Measurement Validity in Physical Therapy Research
Julius Sim and Peggy Arnell
PHYS THER. 1993; 73:102-110.

Measurement Validity in Physical Therapy Research

Julius Sim
Peggy Arnell

This article considers the role of measurement validity within physical therapy research. The concept of measurement validity is identified as a component of internal validity, and it is differentiated from the notion of reliability; these concepts are related to systematic and random sources of error, respectively. Using examples from physical therapy and rehabilitation, four main types of validity are reviewed: face validity, criterion-related validity, content validity, and construct validity. The differing implications of these types of validity for quantitative and qualitative research are discussed. Three principal areas of concern are then addressed, based on a critical discussion of selected examples from the literature. First, it is argued that validity is often poorly distinguished from the allied concept of reliability and that purported claims for validity often only demonstrate reliability. Second, it is claimed that validity is too often neglected in favor of reliability, and specific examples relating to gait analysis are put forward to support this argument. Third, some of the methodological difficulties that may occur when attempts are made to demonstrate validity are considered. The article concludes with a plea for a closer focus on the issue of measurement validity within physical therapy research. [Sim J, Arnell P. Measurement validity in physical therapy research. Phys Ther. 1993;73:102–115.]

Key Words: *Research; Research design; Tests and measurements, general.*

The concept of validity is central to the research process. The precise meaning of this concept, however, is often not fully grasped. Loosely used, the term "validity" may suggest that a piece of research is relevant, that it is worth doing, or that its results are valuable. Such nontechnical uses of the term are perfectly intelligible and should not be dismissed out of hand, but it is important for the producers and consumers of research findings to be aware that the strict meaning of the word, in terms of research methodology, is more specific. It is the aim of this article to elucidate this meaning. We will highlight some of the problems that can result when the true nature of validity is not appreciated (in particular, when it is not sufficiently differentiated from reliability), and we will briefly discuss methodological difficulties that may be encountered when seeking to determine measurement validity.

Although validity is a crucial consideration in both quantitative and qualitative research,[1,2] the specific issues to which it gives rise often differ. Some of these differences will be discussed in this article.

The focus of this article will be on the importance of measurement validity within research. Nonetheless, it is implicit throughout the article that the same importance attaches to the validity of clinical measurements and that the same fundamental principles apply in both spheres of professional activity. Both the clinician and the researcher seek to draw certain inferences from the measurements they take, and the validity of these measurements should be of equal concern in each case.

Validity and Reliability

In relation to research, there are two broad types of validity: (1) external validity and (2) internal validity. *Exter-*

J Sim, PT, is Principal Lecturer in Health Sciences, School of Health and Social Sciences, Coventry University, Priory St, Coventry CV1 5FB, United Kingdom. Address all correspondence to Mr Sim.

P Arnell, PhD, PT, is Manager for Postgraduate Medical Education, Department of Postgraduate Medicine and Dentistry, University of Manchester, Stopford Building, Oxford Rd, Manchester M13 9PT, United Kingdom.

This article was submitted March 3, 1992, and was accepted September 14, 1992.

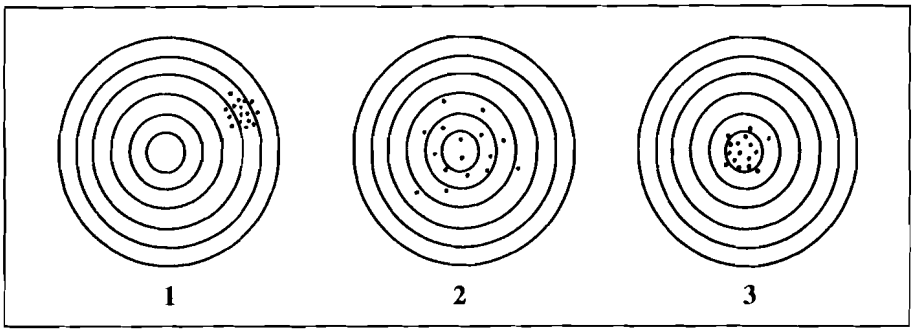

Figure 1. *Target-shooting analogy for validity and reliability: (1) Scores (ie, measurements) can be highly reliable, but not necessarily valid; (2) scores can be somewhat valid, yet have low reliability; (3) scores can be both highly valid and reliable.*

nal validity refers to the extent to which the findings of research conducted on a sample can be generalized to the population from which the sample was drawn. The notion of *internal validity,* in contrast, refers to "the possibility that the conclusions drawn from experimental results may not accurately reflect what has gone on in the experiment itself."[3(p221)] As Polgar and Thomas state,

> If a study is internally valid, this means that any effects/changes or lack thereof in the dependent variable can be directly attributed to the manipulation of the independent variable.[4(p146)]

Internal and external validity relate to the validity of findings, in a general sense. The focus of this article, however, is on a particular facet of internal validity, namely, the validity of measurements. Measurement validity relates to the extent to which an instrument measures what it is intended to measure and can refer equally to nominal, ordinal, interval, or ratio levels of measurement. The question we pose when considering measurement validity is: "Are we measuring what we think we are measuring?"[5(p457)] That is, we are interested in the *relationship* between the instrument and the entity it is supposed to measure and in the "degree to which a useful (meaningful) interpretation can be inferred from a measurement."[6(597)] It follows from this that validity only has meaning within a specified context and is not an inherent property of an instrument. As Payton points out: "A measurement tool is never just valid; it is valid for making a particular measurement."[7(p70)] Accordingly, it is more helpful to think of validity as an attribute of a measurement than of an instrument.

