The reliability of an Arabic translation of the chronic obstructive pulmonary disease assessment test

To the Editor

I have read the interesting and very useful article on "the reliability of an Arabic translation of the chronic obstructive pulmonary disease assessment test" by Al-Moamary et al.[1] The paper appears as a product of a joint work between 2 major medical institutions in Riyadh. Such research type usually takes time to conduct and requires meticulous care and perseverance to collect a sufficient sample size while eliminating major confounders during testing and re-testing. For that, those involved in such a work must be commanded. Despite what have been said, I still have several comments on the study's reliability procedures that the authors may find useful, especially if they are going to revisit their data.

First, the authors have indicated in their methods section that Cronbach's alpha and intraclass correlation coefficient (ICC) were used for internal consistency and test re-test reliability, respectively. However, looking at the results of their study in table 2, there was only interclass (not intraclass) correlation coefficient. The authors should have presented the findings for both coefficients as well as inter-item correlation matrix. Cronbach's alpha, used in this study, is a useful measure for assessing internal consistency (homogeneity); that is how closely related sets of items are as a group. This is because when items are used to form a scale they need to have internal consistency.[2] Although Cronbach's alpha is widely acknowledged as a measure of internal consistency, one can increase alpha coefficient by increasing the number of items (k). Additionally, if the average inter-item correlation is low, alpha will be low. On the other hand, as the average inter-item correlation increases, Cronbach's alpha increases as well (holding the number of items constant).

It is well recognized that a good test is one that assesses different aspects of the trait (such as quality of life) being studied. If a test has a strong internal consistency, it should show only moderate to high correlation among items (0.70 to 0.90). If correlations between items are too low, it is likely that they are measuring different things and therefore should not all be included in a test that is supposed to measure one trait. At the same time, since all the items are intended to measure the same thing, they should be correlated with one another. However, if item correlations are too high, it is likely that some items are redundant and should be removed from the test.

Second, in the statistical analysis section, the authors have also stated that the reliability was tested using ICC. Yet, throughout the results and the discussion sections, the authors kept mentioning interclass correlation coefficient, so, which reliability coefficient had they really used? If the authors did use interclass correlation coefficient (Pearson's or Spearman's rank correlation coefficient), this was not the right choice for the test re-test reliability, because the Pearson r is a bi-variate measure. Instead, the uni-variate measure of reliability (ICC) is more appropriate measure for the test re-test analysis.[3,4] The ICC is the ratio of the variance among subjects (subject variability) over the total variance. These variances are derived from analyses of variance (ANOVA). When using a bi-variate test (for example, Pearson r), we could still get a high correlation coefficient even if the responses in the second test increased (or decreased) by 100% compared to the responses in the first test. Moreover, ICC will produce a value of r = 1.0 only if all observations on each subject are identical and the intercept is at zero. However, ICC, like interclass correlation, sometimes has its shortcomings. Its value is dependent on the range of the variables measured. With larger ranges (a more heterogeneous population), the value of ICC is higher. In addition, the ICC is a ratio of variances and, therefore, difficult to interpret clinically. Therefore, it may be more informative clinically to also calculate the standard errors of measurement (SEMs), or the square root of the error variances, which is expressed in the metric units of the original measurement and is calculated as follow: $SEM = SD \times\sqrt{}\ (1 - r)$, where SD is the standard deviation and r is the correlation coefficient. The disadvantage of the SEM is that no clear criteria for an acceptable value are available, though the smaller the SEM the more reliable the measurements.[4]

Third, in addition to the relative reliability (such as ICC) the authors could have added an absolute reliability test such as %coefficient of variation (CV) or Bland and Altman test of agreement. Using limits of agreement for Bland and Altman would also show if there is any heteroscedasticity in the data.[5,6] The Bland and Altman level of agreement test and the 95% limits of agreement can be obtained by calculating the mean difference (d) between the 2 tests (test and re-test) and the standard deviation (SD) for this difference. The closer d is to zero and the smaller the SD of this difference, the better the test re-test agreement. Differences between the 2 tests can also be plotted against the mean of the measurements made by the 2 tests. The graph would show the size, direction, and range of the differences and indicates whether differences between test re-test are consistent across the range of measurements (no heteroscedasticity). The 95% limits of agreement (as the mean difference between the 2 tests ±1.96 SD of the differences) indicate the total error (both bias and random error). The presence of bias between the test and retest is estimated by calculating the 95% confidence interval (CI) for d. The 95% CI for d can be calculated as d ± tn - 1SEM (d), where n is the number of subjects and SEM is the standard error of the mean (SD/√n). If zero lies outside the 95% CI, systematic differences (bias) between the observers exist.[4-6] These formulas, however, hold if the differences are not dependent on the value of the mean (larger differences with higher means). In this case, if heteroscedasticity exists, transformation of the data (such as log transformation) is required to make the differences independent of the mean.

Fourth, looking at the mean values shown in table 2, one can tell that the mean of the total score in the second test (re-test) is 14% lower than the mean of the total score in the first test. The same thing can be said for the means of the individual items in the same table. This consistent drop in the responses values may suggest a systematic error (this can be confirmed using Bland and Altman test of agreement).

