ACRM
AMERICAN CONGRESS OF REHABILITATION MEDICINE

Archives of Physical Medicine and Rehabilitation

journal homepage: www.archives-pmr.org

Archives of Physical Medicine and Rehabilitation 2016;97:1051-2

FROM THE EDITORS' DESK

What to Do With "Moderate" Reliability and Validity Coefficients?

Marcel W. Post, PhD

From the University of Groningen, University Medical Center Groningen, Department of Rehabilitation Medicine, Groningen; and Brain Center Rudolf Magnus and Center of Excellence in Rehabilitation Medicine, University Medical Center Utrecht and De Hoogstraat, Utrecht, The Netherlands.

Abstract

Clinimetric studies may use criteria for test-retest reliability and convergent validity such that correlation coefficients as low as .40 are supportive of reliability and validity. It can be argued that moderate (.40−.60) correlations should not be interpreted in this way and that reliability coefficients <.70 should be considered as indicative of unreliability. Convergent validity coefficients in the .40 to .60 or .40 to .70 range should be considered as indications of validity problems, or as inconclusive at best. Studies on reliability and convergent should be designed in such a way that it is realistic to expect high reliability and validity coefficients. Multitrait multimethod approaches are preferred to study construct (convergent-divergent) validity.

Archives of Physical Medicine and Rehabilitation 2016;97:1051-2

The inspiration for this editorial on reliability and validity coefficients came from an article on interrater reliability of a new measure that I handled as section editor of *Archives*. In this article, the authors' statement that "We considered agreement to be poor for ICCs <.40, good for .40 to .70 and excellent for >.70" caught my attention because it seemed very lenient.[1(p1987)] The reference for this interpretation was an article by Carod-Artal et al.[2] In that article I found the same figures and interpretation, with a reference to the well-known textbook by Nunnally and Bernstein,[3] but in the 1994 edition of this book I found no such guideline.

To find out how common this interpretation was, I visited the Rehabilitation Measures Database (RMD) (available at: http://www.rehabmeasures.org/rehabweb/rhstats.aspx) and the outcome measurement page of the Spinal Cord Injury Rehabilitation Evidence (SCIRE) project (available at: http://www.scireproject.com/tables/table-3-criteria-rating-properties-of-outcome-measures). The SCIRE project advises to consider a reliability coefficient of .40 to .75 as adequate. The RMD does the same for interrater reliability, but it is more restrictive for test-retest reliability, for which a minimum of .70 for studies at group level is advised.

Regardless of whether an intraclass correlation coefficient (ICC) in the .40 to .70 range is considered good or adequate, the type of ICC used,[4] and whether the study concerns test-retest or interrater reliability, I doubt the usefulness of accepting an ICC as low as .40 as evidence of reliability. In case of a Pearson correlation coefficient, .40 corresponds to only 16% explained variance. An ICC is not a Pearson correlation, but I hope the analogy illustrates that an ICC value in this range indicates such a large amount of measurement error that the test cannot be taken to be reproducible. Even a Pearson correlation of .70 implies that only about half of the variance of the reference test administration is explained.

What do the experts recommend? A minimum of .70 for an ICC for application of a measure at group level is recommended by Fitzpatrick,[5] Terwee,[6] and colleagues, who refer to the same textbook by Nunnally and Bernstein[3] as the aforementioned Carod-Artal study.[2] Streiner and Norman conclude that .75 is a "fairly minimal requirement for a useful measure."[7(p195)] Based on this literature, I feel we should no longer accept reliability coefficients in the .40 to .70 range as evidence for test-retest or interrater reliability of a measure.

Validity, and I restrict myself to convergent or concurrent validity here, is a more complex issue because associations between different measures are examined to establish convergent validity, and perfect correlation therefore cannot be expected.[5] The RMD and SCIRE project both advice to consider a correlation of .31 to .60 as adequate and >.60 as excellent indication of

Disclosures: none.