The meaning of validity can be clarified by contrasting it with the allied concept of reliability. Reliability concerns the extent to which the instrument yields the same measurement on repeated uses, either by the same operator (intraobserver reliability) or by different operators (interobserver reliability). Reliability relates to the *reproducibility* of measurements, whereas validity deals with the *accuracy* (correctness) of inferences drawn from such measurements. As an example, consider the measurement of a person's weight using a pair of bathroom scales. If, on repeated weighings of the same individual (of unchanging mass), the scales produce the same readings, they can be said to be reliable. This is not to say, however, that they necessarily produce valid measurements; it is quite possible that they were inadequately zeroed at the outset and have therefore been consistently underreading or overreading. Although the readings have been the same throughout the weighings (and thus reliable), this does not in itself demonstrate their validity; it does not tell us whether the readings are a true representation of the entity in which we are interested (ie, the weight of the individual). This would require either that additional, independent information be obtained as to the individual's weight or that the scales be calibrated against an object of known mass. In other words, reliability does not presuppose validity.

The relationship between validity and reliability can be further illustrated by analogy with target-shooting.[5] The first target in Figure 1 illustrates the fact that scores (ie, measurements) can be highly reliable, but not necessarily valid. The second target shows that scores can be somewhat valid even if of low reliability. To obtain a *high* degree of validity, we would wish to see the situation depicted in the third target. As the accuracy of the scores improves, so in turn does their reproducibility; that is, validity largely presupposes reliability.[4(p144)] If an instrument yields unreliable measurements, this will set limits to their validity.

Validity and the Concept of Error

Further inspection of Figure 1 reveals something of the relationship between validity and the concept of error. In the first target, the scores are closely grouped, but are some distance from the bull's-eye. This is the result of systematic (nonrandom) error; there is a consistent pattern to the inaccuracy illustrated. In contrast, the error in the second target is random; there is no consistent pattern to the way in which the scores are dispersed. Systematic error is responsible for bias. Random error tends to be self-compensating—errors tend to cancel out one another—and thus does not introduce bias. Random error, however, makes inference difficult by obscuring patterns or relationships and may therefore necessitate a larger sample in order to reveal them.[8]

In order to improve validity, attempts must be made primarily to remove systematic error, or bias. Conversely, in order to improve reliability, attempts must be made to remove random error. The main differences between validity and reliability are summarized in the Table.

Table. *Differentiating Characteristics of Validity and Reliability*

Validity	Reliability
Deals with the accuracy of inferences made from measurements	Deals with the reproducibility of measurements themselves
Concerns the relationship between the measurement and the entity being measured	Is a property of the measurement (and the person performing it)
Requires independent knowledge of the "true" value of the entity being measured	Is not dependent on the "true" value of the entity being measured
Presupposes a certain degree of reliability	Does not presuppose validity
Is undermined by systematic error	Is undermined by random error
Liable, if lacking, to distort or bias relationships among variables	Liable, if lacking, to obscure relationships among variables

Types of Validity

Writers on the subject of validity have identified four principal types of validity: face validity, criterion-related validity, content validity, and construct validity.[5,9,10] In order to clarify their meaning, each of these types will be examined in turn, using examples drawn from the fields of physical therapy and rehabilitation.

Face Validity

Face validity is a very rudimentary form of validity. It concerns the extent to which a test or measure *appears* to measure what it purports to measure. Thus, it is "really based on the personal opinions of those either taking or giving a test."[10(p17)] Indeed, the importance of face validity relates most significantly to the perceptions of research subjects. For example, a questionnaire might be developed to measure client satisfaction with the physical therapy service provided in a hospital. If the questionnaire included a number of items that appeared, to the respondents, to have little relevance to the stated purpose of the survey, it is likely that the quality of responses would be adversely affected. This might be because of a certain amount of confusion on the part of respondents as to the exact responses required or because the perceived irrelevance of the items caused them to make little effort to answer thoughtfully.

Some writers, either explicitly[4(p145)] or implicitly,[5] identify face validity with content validity. Despite similarities between the two concepts, however, they are more usefully considered separately, not least because face validity is less amenable to formal scientific testing.

Criterion-Related Validity

Criterion-related validity is the type of validity that frequently underpins quantitative research. Evidence for criterion-related validity is obtained by comparing the readings or measurements obtained by the investigator with a measurable criterion that is accepted as a standard indicator of a concept or variable. If the instrument gives an accurate representation of the concept or variable, criterion-related validity has been demonstrated.

