, kidney failure, and so on.

Despite reservations such as those just expressed, and their justification seems strong, measured blood pressure has such a definite and elaborated theoretical underpinning, it is related with sufficient consistency to a wide range of health problems, and it is, under proper conditions, measured with such dependability, that its construct validity (that of *measured blood pressure*, not of the sphygmomanometer) can be regarded as well established and substantial. Note, however, that no numerical value can be attached to the estimate of construct validity. No one can say how valid the construct itself is, let alone any measure of it.[3]

Validity of Social Science Measures

The long digression on the measurement of blood pressure and its validity was intended to illustrate the complexity of the problem of validity in the context of a measure that is widely accepted and whose validity is scarcely ever questioned. If it is possible, indeed necessary, I think, to raise questions about the meaning and extent of validity of measured blood pressure, then even more serious questions are unavoidable in relation to measures commonly used in social science.

Suggestions are made from time to time about the desirability of having strict standards for the validity of psychosocial measures. For example, Messick's scheme for conceptualizing validity might be taken as a starting point. Or so might the Joint Standards (AERA, APA, NCME 1999). Any such standards would probably be unworkable in light of the difficulties that are inherent in specifying the relevant constructs, which would, of necessity be latent and, therefore, only measurable in principle. The various aspects of validity may not always be equally applicable or assessable, and it would be quite a task to quantify any of them.

It is reasonably easy to decide that commercial test publishers should be responsible for demonstrating "the validity" of their tests, whatever that might

mean. Nearly everyone, after all, is in favor of truth in advertising. Perhaps test publishers might even be held to standards such as those proposed by Messick, although I, and some others who reviewed the Joint Standards (AERA, APA, NCME 1999), have serious reservations about the applicability of the idea of consequential validity.

The situation is very different, however, for measures that are developed for scientific purposes and that are not meant to be commercialized. Major test companies spend thousands, even hundreds of thousands, of dollars on the development of the measures they publish. No money at all might be available for an investigator who wished to develop a "Hope for Recovery Scale" to be used with cancer patients. (And just how would the validity of such a construct/scale ever be firmly established?) A large proportion of all of our research measures are developed like open source software, as the result of iterative contributions of multiple scientists interested in and willing to work in independent collaboration to probe and improve the measure—and the underlying construct. That also means, of course, that development of our measures is not likely to be systematic and optimal.

Measurement efforts also benefit from the processes encompassed by the idea of "validity generalization" proposed by Schmidt and Hunter (1977). Through multiple instances of use, measures gradually come to be accepted as "valid," even for new populations, in new contexts, and for new purposes. Each time a new measure is developed and shown to be useful, that contributes in at least some small way to the presumption of validity for similar measures, even ones not yet developed. If there is one thing that psychologists, and to some extent other social scientists, have learned how to do, it is to develop "scales." We know how to write items and put them together in ways that produce results that are from modestly to highly dependable (reliable), even if their validity for the intended construct cannot be guaranteed.

From the standpoint of measurement, the social sciences are different in one important way from the harder sciences, viz., in their inclination to develop measuring instruments and then set out to find out what, specifically, they are measures of, a process lamented by Borsboom, Mellenbergh, and van Heerden (2004). Nonetheless, to some extent that is inevitable in the softer sciences. The social sciences have an extraordinarily wide range of constructs, many of which cannot be quite exactly defined, let alone be exactly represented in measurement operations. It is necessary, then, once measures are proposed, developed, and in use to continue efforts to understand them and their relationships to other measured variables and underlying constructs that can only be inferred. The SF-36, for example, was developed over a long

period of time and with great care, but elaboration of understanding of just what it measures and for what it might be useful has been equally gradual, with progress having been, in all probability, much greater since its publication and wide adoption than in all the preceding years of its formulation.

