

Reliability of the Clinical Examination: How Close is "Close Enough"?

LtCol Robert S. Wainner, PT, PhD, OCS, ECS, FAAOMPT¹

It is has been said that close only counts in a game of horseshoes. The reason being, of course, that when playing horseshoes you get points for having the shoe closest to the stake even if it isn't touching. In reality, this saying is true for many psychomotor activities. For example, a baseball pitch only counts as a strike if it is within the strike zone, a thrown dart counts as a bull's-eye only if it hits somewhere within the red center, and a field goal in football is only worth 3 points if it passes between the uprights. In each of these tasks, there is a certain margin of error that is acceptable, and the margin of error for each has been operationally defined. Likewise, the clinical examination, which is comprised of the history and physical examination, consists of psychomotor tasks that are performed on a regular basis by all practicing clinicians. The ultimate goal, of course, is to accurately establish a diagnosis, direct the choice of intervention, and establish a prognosis. The practical question that the clinician must ask is, "Should I include this clinical test or measure as part of my examination?" Often the answer is based on whether the test or measure being considered has a reliability coefficient that surpasses some predefined threshold. But is this the proper approach? If only one could get a straight answer!

Although it is tempting to think of reliability as a yes-or-no decision, reliability actually exists along a continuum and is not an action-potential concept (all or none). Rather, reliability is a statistical model that characterizes the amount of error contained when repeated measurements are obtained. The key words here are *model* and *amount of error*. A model is merely an imperfect, tangible framework that helps us understand some construct or concept. The statistical coefficients most commonly used to characterize the reliability of a test or measure are the intraclass correlation coefficient (ICC) and the kappa statistic (K), both of which are based on statistical models. As with other models, these statistics have certain assumptions (like restriction in range and base rates) that are critical and that need to be met if we are to have confidence in their application and results.^{2,6,7} When assumptions are violated, the coefficients become suspect and the true reliability of the test or measure remains unknown and must be further examined. Let's consider the second key word, *amount of error*. The results of all clinical tests and measures contain some amount of error.² Continuing the game analogy this editorial started with, a clinician scores when an accurate diagnosis, prognosis, or measure is obtained. Unfortunately, unlike the games mentioned above, the operational definition as to how much error is acceptable depends on the clinical context; there isn't a fixed operational definition of how much error is acceptable. True, there are qualitative descriptions of reliability coefficients such as "poor," "moderate," "good," "excellent" that have been proposed by a number of authors.^{2,6,8} Unfortunately, the prevailing opinion seems to be that a test or measure must at least approximate, or better yet, surpass one of the upper thresholds, without which a test could not possibly have any validity. However, the authors who coined these qualitative terms stated that the reliability of a particular test or measure must be interpreted within the context of its intended use.^{6,8} In other words, these general

¹ Assistant Professor, US Army-Baylor Graduate Program in Physical Therapy, US Army-Baylor University, Ft. Sam Houston, TX. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Department of the Air Force, Department of the Army, or the Department of Defense.

guidelines should not be viewed as a gold standard filter through which all tests and measures must pass before the validity of a test or measure can be considered.

While reliability is important, it is more important to demonstrate whether or not the measure has actually been shown to be useful.⁹ That brings us to the related concept of validity. Although there are many forms of validity, establishing the criterion validity of a test or measure (often expressed as sensitivity/specificity and likelihood ratios) is helpful in the clinical decision-making process to establish a diagnosis, direct the choice of intervention, and establish a patient's prognosis. The reliability of the results from a test or measure for the purposes of diagnosis and prognosis is certainly related to validity, but how close is close enough? Is there an operational definition of reliability that tells us whether or not a test or measure will be helpful for making the proper diagnosis or prognosis, if you would? Consider the examples from the rehabilitation literature in Table 1.

One can clearly see that items with moderate and even poor reliability coefficients based on commonly accepted guidelines may be useful in the clinical decision-making process. For example, although the kappa coefficient for hypermobility spring testing of the lumbar spine was low ($K = 0.30$), the associated positive likelihood ratio was useful ($+LR = 9.2$) for predicting which patients would fail to respond to a 4-week lumbar stabilization program.⁴ Lest one think these examples are a weak attempt by rehabilitation professionals to make excuses for the tools of our trade, consider the following examples from the medical literature in Table 2.

Once again, the findings are similar. So what gives? Has everything we ever learned about reliability and validity become invalid? No. But it is safe to say that we still have much to learn about the relationship between reliability and validity. In a few of these examples, model assumptions were violated; in others, a different explanation is required. Interestingly enough, in most of the rehabilitation examples the variables were also identified in a separate analysis as being part of a cluster of findings that, taken as a whole, were useful for establishing a patient's diagnosis or prognosis. Considering our previous examples, does reliability have to be qualified as excellent or will moderate suffice? Is a test or measure worthless if associated with a poor reliability coefficient? In the case of criterion validity, we should be cautious

TABLE 1. Test examples from rehabilitation literature.