The American Physical Therapy Association's Task Force on Standards for Measurement in Physical Therapy distinguishes three varieties of criterion-related validity: (1) concurrent validity, (2) predictive validity, and (3) prescriptive validity. *Concurrent validity* exists when "an inferred interpretation is justified by comparing a measurement with supporting evidence that was obtained at approximately the same time as the measurement being validated."[6(p597)] In *predictive validity,* this supporting evidence is gained at a later time, and we are concerned with "the justification of using a measurement to say something about future events or conditions."[6(p597)] Finally, *prescriptive validity* exists when "the inferred interpretation of a measurement is the determination of the form of treatment a person is to receive . . . [and is] justified based on the successful outcome of the chosen treatment."[6(p597)]

Currier,[11(p171)] when discussing criterion-related validity, gives the example of heart rate monitoring by palpation of the radial artery. In order to validate this method as a means of measuring the underlying variable (ie, heart rate), it must be compared with a generally accepted criterion measurement such as cardiac catheterization. Currier points out that, assuming the accuracy of catheterization has been established, the greater the agreement between the two sets of results, the greater the concurrent validity of the palpation technique.

Similarly, Beattie et al[12] have attempted to determine the criterion-related validity of measurements of leg-length difference, utilizing measurements obtained radiographically as the criterion. These investigators qualify their results by pointing out that both measurement methods were used on supine subjects. Therefore, although the mean of two measurements of leg-length difference appears to be a valid measure of the criterion, the tape measure and radiographic measurements may not reflect functional leg-length difference. For example, leg-length difference resulting from structural or biomechanical asymmetries of the foot and ankle—which are only evident in activities such as standing and walking—would not be defined by either measurement procedure. This limitation underlines the need to consider validity within a specific context.

Another example of concurrent validity—assessment of hamstring muscle length—is provided by Gajdosik and Bohannon.[13] In their example, goniometry was the instrument and straight leg raising (SLR) was the criterion for the underlying concept

of hamstring muscle length. It is important to note, however, that although goniometry may be a valid measure of the criterion (SLR), SLR itself is not necessarily a valid indicator for hamstring muscle length. Indeed, Gajdosik and Bohannon[13] suggest that SLR is not a valid indicator of hamstring muscle length because of the pelvic movement that may accompany the SLR test. In addition, SLR may be limited by structures other than tight hamstring muscles.

Thus, establishment of validity should not be seen as an "all-or-nothing" process. Because the criterion used is, in most instances, likely to be less than perfect, any resulting assessment of validity will only be partial. Rose and Barker illustrate this clearly:

> The validity of a questionnaire for diagnosing angina cannot be fully known; the best clinical opinion is subject to observer variation, and even coronary arteriograms may be normal in true cases or abnormal in symptomless people. The pathologist can describe postmortem structural changes, but these may say little of the patient's symptoms or functional state. Measurements of disease in life, whether clinical or epidemiological, are often incapable of full validation.[14(p1071)]

At times, the criterion used for assessing the validity of a measurement may itself be another measurement for which validity has already been established. In an attempt to determine the concurrent validity of Pediatric Evaluation of Disability Inventory (PEDI) scores, Feldman et al[15] tested PEDI scores obtained from a sample of children against scores obtained with the Battelle Developmental Inventory Screening Test (BDIST), which had already been validated.[16]

Establishing criterion-related validity is also an issue in more qualitative research, but it is often considerably less straightforward to identify an appropriate criterion measure. Consider, for example, the predictive validity of information gained from a questionnaire devised to assess patients' suitability for rehabilitation; for such information to be valid, it would have to predict successful rehabilitation. In this example, the questionnaire would be the instrument and the underlying concept would be successful rehabilitation. The criterion chosen might be, for instance, completion of the rehabilitation program. To the extent that scores on the questionnaire predict completion or noncompletion of the rehabilitation program, the questionnaire can indeed be said to possess criterion-related validity. As we observed previously, however, this does not guarantee that the criterion is itself a valid indicator of the underlying concept.

Other types of qualitative research are more challenging. For example, an observational study might be undertaken in a rehabilitation counseling setting in order to analyze practitioners' counseling strategies. In such an instance, the researcher would be concerned not so much with a descriptive account of what occurred during sessions, as with the counselors' underlying purposes (eg, the meaning that they attached to their behavior). No independent criterion for the researcher's inferences is available. One means of attempting to validate an interpretation of the data gathered, however, is for the researcher to present this interpretation to the individuals who were the subjects of the study to determine whether they endorse it. This process is often referred to as "member validation"[17] or "respondent validation."[18] Underlying this technique is the idea that the subject possesses access to information denied to the researcher:

> . . . participants involved in the events documented in the data may have access to additional knowledge of the context—of other relevant events, of thoughts they had or decisions they made at the time, for example—that is not available to the ethnographer.[18(p196)]

Content Validity

In this form of validity, it is necessary to define what is known as the "domain of content" of the concept being measured and then to determine whether this domain of content is adequately covered by the instrument. The more elements within the concept that are actually assessed by the instrument, the greater the instrument's content validity. In sociology, for example, the concept of alienation may be thought of as having a domain of content consisting of five elements: powerlessness, normlessness, meaninglessness, social isolation, and self-estrangement.[19] A valid indicator of alienation should represent or "sample" each of these elements. Similarly, in a rehabilitation context, the concept of fitness may be regarded as having several elements, including strength, speed, stamina, skill, and spirit or motivation.[20] A fitness test that measures only stamina, and ignores strength, speed, skill, and so forth, would insufficiently sample the full domain of content of the concept fitness and would therefore have low content validity.