Social Science Constructs

Evidence for the construct validity of a measure is only as acceptable as the construct is acceptable. Many people appear not to "believe in" the construct of "general intelligence." If they do not, then it is futile to present evidence to them for the construct validity of a general intelligence measure. In a very important sense, it is true that "intelligence is what intelligence tests measure." Virtually all (maybe really all) psychological constructs are just that, constructs, and they have no independent, verifiable reality beyond the specifications of their definitions and the operations proposed for measuring them. Therefore, no hard, absolute evidence for the construct validity of any psychological measure can be produced. Acceptance of the construct validity of a measure requires acceptance of the reasonableness or "truth" of the construct itself. Hence, test developers can only appeal to consensus in the community of psychological theorists and researchers in summarizing the evidence for the "validity" of their measures. (That applies, however, only to construct validity, for empirical or predictive validity can be convincingly supported by simply showing a correlation of useful magnitude between a potential predictor and some variable of more fundamental theoretical or practical interest.)

Construct validity of measures can be established only incrementally as evidence accrues that the measures are related in theoretically interesting ways to a range of other measures and phenomena. In fact, as Loevinger (1957) persuasively argued nearly half a century ago—demonstrating our slow uptake of truly seminal ideas—theory and measurement are mutually intertwined, for our attempts to measure constructs almost always can help us better to understand and revise our constructs. There can be no point at which, abruptly, construct validity can be said to have been established; construct validity is a matter of more or less. In addition, as noted, evidence for construct validity will be accepted by critics only to the extent that they "buy into" the construct in the first place. Hence, the status of any measure of any construct in the social sciences will always be uncertain. Could one have a "valid" measure of, let us say, QoL? Only to the extent that a potential audience believes (1) that the construct had been defined in a satisfactory way, (2) that the measure

seems to capture what is implied by the definition, and (3) that scores on the measure are related to broader phenomena implied by the idea of QoL. Could one have a "valid" measure of poverty? Of the cost of living? Of anxiety? The same strictures apply.

Return to the example of blood pressures. Whether a conventionally constructed instrument to measure diastolic and systolic blood pressures is a "valid" measure is to some extent beside the point: blood pressure is what a sphygmomanometer measures. Blood pressure as measured has reasonably good empirical/predictive validity. High values of blood pressure are predictive of the occurrence of stroke, heart attack, and kidney disease. But the latent variable of "blood pressure health" is another matter. That is, whether we can say more about a person with "high" blood pressure than that, statistically, he or she is at increased risk of some specified list of bad health outcomes could certainly be questioned.

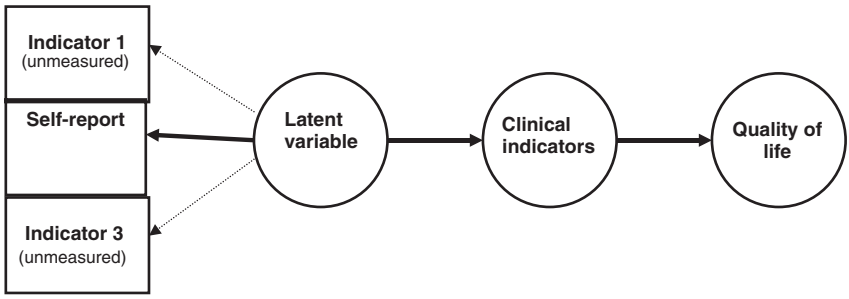
Blood pressure is, as we know, only modestly correlated with health outcomes such as incapacitation and death. I do not know what the correlation might be; it would vary according to the ages of samples, for one thing. But although the odds ratios may be 2.0 or so, the correlations are almost certainly no greater than 0.1 or 0.2. All the bad outcomes depend on many other things such as the elasticity of arterial walls and comorbidities. So is blood pressure a "valid" measure of health or ill health? I do not think we can say that. It makes a lot of difference whether we think about the validity of measures in terms of their correlates (predictive power) or their theoretical coherence (construct validity).

Constructs and Self-Reported Constructs: Maybe Not the Same

A fundamental limitation on measurement in social science, and certainly the limitation applies strongly to psychology, is the extensive reliance on self-reports. That reliance is, it seems, unavoidable. But that reliance adds another layer of complexity to considerations regarding validity of measures. The required, but not always explicitly recognized, model is as shown in Figure 2.