0003-9993/16/$36 - see front matter © 2016 by the American Congress of Rehabilitation Medicine
http://dx.doi.org/10.1016/j.apmr.2016.04.001

convergent validity. Again one wonders why such correlations, reflecting 10% and 36% explained variance, respectively, would be satisfactory evidence of the similarity of 2 measures. In our studies, we frequently find correlations in the range of .40 to .60 between measures of clearly divergent but related constructs (eg, perceived health and well-being, activities and participation, coping and health-related quality of life). An association within this range therefore may show that the measures under investigation can be valid measures of different—but related—concepts, invalid measures of similar concepts, or a combination of both and is therefore not useful for this goal.

The low standards for convergent validity seem to be based on correlations typically found in the literature in the past. Nunnally stated in an earlier edition that "correlations as high as 0.7 are rare and the average of all correlations reported in the literature are <0.4"[8(p143)] McDowell and Newell wrote that concurrent validity correlations between the tests reviewed in their book were low, "typically falling between 0.2 and 0.6, with only occasional correlations between very similar instruments (such as the Barthel and PULSES scales) falling above 0.7"[9(p30)] Fitzpatrick et al referred to McDowell and Newell and wrote, "…given typical levels of reliability of patient-based variables, a correlation coefficient of 0.6 may be strong evidence in support of construct validity."[5(p26)] However, none of these authors argue that correlations <.60 support convergent validity.

Given the current plethora of measures, it should be possible to find a similar measure as reference for almost any construct. Therefore, there seems to be no reason to validate a new measure against a more or less similar reference measure anymore. Therefore, we may expect from authors of a convergent validity study to include a reference measure with which a high correlation may be expected.

Most authors argue that we should keep in mind that validity of a measure is established over a number of studies, and it is not possible to define any single appropriate cutoff for convergent validity.[3,5-9] In contrast, a multitrait multimethod is preferred to study construct (convergent-divergent) validity.[3,5,6] In this method, measures reflecting similar and dissimilar concepts are chosen as reference measures, and it is expected that the measure under study correlates strongly (eg, >.60 or .70) with the similar measure and does not correlate (eg, <.30 or .40) with the dissimilar measure. I think that even in this method correlations in the range of .40 to .60 indicate validity problems or are inconclusive at best; however, others argue that the size of the coefficients is less important as long as researchers provide specific a priori hypotheses about the directions and strengths of these correlations and that at least 75% of these hypotheses are confirmed.[6]

I think we should become stricter when reviewing validation studies. In particular, we should stop considering moderate correlations as evidence of reliability or validity. Researchers can be expected to design their clinimetric study in such a way that it is realistic to expect stronger correlations (eg, by avoiding a long time between test occasions, unclear test instructions, only partly similar reference measures).

List of abbreviations:

ICC	**intraclass correlation coefficient**
RMD	**Rehabilitation Measures Database**
SCIRE	**Spinal Cord Injury Rehabilitation Evidence**

Keywords

Outcome assessment (health care); Psychometrics; Rehabilitation; Validation studies as topic

Corresponding author

Marcel W. Post, PhD, De Hoogstraat Rehabilitation, Rembrandtkade 10, 3583TM Utrecht, The Netherlands. *E-mail address:* m.post@dehoogstraat.nl.

References

1. Kozlowski AJ, Singh R, Victorson D, et al. Agreement between responses from community-dwelling persons with stroke and their proxies on the NIH Neurological Quality of Life (Neuro-QoL) Short Forms. Arch Phys Med Rehabil 2015;96:1986-92.
2. Carod-Artal FJ, Ferreira Coral L, Stieven Trizotto D, Menezes Moreira C. Self- and proxy-report agreement on the Stroke Impact Scale. Stroke 2009;40:3308-14.
3. Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill; 1994.
4. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30-46.
5. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess 1998;2:i-iv. 1-74.
6. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
7. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 5th ed. Oxford: Oxford Univ Pr; 2008.
8. Nunnally JC. Psychometric theory. New York: McGraw-Hill; 1978.
9. McDowell I, Newell C. Measuring health. A guide to rating scales and questionnaires. Oxford: Oxford Univ Pr; 1987.