To return to an earlier example, we can reconsider the relationship between the concept of successful rehabilitation and the criterion used to represent it—completion of the rehabilitation program. It might be felt that there is much more to successful rehabilitation than simply completing the program and therefore that the criterion lacks content validity as an indicator of the underlying concept. The domain of content of successful rehabilitation might be felt to include a variety of additional factors; however, there would not necessarily be a definitive "correct" list of such factors. Thus, whatever criteria were chosen, there would never be a stage at which "total" content validity could be established.

The findings of a study of patients with rheumatoid arthritis (RA) by Bellamy et al[21] illustrate the necessity for rigor when defining the domain of content of the concept and the way in which it is to be measured. Pain, stiffness, and physical function are key measures in clinical trials involving patients with RA. Together, these three measures are commonly regarded as defining the domain of content of disability in this patient group. Bellamy et al were able to

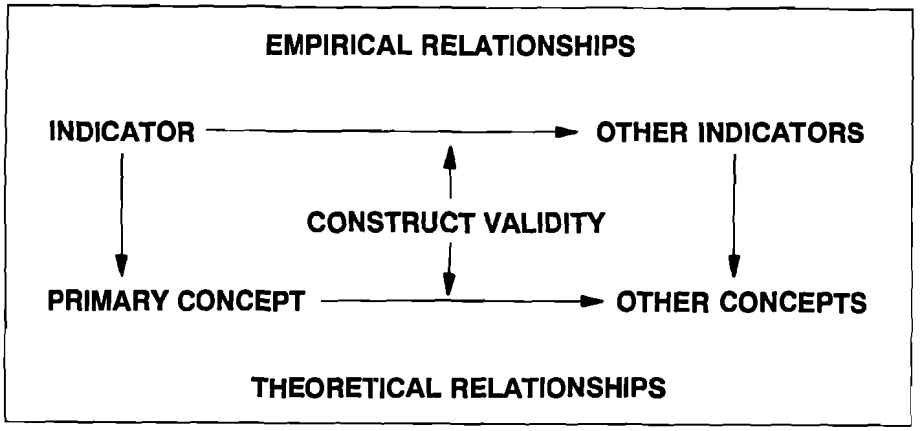

Figure 2. *Framework for establishing construct validity.*

describe significant circadian rhythms, for the group of patients as a whole, for pain, stiffness, and manual dexterity. Such findings clearly have implications for the valid use of these measures as clinical indicators. In particular, Bellamy and colleagues recommend that the time of day at which such measurements are taken be kept constant during the course of a clinical trial. In addition, it is interesting to note that, if there were to be asynchronous fluctuations in the elements within the domain of content, individual elements would be assessed to differing degrees on separate occasions. Accordingly, content validity would be undermined, even though the full domain of content would be consistently represented.

Construct Validity

In this type of validity, which is perhaps most applicable to research conducted within a social science framework, theoretical relationships are established between the primary concept to be measured and one or more other concepts. The researcher then tests the instrument to determine whether it confirms these relationships at the level of systematic empirical observation (Fig. 2). Construct validity, therefore, is demonstrated within a particular theoretical context.[9(p23)]

For example, an index could be developed to measure adjustment to physical disability. Theoretical relationships would need to be proposed between adjustment to disability and other concepts such as self-worth, internal locus of control, optimism, inappropriate help-seeking, and so forth. Thus, one might propose that individuals who have adjusted successfully to physical disability would have a high degree of self-worth, would have an internal rather than an external locus of control, would be optimistic rather than pessimistic regarding their future, and would exhibit little in the way of inappropriate help-seeking behavior. To enable the testing of individuals, indicators would need to be identified for each of these allied concepts and then measured by means of other instruments (preferably instruments that had already been validated).

It would be hoped that the *theoretical* relationships between adjustment and these allied concepts would correspond to the *empirical* relationships between the instrument used to measure adjustment and the indicators representing the other concepts. If, however, the empirical relationships between indicators did not reflect the underlying theoretical relationships, construct validity could not be claimed. Provided that the proposed theoretical relationships were not mistaken, the instrument could not be accepted as a valid measure of individuals' adjustment to physical disability.

Thus, evidence of construct validity can be gained by seeking a positive correlation between measures of the original concept and those of other concepts to which the original concept is known to be positively related. Such evidence can be strengthened by also seeking an absence of correlation with measures of other concepts to which the original concept is known *not* to be related (or indeed a negative correlation with those with which it is known to be inversely related). Campbell and Fiske[22] have referred to the underlying principles as convergence and discrimination, respectively. Jette[23] provides an example of convergent validation of the Functional Status Index (FSI). On theoretical grounds, functional status would be expected to be directly related to stage of disease and degree of disease activity. A sample of 81 adult patients with RA were assessed on measures that included the American Rheumatology Association (ARA) stage of disease and professional global assessment of disease activity. The positive correlation between scores on these measures and those obtained by the FSI mirrored the proposed theoretical relationship and thereby provided evidence of the convergent validity of the FSI.

The need for construct validity consists in the fact that there is usually no *direct* way of testing the relationship of the instrument to the underlying concept. For example, it is difficult to understand how a criterion measure could be found for a concept such as functional performance or adjustment to physical disability. Hence, there is a need to approach the issue indirectly, by examining relationships rather than properties or characteristics. Rothstein[10(p18)] suggests that construct validity may be perceived as more of a concern for researchers in education or psychology, who may be dealing with "what we may consider vague concepts such as intelligence, anxiety, emotional state, [and so on]," than for those working in areas such as physical therapy, which is concerned with more straightforward concepts. Rothstein argues, however, that this perception is largely mis-

placed, and he views construct validity as a significant area of concern in physical therapy research.