Fifth, the values of SD for the 2 means of the total score that are shown in table 2 indicate a considerable variability. Calculating values for the CV for the test and re-test confirmed a fairly high variability (54.2% for the test and 49.9% for the re-test). This means that assuming the data are normally distributed, 68% of the differences between the test and re-test lie within at least 50% of the mean of the data, something does not reflect a good absolute reliability.

Finally, I understand that the authors are aware of the importance of future validation of their Arabic instrument (the Chronic Obstructive Pulmonary Disease [COPD] Assessment Test [CAT]) against objective measure, and we are looking forward to seeing such validity study realized in the near future. After all, a test can be reliable and not valid; however, a test cannot be valid and not reliable.

Hazzaa M. Al-Hazzaa
Exercise Physiology Laboratory
King Saud University
Riyadh, Kingdom of Saudi Arabia

Reply from the Author

We are writing this letter in reply to the "Letter to the Editor" written by Dr. Hazzaa M. Al-Hazzaa, pertaining to the article "The reliability of an Arabic translation of the chronic obstructive pulmonary disease assessment test" (Al-Moamary et al).[1] We would like to thank Dr. Al-Hazzaa for his interest in our paper, as well as his detailed review and critique of the statistical analyses and the results' presentation. The intra-class correlation coefficient was the measure used to assess the test-retest reliability, although, referring to it as interclass correlation coefficient was a typo. The intra-class correlation coefficient, introduced in the late 1970's[3,7] has been widely used as the preferred measure of test-retest reliability since few decades up to this date.[8-11] The appropriateness of using the intra-class correlation coefficient as the statistical test for assessing the test-retest reliability in our paper is not justifiably questioned. Although the use of the Bland-Altman test is appropriately described by Dr. Al-Hazzaa as a good supportive test to carry out for the reliability analyses, it does not undermine the importance of the intra-class correlation coefficient as the superior test for test-retest reliability. Moreover, the other suggested alternative was the Coefficient of variation, which has been found not to be a proper measure of reliability.[12]

Again, I would like to thank Dr. Al-Hazzaa for his critical appraisal of the statistical analyses carried out in the paper, and I may take his suggestions in consideration for future work. Finally, despite all, I still believe that the statistical analyses carried out in this paper were appropriate for the question and the data we addressed.

Mohamed S. Al-Moamary
Hani M. Tamim
Department of Clinical Affairs
College of Medicine
King Saud Bin Abdulaziz University for Health Sciences
Riyadh, Kingdom of Saudi Arabia

References

1. Al-Moamary MS, Al-Hajjaj MS, Tamim HM, Al-Ghobain MO, Al-Qahtani HA, Al-Kassimi FA. The reliability of an Arabic translation of the chronic obstructive pulmonary disease assessment test. ***Saudi Med J*** 2011; 32: 1028-1033.
2. Bland JM, Altman DG. Cronbach's alpha. ***BMJ*** 1997; 314: 572.
3. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. ***Psychol Bull*** 1979; 86: 420-428.
4. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. ***Sports Med*** 1998; 26: 217-238.
5. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. ***Lancet*** 1986; 1: 307-310.
6. Bland JM, Altman DG. Measuring agreement in method comparison studies. ***Stat Methods Med Res*** 1999; 8: 135-160.
7. Koch GG. Intraclass correlation coefficient. In: Kotz S, Johnson NL. Encyclopedia of Statistical Sciences. 4th ed. New York (NY): John Wiley & Sons; 1982. p. 213–217.
8. Stubbings S, Robb K, Waller J, Ramirez A, Austoker J, Macleod U, et al. Development of a measurement tool to assess public awareness of cancer. ***Br J Cancer*** 2009; 101 Suppl 2: S13-S17.
9. O'Donnell DE, Travers J, Webb KA, He Z, Lam YM, Hamilton A, et al. Reliability of ventilatory parameters during cycle ergometry in multicentre trials in COPD. ***Eur Respir J*** 2009; 34: 866-874.
10. Nyitray AG, Kim J, Hsu CH, Papenfuss M, Villa L, Lazcano-Ponce E, et al. Test-retest reliability of a sexual behavior interview for men residing in Brazil, Mexico, and the United States: the HPV in Men (HIM) Study. ***Am J Epidemiol*** 2009; 170: 965-974.
11. De Vera MA, Ratzlaff C, Doerfling P, Kopec J. Reliability and validity of an internet-based questionnaire measuring lifetime physical activity. ***Am J Epidemiol*** 2010; 172: 1190-1198.
12. Lachin JM. The role of measurement reliability in clinical trials. ***Clin Trials*** 2004; 1: 553-566.

Related Articles

Al-Qahtani MM, Hagr AA. A preliminary study of endoscopic acoustic stapedial reflex in chronic otitis media. ***Saudi Med J*** 2010; 30: 900-903.

Abu-Alshiekh NK, Kofahi SM, Nusair ZM. The use of sweat chloride test for screening cystic fibrosis among malnourished children suffering from frequent respiratory infections. ***Saudi Med J*** 2009; 30: 1526-1531.

Noori NM, Mehralizadeh S, Khaje A. Assessment of right ventricular function in children with congenital heart disease. *Doppler tissue imaging.* ***Saudi Med J*** 2008; 29: 1168-1172.