Briefly, the model is that QoL is affected by Clinical Outcomes, which are affected by a Latent Variable that has three (and potentially many) indicators, but of which only one, Self-Report, is measured. To give this example some substance, let us suppose that the latent variable is "Severity of Con-

Figure 2: Latent Variable Model



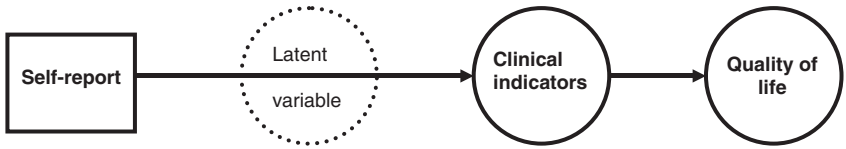
dition" and that the patients in the study are to receive some treatment the results of which should depend somewhat on initial severity.

In fact, however, under these circumstances, the Latent Variable, Severity, is transformed into a single measured variable that is interpreted as if it were the latent variable of "Severity." The recollection that it is really only a measured variable is lost. The Latent Variable, which we would think of as "Real, underlying severity," disappears from the model (Figure 3).

This example is hypothetical but not unrealistic. It is very common in social science research to find that investigators make statements such as "We included a measure of . . .," or "Our measure of . . . was . . .," when what they refer to is a single, self-report *indicator* that often cannot possibly capture the richer meaning that is intended by the construct in question.

Self-reported severity might be expected to reflect that actual severity of the condition (the latent variable) of course, but also personal tolerance for pain, comparisons with other persons, stoicism, desire to be perceived as strong, desire for sympathy, and so on. It is not that self-reports do not include variance from the variable of interest, it is that they almost always include so many other things.

Figure 3: Self-Report Model with Latent Variable Omitted



Therefore, even if it is generally true that psychologists and other social scientists can readily come up with scales to "measure" all sorts of things, reliance on self-reports can, and probably usually does, alter the construct involved in ways that are important for the research and conclusions drawn from it.

Some Recommendations for the Time Being

So, if we do not have a useful framework for construct validity, what ought we to do? The most important thing by far is to promote generally better understanding in the research community and in all its audiences of the nature of measurement, the special problems posed by measurement in the social sciences, and the inevitable limitations on inferences in science (so that results of any sort are not overinterpreted). How to do that promotion is worth consideration, e.g., by the Measurement Excellence and Training Resource Information Center (METRIC, www.measurementexperts.org) established by the Veterans Administration Health Services Research and Development Program. In the meantime, and in addition, we have some protections that we should cling to closely—let us say in the name of safe measurement.

1. We very much need to advocate for, lobby for, and assist in the improvement of training in measurement in all our programs and in professional development and continuing education efforts. That improvement would be most affected by the restoration of concern for measurement, which has clearly declined very seriously over the past four or five decades, at least, and most grievously, in psychology. Knowledge about and general understanding of measurement has increased greatly during that same five decades, but it is not being recognized and capitalized on in training programs (Aiken et al. 1990).
2. We are protected somewhat by a general consensus about many measurement questions, and we should capitalize on that consensus. Most researchers most of the time think about measurement in very conventional ways and use conventional measures. That may not always be optimal, but it probably protects us against more egregious errors. We might be even better protected if we developed and promulgated consensual views about what is good measurement and what are good measures. For example, great benefit might come from

"white papers" that would formulate what is the best current thinking about different topics. If those papers are written for a general audience, rather than for those sophisticated in measurement already, they could provide a justification for investigators to "do the right thing." It might be additionally helpful to have a series of consensus statements presenting the best current thinking about measurement of specific constructs, e.g., depression, patient satisfaction, and so on. Not so much which instruments to use, but how to set about understanding the construct and approaches to its measurement. Such consensus summaries might be seen as stifling creativity, but that, I think, is not a serious risk. In any case, we do not want people who do not know what they are doing to be leading the way.