Rothstein also suggests that both construct and content validity "are forms of 'theoretical validity' [and] may be contrasted to the criterion-related validities that are demonstrated through direct research."[10(p21)] This suggestion highlights an important point, namely, that the process of determining construct and content validity cannot be wholly empirical. In content validity, for example, whereas the relationship of an instrument to a number of concepts or variables can be established by empirical testing, the extent to which these concepts or variables in turn adequately represent an underlying domain of content can only be judged within a given theoretical framework. In a similar way, the key element in construct validity is the theoretical relationship between the primary concept and other allied concepts (Fig. 2).

Areas of Concern

There are a number of problematic issues related to measurement validity that are apparent in the physical therapy research literature. The foremost among these issues are (1) a failure to distinguish clearly between validity and reliability; (2) a concentration on issues of reliability to the exclusion of validity; and, when validity is indeed addressed, (3) methodological difficulties in establishing it. These issues will be briefly addressed with reference to specific examples.

Validity or Reliability?

The key differences between validity and reliability are not always fully appreciated or given sufficient attention. For example, in the report of their study of the use of an inclinometer for measuring cervical spine movement, Klaber Moffett et al discuss existing devices for measuring neck movement:

> Most other instruments are unwieldly [sic] for the operator and cumbersome for the patient. This in itself could affect the person's willingness to move, especially if they have a neck problem, and would therefore jeopardise the reliability of the measure.[24(p309)]

It is not clear that reliability is indeed most at stake here. Surely, the effect of a cumbersome and unwieldy instrument is likely to be that patients being tested will perform neck movements in a range that is less than their actual available range of movement. The instrument, by distorting the element it purports to measure, can introduce systematic error and therefore fail to provide a valid indication of neck movement. It may do so quite consistently, however, and thus be reliable. It is unlikely that the instrument would introduce random error in an example such as this, but were it to do so, then both reliability and validity would be diminished.

Pomeroy[25] reports an investigation of the reliability and validity of scores obtained with a mobility assessment scale for elderly people with dementia. Six subjects were videotaped performing four types of activity: coming from a sitting to a standing position, standing balance, gait, and moving from a standing to a sitting position. The tapes were then viewed, on two separate occasions, by 10 therapists experienced in the care of elderly people.

Interrater reliability was assessed by comparing the 10 therapists' scores on each item in the assessment, and intrarater reliability was assessed by comparing the scores given by each rater to each subject on the two viewings.[25] In an attempt to assess content validity, each rater's total assessment scores for each subject were rank-ordered and then compared with the other raters' orderings for each subject, so as to determine interrater agreement. It is dubious, however, whether validity was actually being tested. If, as the author suggests, content validity was the issue in question, this would have required the *scope* of the assessment to be examined (ie, the extent to which the instrument assessed all the necessary components of a mobility assessment), rather than the relative values of the scores it generated on different subjects. If criterion-related validity was the object, however, this cannot justifiably be claimed, as there was no mention of any independent standard with which the instrument's performance was being compared.

It is essential, therefore, to establish clearly whether a study will provide evidence of the inferential *accuracy* (ie, validity) of a measuring instrument or its *consistency* (ie, reliability). This is crucial, for, as we have seen, a high level of reliability gives no firm evidence that the instrument is measuring what it is supposed to. It is equally important to consider the clinical context to which any findings are likely to be applied. At times, we may be most concerned with the *absolute* value of a variable. If measuring heart rate, blood pressure, or peak expiratory flow rate, for example, it may be important to know whether a pathologically significant threshold has been exceeded; in such instances, it is necessary above all to be confident of the validity of measurements obtained with the instrument. If, in the course of treatment, we are carrying out serial goniometric measurements at a joint, however, our chief concern is likely to be with the *relative* assessment of consecutive measurements, as an indicator of improvement or deterioration in range of movement. The absolute value of joint motion is probably of little clinical significance; consequently, in this instance, it is most important to be confident of the reliability of measurements obtained with the instrument. Studies of either validity or reliability, therefore, should be evaluated as to their appropriateness for the clinical situations in which the tested instrument is likely to be used.

The Neglect of Validity

Within the physical therapy and rehabilitation literature, studies of reliability appear to be considerably more popular than those of validity. An example of a specific field in which considerable attention has been paid to the problem of reliability, but comparatively little to validity, is gait as-

sessment. There are a number of studies that have examined the reliability of visual assessments of gait. In studies of the interrater reliability of clinicians' assessments of hemiplegic gait deviations, Goodkin and Diller[26] investigated the extent to which three physical therapists agreed, whereas Miyazaki and Kubota[27] compared ratings by a physical therapy student, a physical therapist, and two physicians. DeBruin et al[28] reported an intrarater reliability study in which six orthopedic residents assessed videotaped walking sequences of children with cerebral palsy and then reexamined these sequences after an interval of a week. In a more fully reported study by Krebs et al,[29] three physical therapists rated gait deviations in children with neurological impairments while the children walked with the assistance of bilateral knee-ankle-foot orthoses. In contrast to the three previously mentioned studies, Krebs et al explored both interrater and intrarater reliability. More recently, Eastlack et al[30] have examined interrater reliability of 54 physical therapists' visual ratings of gait characteristics in patients with RA.

