Commentary: Reliability in research

Clinicians perform several measurements for assessing the disease severity, deciding the treatment plan, and assessing the treatment outcome. The trustworthiness of a test is measured by its reliability as well as the ability of clinicians to replicate it. The "Reliability" of a scoring system is defined as the stability or consistency of the measurement method. In other words, it describes the ability of a particular test to produce similar results in different circumstances.[1]

The clinical measurements are rarely perfect as all the instruments and observers have some internal inconsistency. Hence, any observed score (O) can be considered as a function of two components, that is, O = T (true score) ± E (measurement error). Hence, a test with a reliability of 0.9 means that 90% of the observed score is true, whereas the rest 10% is due to error. Measurement errors are of two types, that is, systematic and random. Systematic errors are predictable errors that are usually unidirectional, constant, and biased, for example, the learning effect. Usually, re-tests in such a situation are consistently higher than the prior tests. Such errors affect the validity of a test and not its reliability. Random errors occur due to chance and are unpredictable and affect the reliability of a test.[1]

There are four main types of reliability, that is, test–retest, interrater, parallel forms, and internal consistency. *"Test–retest"* reliability measures the consistency of a test when it is repeated on the same sample at different points of time. It is applicable in situations that either do not involve raters or the rater effect is neglectable, for example, questionaries. *"Intrarater reliability"* measures the consistency of a test when it is repeated by the same rater. *"Interrater reliability"* measures the degree of agreement between different researchers assessing the same thing. *"Parallel forms"* reliability measures the correlation between two tests that are designed to measure the same thing. *"Internal consistency"* measures the correlation between multiple items of a test that are intended to measure the same variable.[2]

We congratulate the authors for describing and validating their inflammatory score system for grading infectious endophthalmitis.[3] They measured the intrarater and interrater reliability using the interclass correlation coefficient (ICC). ICC was first introduced in 1954 by Fisher *et al*.[2,4] McGraw and Wong have defined 10 forms of ICC on the basis of their "model," "type," and "definition."[2,4,5] It is important to select the correct form of ICC for any study evaluating the reliability of a test/score.

There are three types of "*models,*" the selection of which depends on the characteristics of the raters. *One-way*

random-effects model (Model 1) is used when each group of subjects is rated by a different set of raters. This model is rarely used as most of the studies involve the same set of raters, except for multicentric trials. A *two-way random-effects model (Model 2)* is used when the raters are selected randomly from a larger set of raters with similar characteristics. It is used for generalizing the reliability results to all the raters with similar characteristics such as retina surgeons with the same years of experience. This is the most commonly used model. A *two-way mixed-effects model (Model 3)* is used if the selected raters are the only raters of interest and can be used to assess intrarater reliability.[2,4,5]

The selection *of "type"* depends on the number of readings taken for the measurement, that is, *"single"* or *"mean of multiple measurements."* Averaging is expected to reduce the variability among scores, thus ICC values based on the mean of multiple measurements will always be higher than values based on single measures. Hence, inappropriate "type" selection can produce inaccurate are perhaps better and impressive results, leading to the well-known positive publication bias.[2,4,5]

The selection of *"definition"* depends on the type of relationship, that is, *absolute agreement* (if raters assign exactly the same score) and *consistency* (if raters' scores are correlated in an additive manner). In other words, consistency merely reflects the extent to which the two sets of scores have a similar sequence when arranged in ascending order. For example, the score set "10, 20, 30, and 40" has the same sequence as the set "20, 30, 40, and 50." They are perfectly correlated; however, the two sets are not identical to each other. This difference is seen when calculated is done with the absolute agreement option. The value produced with the consistency option is usually higher than with the absolute option and can be misleading. Both types of definitions, that is, consistency and absolute, are available within models 2 and 3, whereas only the absolute option is relevant for model 1.[2,4,5]

The calculation in different forms of ICC is based on different assumptions and leads to different interpretations. The same set of data can produce different results depending on the selections made for the analysis. Hence, it is imperative to mention details related to the software, model, type, and definition while reporting ICC. The readers should check for all this information while interpreting ICC values reported in any research article. It has been suggested that at least three raters should perform the measurement on a sample size of at least 30 to get meaningful results. ICC values <0.5, 0.5–0.75, 0.75–0.9, and >0.9 are indicative of poor, moderate, good, and excellent reliability, respectively. The 95% confidence interval (CI) range may involve two or more grades of reliability. For example, the level of reliability for 95% CI of an ICC score of 0.76–0.94 can be regarded as "good' to "excellent."[2,4,5]

Naresh Babu, Piyush Kohli[1]

Department of Vitreo-Retinal Services, Aravind Eye Hospital and Post Graduate Institute of Ophthalmology, Madurai, Tamil Nadu, [1]Department of Vitreo-Retinal Services, C. L. Gupta Eye Hospital, Moradabad, Uttar Pradesh, India

Correspondence to: Dr. Naresh Babu, Department of Vitreo-Retinal Services, Aravind Eye Hospital and Post Graduate Institute of Ophthalmology, Madurai, Tamil Nadu, India.
E-mail: cauveryeye@gmail.com

References

1. Bruton A, Conway J, Holgate S. Reliability: What is it, and how is it measured? Physiotherapy 2000;86:94-9.
2. Koo TK, Li MY.A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155-63.
3. Dave VP, Belenje A, Dogra A, Das T, on behalf of the EMS working group. Application and validation of a novel inflammatory score in the clinical grading of infectious endophthalmitis: The endophthalmitis management study – Report 2. Indian J Ophthalmol 2023;71:396-400.
4. McGraw KO, Wong SP. Forming inferences about some intraclasscorrelation coefficients. Psychol Methods 1996;1:30-46.
5. Trevethan R. Intraclass correlation coefficients: clearing the air, extending some cautions, and making some requests. Health Serv Outcomes Res Method 2017;17:127-43.

Access this article online

Quick Response Code:

Website: www.ijo.in

DOI: 10.4103/ijo.IJO_2016_22

Cite this article as: Babu N, Kohli P. Commentary: Reliability in research. Indian J Ophthalmol 2023;71:400-1.