3. As part of the consensus process, we need to think about some ways of stipulating the construct validity of specific existing measures for specific purposes, e.g., perhaps expert panels. If construct validity is a complex matter, then it follows that concluding something about construct validity is also complex. But we do not want every author who elects to use, say, the Beck Depression Inventory (Beck et al. 1961) or the SF-36 (Ware, Kosinski, and Gandek 1993), to do a separate analysis and exposition of evidence for validity. Journal space is precious, and so is reading time.
4. We should pay close attention to the conditions and context in which measurement occurs. It might be a good idea, for example, to develop protocols for the use of at least some instruments, with the protocols specifying contextual variables, conditions under which measurement is carried out, instructions or preparation given to subjects, and training of persons involved in the measurement process. The protocol described by Goldman (2002) for diagnosis of hypertension provides a good example. With few exceptions for published measures of intelligence, achievement, and personality, one will look in vain for protocols for the administration of almost any measure in the social sciences. Perhaps there might be several alternative protocols relevant to different research situations. Investigators could then choose the protocol most appropriate to their interests and capabilities, they could refer to those protocols in describing the methods for their study, and they could describe departures from protocols made necessary or desirable by the particular circumstances of their projects. At least we would know a good bit about how measurement was actually carried out, which is not true under present research conventions.

5. Peer review is an important protection in all aspects of science. The protection afforded by peer review, however, depends on how well informed peer reviewers are. Special efforts should be made to reach out to peer reviewers and offer assistance to them. That assistance might take special forms. One would be to ensure that peer reviewers (peer review panels, editorial panels) are offered early access to materials, perhaps even being offered special versions of papers that are more oriented toward review rather than toward the doing of research, e.g., issues to be raised, questions to be asked, qualifications to be expected. Another form of reaching out that might be tried is the offer of consultation on measurement issues. For example, if, in the course of reviewing proposals a review panel develops a sense of something they need to know, they might submit their concerns to some individual or group consultant for a response to be delivered before their next set of reviews. A panel might ask some question such as "What is the difference between Rasch and item response theory analyses and why should we care?" Or, "Investigators sometimes claim that a reliability of 0.65 or something of that sort is 'sufficient' or 'acceptable,' but others sometimes say that a reliability of 0.85 is 'required.' What should be the standard for reliability?" (Not that there is a standard.)
6. We can and should rely on postpublication critique in science. Very few "findings" in science get at once absorbed into the current packets of wisdom about important matters. Generally, the more important the issue, the slower that rate of absorption, and there is sufficient time to bring to bear the collective expertise of the field on problems at hand. We do, of course, need knowledgeable and dependable critics to carry out those tasks of criticism. One of the useful initiatives of editors and professional groups might very well be to foster the preparation of critical review papers related to measurement issues, practices, or instruments.
7. Finally, and without being exhaustive, we need very much in the social sciences to promote "critical multiplism" as a hallmark of our science. This is not the place to try to explain critical multiplism; excellent and provocative expositions are available and should be consulted by all social scientists (Cook 1985; Shadish 1994). Basically, although, in relation to measurement, critical multiplism requires that we not rely too heavily on any one measure or any one measurement procedure.

In summary, we social scientists will do ourselves a favor and our audiences a favor if we think more about, do more about, and write more about the validity of the data we produce and less about the validity of specific instruments.

ACKNOWLEDGMENTS

The author would like to thank members of the Evaluation Group for Analysis of Data (EGAD) for their helpful comments on an earlier draft of this paper. Thanks are also owed to two anonymous reviewers who provided helpful suggestions for revisions.

NOTES

1. Roughly: content = the nature of the construct; substantive = the theory underlying it; structural = the proper relationship between the scoring and the construct; generalizability = populations, settings, tasks, etc., to which results apply; external = convergent and discriminant validity; and consequential = legitimacy of conclusions and interpretations from the standpoint of values and ethics.
2. The number of different devices for measuring blood pressure is very large, and the literature comparing the results obtained from various ones under various conditions is voluminous; hundreds of such studies may be found in the literature.
3. Westen and Rosenthal (2000) proposed two summary correlational indexes of the magnitude of construct validity, but they require data and judgments that may not often be realistic, and their usefulness remains to be demonstrated.

REFERENCES

Aiken, L. S., S. G. West, L. Sechrest, and R. R. Reno. 1990. "Graduate Training in Statistics, Methodology, and Measurement in Psychology." *American Psychologist* 45 (6): 721–34.