The validity of visual assessments of gait has received comparatively little emphasis in formal studies. Among the few such studies is that of Saleh and Murdoch[31] in which a range of health professionals' qualitative assessments of gait were examined. The investigators determined the validity of the observers' visual ratings of gait deviations in a group of five individuals with unilateral below-knee amputations. Each subject was observed walking with six different prosthetic alignments, four of which were deliberate misalignments. Observers recorded the presence or absence of seven gait deviations for each of the six prosthetic alignment configurations. These qualitative ratings were compared with a criterion in the form of the gait deviations that were predicted by biomechanical analysis[32] for the different misalignments.

Saleh and Murdoch's study[31] did not, however, address the question of the relative severity of given gait deviations. Larsson et al[33] have briefly explored this issue using a single gait characteristic. These investigators compared quantitative footswitch recordings of children with cerebral palsy with clinicians' ratings of the degree of these children's foot-strike abnormalities.

In both these studies,[31,32] although validity has been explored, the focus of the investigations has been somewhat limited. As we noted, Saleh and Murdoch[31] reported on the presence or absence of gait deviations, but not on their relative severity. Furthermore, neither of these two studies presented data on reliability in tandem with their examination of validity.

Recently, Arnell and Bowker[34] reported the validity and reliability of three orthopedic surgeons', two physical therapists', and a physical therapy student's visual assessments of the degree of normality of sagittal-plane rotations of lower-extremity joints in patients with degenerative joint disease. Observers' ratings were compared with criterion measurements, obtained by instrumented gait analysis, which were expressed as percentages of mean normal ranges measured in a control group.

Ten videotape records of patients walking at free speed were viewed on two separate occasions by the six clinicians. Six ratings were obtained from each clinician for each record viewed. In terms of group intrarater reliability, 213 of the total of 360 ratings were agreed between sessions. The reliability analysis was then restricted to the ratings that were valid on both occasions, that is, concordant with the instrumented gait data. This analysis resulted in a marked decrease in group intrarater reliability, such that only 30 ratings were agreed between sessions. These findings illustrate that the ability to assign a rating to an observation consistently is of no clinical value if the rating itself is incorrect; reliability does not guarantee validity.

Methodological Difficulties

Probably the foremost difficulty to be encountered in any attempt to establish measurement validity is identifying an independent standard for the "true" value of the entity to be measured. In interpretive or qualitative research, such a standard may simply not exist. Rather than seek a predetermined criterion of validity, researchers dealing with qualitative data are more likely to use key informants or other sources of expert opinion. Thus, when exploring such areas as the cognitive and affective dimensions of the healing process or the concept of adolescent hopefulness, Hinds et al submitted their data to "a panel of reviewers whose members [were] selected for their theoretical sensitivity to the studied phenomena."[35(p432)]

Even in quantitative research, however, it is not always an easy process to identify an independent criterion. A recent study of the validity and reliability of physical therapists' auscultation of lung sounds[36] demonstrates some of the methodological hurdles that may be encountered. These researchers were aware of the inherent difficulty of finding a "true" source of lung sounds against which to evaluate physical therapists' findings. They note that, however accurate the identification of certain lung sounds, there is no certain relationship between these acoustic findings and underlying pathophysiological processes. Even if no diagnostic claims are made for the sounds being examined, an acceptable source of the required lung sounds must still be identified. A possible approach would seem to be for the therapists studied to have auscultated the same patient(s). There are, however, three problems inherent in this approach. First, it would have been difficult to identify a criterion against which to measure each therapist's identification of in vivo sounds. A "true" classification would need to have been assigned to each sound with which the therapists were presented. Second, as the authors note,

> ... the changes that occur in lung sounds with repeated breaths would dictate that all subjects listen simultaneously at the same point on the patient's chest wall.[36(p282)]

As a result, the study used tape-recorded lung sounds to circumvent these difficulties. The tape provided an unchanging source of lung sounds and represented a fairly authoritative criterion. Unfortunately, this method introduces the third problem, which is one of external validity. The findings of the study relate to recorded sounds, and it cannot be assumed that they can be extrapolated to auscultation in vivo.

Despite seemingly rigorous selection of a criterion for assessing the validity of measurements obtained with an instrument, investigators may overlook factors that contaminate the validity test. Johnston and Smidt,[37] for example, used an electrogoniometer to measure hip joint motion during walking in a patient with a surgically fused hip. They concluded that the 5 degrees of sagittal-plane hip motion recorded in this patient defined the error in hip joint measurements attributable to slipping of the exoskeletal device on the patient. Moreover, they suggested that the amount of motion that was recorded represented the upper limit for measurement error attributable to slipping, as this type of patient was likely to produce greater slipping of the goniometer pelvic band than healthy subjects because of greater motion in these patients' lumbar spines.

Arnell[38] conducted a similar experiment, using a polarized light goniometer, to obtain hip motion recordings. The assumption made in this investigation was that the recorded range of motion would reflect error from slipping of the goniometer's pelvic band or movement of the soft tissue exoskeleton of the thigh. The patient's recorded hip movement was 6 degrees.