Test or Measure	Reliability Coefficient	Validity Coefficient	Condition (Reference Criterion)
Hypomobility present ³	$K = 0.13$	$Sn = 0.97$ $Sp = 0.23$ $-LR = 0.13$	Acute low back pain (50% improvement in Oswestry over a 4-d period)
Hypermobility absent ⁴	$K = 0.30$	$Sn = 0.28$ $Sp = 1.0$ $+LR = 9.2$	Chronic low back pain (less than 6-point improvement in Oswestry over a 4-wk period)
Cervical ROM of painful side (inclinometer) ¹³	$ICC = 0.75$ $SEM = 6.6^\circ$	$Sn = 0.89$ $Sp = 0.23$ $-LR = 0.23$	Cervical radiculopathy (+ needle electromyography)
Repeated timed trunk flexion (inclinometer) ¹⁰	$ICC = 0.45$ $SEM = 6.7$ s	Different between groups ($P < .001$)	Patients with low back pain and asymptomatic subjects (NA)
Straight leg raise discrepancy $\geq 20^\circ$ ¹²	$K = 0.23$	$OR = 5.9$	Low back pain (sick leave ≥ 15 d)
Pain provocation in buttock with sidebending ¹²	$K = 0.23$	$OR = 3.3$	Low back pain (sick leave ≥ 15 d)

Abbreviations: ROM, range of motion; K, kappa statistic; ICC, intraclass correlation coefficient; SEM, standard error of measure; Sp, specificity; Sn, sensitivity; -LR, negative likelihood ratio; OR, odds ratio; +LR, positive likelihood ratio; NA, not applicable.

TABLE 2. Test examples from medical literature.

Test or Measure	Reliability Coefficient	Validity Coefficient	Condition (Reference Criterion)
Position sense ¹¹	K = 0.28	+LR = 12.9	Diabetic neuropathy (monofilament examination)
Slow cardiac upstroke ¹	K = 0.26	+LR = 9.2	Moderate to severe aortic stenosis (Doppler echocardiography)
Breath sound intensity ⁵	K = 0.23	+LR = 3.7	COPD (FEV ₁ and FEV ratios)
Palpation of subxyphoid cardiac apical impulse ⁵	K = 0.30	+LR = 4.6	COPD (FEV ₁ and FEV ratios)

Abbreviations: K, kappa statistic; LR, Likelihood ratio; COPD, chronic obstructive pulmonary disease; FEV, force expiratory volume; FEV₁ = 1 second forced expiratory volume.

before dismissing a test or measure based solely on the basis of a low reliability coefficient. In the author's opinion, unless reliability studies are performed and interpreted in the context of validity, we run the risk of impeding the discovery process involved in clinical research. A test or measure can possess an "excellent" reliability coefficient and have no validity whatsoever.³ Similarly, a test or measure may have a somewhat lower reliability coefficient yet still have useful validity. In most cases, studies of reliability should be embedded in the context of a broader study whose aims also address the validity of the test or measure of interest.

Reliability is an unarguably important measurement property, but it isn't always as straightforward as we have been led to believe. That being the case, what are clinicians and clinician/researchers to do so we can excel at the task of the clinical examination and the judgments we make from it? Several things: first, when reporting the reliability of results for a particular test or measure, authors should critically examine and report the extent to which statistical model assumptions (eg, restriction of range and prevalence) were satisfied; second, validity (eg, sensitivity, specificity, and likelihood ratios) and precision coefficients (eg, standard error of measure) should be reported when applicable; third, consumers of the literature should demand that authors include the information mentioned in the previous 2 points or withhold judgment until it is supplied; finally, the reliability of results for a particular test or measure must be interpreted within the context of its intended use. Otherwise, how are we to know how close is close enough?

REFERENCES

1. Etchells E, Glens V, Shadowitz S, Bell C, Siu S. A bedside clinical prediction rule for detecting moderate or severe aortic stenosis. *J Gen Intern Med.* 1998;13:699-704.
2. Fleiss, JL. The measurement of interrater agreement. In: *Statistical Methods for Rates and Proportions.* New York, NY: John Wiley and Sons; 1981.
3. Flynn T, Fritz J, Whitman J, Wainner R, Magel J, Rendeiro D, Butler B, Garber M, Allison S. A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. *Spine.* 2002;27:2835-2843.
4. Hicks, GE, Fritz JM, Delitto A, Mischock J. Interrater reliability of clinical examination measures for identification of lumbar segmental instability. *Arch Phys Med Rehabil.* In press.
5. Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA.* 1995;273:313-319.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-174.
7. Mitchell SK. Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychol Bull.* 1979;86:376-390.

8. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 2nd ed. Norwalk, CT: Appleton and Lange; 2003.
9. Rothstein JM. Sick and tired of reliability? *Phys Ther*. 2001;81:774-775.
10. Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, Radwan H. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. *Spine*. 1998;23:2412-2421.
11. Smieja M, Hunt DL, Edelman D, Etchells E, Cornuz J, Simel DL. Clinical examination for the detection of protective sensation in the feet of diabetic patients. International Cooperative Group for Clinical Examination Research. *J Gen Intern Med*. 1999;14:418-424.
12. Viikari-Juntura E, Takala EP, Riihimäki H, Malmivaara A, Martikainen R, Jappinen P. Standardized physical examination protocol for low back disorders: feasibility of use and validity of symptoms and signs. *J Clin Epidemiol*. 1998;51:245-255.
13. Wainner RS, Fritz JM, Irrgang JJ, Boninger ML, Delitto A, Allison S. Reliability and diagnostic accuracy of the clinical examination and patient self-report measures for cervical radiculopathy. *Spine*. 2003;28:52-62.