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

Baccellis, G., P. Reggiani, A. Mattioli, E. Corbellini, S. Garducci, M. Catalano, and S. Omboni. 1997. "Hemodynamic Changes in the Lower Limbs during Treadmill Walking in Normal Subjects and in Patients with Arteriosclerosis Obliterans." *Angiology* 48 (9): 795–803.

Beck, A. T., C. H. Ward, J. Mock, and J. Erbaugh. 1961. "An Inventory for Measuring Depression." *Archives of General Psychiatry* 4: 561–71.

Borsboom, D., G. J. Mellenbergh, and J. van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111: 1061–71.

Cook, T. D. 1985. "Post-Positive Critical Multiplism." In *Social Science and Social Policy*, edited by L. Shotland and M. Mark, pp. 21–62. Beverly Hills, CA: Sage.

Cronbach, L. J., and P. M. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52: 281–302.

DeStefano, A. L., M. G. Larson, G. F. Mitchell, E. J. Benjamin, R. S. Vasan, J. Li, D. Corey, and D. Levy. 2004. "Genome-Wide Scan for Pulse Pressure in the National Heart, Lung and Blood Institute's Framingham Heart Study." *Hypertension* 44 (2): 152–5.

Engvall, J., C. Sonnhag, E. Nylander, G. Stenport, E. Karlsson, and B. Wranne. 1995. "Arm-Ankle Systolic Blood Pressure Difference at Rest and after Exercise in the Assessment of Aortic Coarctation." *British Heart Journal* 73 (3): 270–6.

Godwin, M., D. Delva, R. Seguin, I. Carson, S. MacDonald, R. Birtwhisltle, and M. Lam. 2004. "Relationship between Blood Pressure Measurements Recorded on Patients' Charts in Family Physicians' Offices and Subsequent 24 Hour Ambulatory Blood Pressure Monitoring." *BMC Cardiovascular Disorders* 4 (1): 2.

Goldman, A. G. 2002. "The Diagnosis of Hypertension." *Journal of Clinical Hypertension* 4 (3): 166–80.

Hamilton, M. 1967. "Development of a Rating Scale; for Primary Depressive Illness." *British Journal of Social and Clinical Psychology* 6 (4): 278–96.

Kay, L. E. 1998. "Accuracy of Blood Pressure Measurement in the Family Practice Center." *Journal of the American Board of Family Practice* 11 (4): 22–8.

Loevinger, J. 1957. "Objective Tests as Instruments of Psychological Theory." *Psychological Reports* 3: 635–94.

Messick, S. 1989. "Validity." In *Educational Measurement*, 3d edition, edited by R. L. Linn, pp. 13–103. New York: Macmillan.

———. 1995. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist* 50: 741–9.

Parati, G. 2004. "Blood Pressure Measurement in Research and in Clinical Practice: Recent Evidence." *Current Opinion in Nephrology and Hypertension* 13 (3): 343–57.

Pocock, S. J., V. McCormack, F. Gueyffier, F. Boutitie, R. H. Fagard, and J-P. Boissel. 2001. "A Score for Predicting Risk of Death from Cardiovascular Disease in Adults with Raised Blood Pressure Based on Individual Patient Data from Randomized Controlled Trials." *British Medical Journal* 323 (7304): 75–81.

Rogosa, D. R. 1987. "Casual Models Do Not Support Scientific Conclusions." *Journal of Educational Statistics* 12: 185–95.

Schmidt, F. L., and J. E. Hunter. 1977. "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62: 529–40.

Shadish, W. 1994. "Critical Multiplism: A Research Strategy and Its Attendant Tactics." *New Directions for Program Evaluation* 60: 13–57.

Thompson, B., and T. Vacha-Haase. 2000. "Psychometrics Is Datametrics: The Test Is Not Reliable." *Educational & Psychological Measurement* 60 (2): 174–95.

Ware, J. E., M. Kosinski, and B. Gandek. 1993. *SF-36 Health Survey: Manual and Interpretation Guide.* Lincoln, RI: QualityMetric Incorporated.

Westen, D., and R. Rosenthal. 2000. "Quantifying Construct Validity: Two Simple Measures." *Journal of Personality and Social Psychology* 84: 608–18.