Gore,[39] however, offers an alternative explanation for recordings of sagittal-plane rotations in surgically fused hips. Using a biomechanical model, he calculated that the deflection of the femoral shaft of a limb with a rigidly fused hip could be of the order of 4 to 5 degrees during walking. His proposal challenges the assumption that motion recordings in patients with surgically fused hips are artifacts attributable to the mode of attachment of a goniometric device or to movement of soft tissues.

Finally, difficulties may be encountered when attempting to use a previously validated instrument, either for use in its own right or as a criterion to validate another instrument. As demonstrated in the leg-length study by Beattie et al,[12] the validity of a measurement relates to a specific measurement context. Similarly, the validity of measurements that has been demonstrated with a given group of patients or subjects cannot necessarily be extrapolated to other groups. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), for example, was validated on a population of patients with osteoarthritis[40] and should not be assumed to be equally valid for individuals with other arthropathies. Chappell,[41] when validating health and disability indexes on an elderly population, found that one of the instruments tested yielded valid and reliable measurements when used for individuals living in the community or in institutions, but not for those living in subsidized housing. She concludes that her findings

> ... strongly suggest the need to assess our measurement tools among different subsamples of elderly persons, rather than assuming their applicability to all persons 65 [years of age] and over.[41(p101)]

Conclusions

The foregoing exploration of the concepts of validity and reliability demonstrates the key differences between these two measurement features and illustrates the four principal types of validity with examples from the literature. Some of the methodological difficulties that confound the process of validating measurements are outlined.

In each measurement instance, the therapist needs to determine whether accuracy (validity) or consistency (reliability) is the requisite measurement feature of the instrument being used. This can only be determined by the clinical context in which the measurements will be applied.

Evaluation of the effectiveness of physical therapy interventions has been hampered by the lack of appropriate assessment instruments. Where instruments do exist, too often the exact referents of the measurements yielded by the instrument are not fully defined, leading to inappropriate assumptions about the nature of the entity being measured.

As has been illustrated in this article, the evaluation of measurements yielded by assessment instruments often focuses solely on reliability. To enable inferences to be made on the basis of these measurements, evidence of measurement validity must also be provided. This is an area on which physical therapy research could profitably focus.

References

1 LeCompte MD, Goetz JP. Problems of reliability and validity in ethnographic research. *Review of Educational Research.* 1982;52(1):31–60.

2 Kirk J, Miller ML. *Reliability and Validity in Qualitative Research.* Newbury Park, Calif: Sage Publications Inc; 1986.

3 Babbie E. *The Practice of Social Research.* 5th ed. Belmont, Calif: Wadsworth Publishing Co; 1989.

4 Polgar S, Thomas SA. *Introduction to Research in the Health Sciences.* 2nd ed. Melbourne, Victoria, Australia: Churchill Livingstone; 1991.

5 Kerlinger FN. *Foundations of Behavioral Research.* 2nd ed. New York, NY: Holt, Rinehart and Winston, Inc; 1973.

6 Task Force on Standards for Measurement in Physical Therapy. Standards for tests and measurements in physical therapy practice. *Phys Ther.* 1991;71:589–622.

7 Payton OD. *Research: The Validation of Clinical Practice.* 2nd ed. Philadelphia, Pa: FA Davis Co; 1988.

8 Wood-Dauphinee S, Williams JI. Much ado about reliability. *Physiotherapy Canada.* 1989;41:234–236.

9 Carmines EG, Zeller RA. *Reliability and Validity Assessment.* Newbury Park, Calif: Sage Publications Inc; 1979.

10 Rothstein JM. Measurement and clinical practice: theory and application. In: Rothstein JM, ed. *Measurement in Physical Therapy.* New York, NY: Churchill Livingstone Inc; 1985:1–46.

11 Currier DP. *Elements of Research in Physical Therapy.* 3rd ed. Baltimore, Md: Williams & Wilkins; 1990.

12 Beattie P, Isaacson K, Riddle DL, Rothstein JM. Validity of derived measurements of leg-length differences obtained by use of a tape measure. *Phys Ther.* 1990;70:150–157.

13 Gajdosik RL, Bohannon RW. Clinical measurement of range of motion: review of goniometry emphasizing reliability and validity. *Phys Ther.* 1987;67:1867–1872.

14 Rose G, Barker DJP. Repeatability and validity. *BMJ.* 1978;2:1070–1071.

15 Feldman AB, Haley SM, Coryell J. Concurrent and construct validity of the Pediatric Evaluation of Disability Inventory. *Phys Ther.* 1990;70:602–610.

16 Guidubaldi J, Perry J. Concurrent and predictive validity of the Battelle Developmental Inventory at the first grade level. *Educational and Psychological Measurement.* 1984;44:977–985.

17 Silverman D. *Qualitative Methodology and Sociology: Describing the Social World.* Aldershot, England: Gower Publishing Co Ltd; 1985.

18 Hammersley M, Atkinson P. *Ethnography: Principles in Practice.* London, England: Tavistock Publications Ltd; 1983.

19 Seeman M. On the meaning of alienation. *American Sociological Review.* 1959;24:783–791.

20 Thomas V. Fitness within sport. In: Reilly T, ed. *Sports Fitness and Sports Injuries.* London, England: Faber & Faber Ltd; 1981:19–22.

21 Bellamy N, Sothern RB, Campbell J, Buchanan WW. Circadian rhythm in pain, stiffness, and manual dexterity in rheumatoid arthritis: relation between discomfort and disability. *Ann Rheum Dis.* 1991;50:243–248.

22 Campbell DT, Fiske DW. Convergent and discriminant validation by the Multitrait-Multimethod Matrix. *Psychol Bull.* 1959;56:81–105.

23 Jette AM. The Functional Status Index: reliability and validity of a self-report functional disability measure. *J Rheumatol.* 1987;14(suppl 15):15–19.

24 Klaber Moffett JA, Hughes I, Griffiths P. Measurement of cervical spine movements using a simple inclinometer. *Physiotherapy.* 1989;75:309–312.

25 Pomeroy V. Development of an ADL oriented assessment-of-mobility scale suitable for use with elderly people with dementia. *Physiotherapy.* 1990;76:446–448.

26 Goodkin R, Diller L. Reliability among physical therapists in diagnosis and treatment of gait deviations in hemiplegics. *Percept Mot Skills.* 1973;37:727–734.

27 Miyazaki S, Kubota T. Quantification of gait abnormalities on the basis of continuous foot-force measurement: correlation between quantification indices and visual rating. *Med Biol Eng Comput.* 1984;22:70–76.

28 DeBruin H, Russell DJ, Latter JE, Sadler JTS. Angle-angle diagrams in monitoring and quantification of gait patterns for children with cerebral palsy. *Am J Phys Med.* 1982;61:176–192.

29 Krebs DE, Edelstein JE, Fishman S. Reliability of observational kinematic gait analysis. *Phys Ther.* 1985;65:1027–1033.

30 Eastlack ME, Arvidson J, Snyder-Mackler L, et al. Interrater reliability of videotaped observational gait-analysis assessments. *Phys Ther.* 1991;71:465–472.

31 Saleh M, Murdoch G. In defence of gait analysis. *J Bone Joint Surg [Br].* 1985;67:237–241.

32 Radcliffe CW, Foort J. *The Patellar-Tendon-Bearing Below Knee Prosthesis.* Berkeley, Calif: Biomechanics Laboratory, University of California; 1961.

33 Larsson LE, Miller M, Norlin R, Thaczuk H. Changes in gait patterns after operations in children with spastic cerebral palsy. *Int Orthop.* 1986;10:155–162.

34 Arnell P, Bowker P. The accuracy of visual gait assessment. In: *Proceedings of the 11th International Congress of the World Confederation for Physical Therapy.* London, England: World Confederation for Physical Therapy; 1991:431.

35 Hinds PS, Scandrett-Hibden S, McAuley LS. Further assessment of a method to estimate reliability and validity of qualitative research findings. *J Adv Nurs.* 1990;15:430–435.

36 Aweida D, Kelsey CJ. Accuracy and reliability of physical therapists in auscultating tape-recorded lung sounds. *Physiotherapy Canada.* 1990;42:279–282.

37 Johnston RC, Smidt GL. Measurement of hip-joint motion during walking: evaluation of an electrogoniometric method. *J Bone Joint Surg [Am].* 1969;51:1083–1094.

38 Arnell MM. *Numerical Descriptors of the Intersegmental Kinematics of Gait.* Manchester, England: University of Manchester; 1988. Doctoral thesis.

39 Gore TA. *The Kinematics of Normal and Pathological Hip Joints.* Durham, England: University of Durham; 1980. Doctoral thesis.

40 Bellamy N, Buchanan WW, Goldsmith CH, et al. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol.* 1988;15:1833–1840.

41 Chappell NL. Measuring functional ability and chronic health conditions among the elderly: a research note on the adequacy of three instruments. *J Health Soc Behav.* 1981;22:90–102.

Commentaries

Following are three commentaries on "Measurement Validity in Physical Therapy Research."

I would like to congratulate the authors of this article for pointing out that the validity of physical therapy measurements is in need of careful scrutiny. Sim and Arnell have drawn careful distinctions between reliability and validity. On close inspection, however, these concepts are highly interrelated, and each can affect the other. The authors have made the point that it is not possible to have a highly valid measure if reliability is low. On the other hand, it is also possible for the calculated reliability of a measure to be affected by problems with validity when systematic error is operating.[1]

The internal consistency coefficient—the definition of reliability in classical test theory—presents a second example of the close relationship of these concepts. The internal consistency coefficient (typically Cronbach's alpha) identifies the extent to which a scale has a single underlying dimension.[2] This aspect of reliability is highly related to the measure's content and construct validity. I believe it is best to think of the concepts of reliability and validity as being equally important and as each having multiple components that interact in complex ways. The crux of the matter is that both researchers and clinicians must define what they are attempting to do within a clearly articulated theoretical context. The authors have generated many interesting examples to stimulate thought in these directions.

Despite the excellent presentation of concepts and examples, I regret that the title of this article and the con-

Measurement Validity in Physical Therapy Research
Julius Sim and Peggy Arnell
PHYS THER. 1993; 73:102-110.