Different types of reliability

Table 1. Different types of reliability, when they are used, how they are computed, and what they mean

Type of reliability	When you use it	How you do it – an example	The result	What question you should answer when you got the result?
Test-retest reliability	When you want to know whether a test is reliable and stable over time. Test-retest reliability is necessary when you are exploring differences or changes over time.	You examine preferences of a group of 10 pupils on the 8th grade for different types of vocational programs and you administer a test in September and then repeat the same test again in June on the same pupils. Then, the two sets of scores (at time 1 varying from 54 to 98 and at time 2, ranging from 56 to 99) are correlated to measure reliability.	Procedure: Correlate the scores from a test given at Time 1 with the same test given at Time 2. In this case, computing the Pearson correlation coefficient is a measure of the test-retest reliability of the instrument. The Pearson product-moment correlation is equal, let's say, with 0.90.	Is the test measuring the preferences for different types of vocational programs reliable over time?
Parallel forms reliability	When you want to examine the equivalence or similarity between two different forms of the same test or assessment tool. Both versions must contain items that probe the same construct, skill, knowledge base, etc., to the same group of individuals to evaluate the consistency of results across alternate versions.	Suppose you apply a test in order to train memory skills, called I Remember Memory Test (IRMT) on a group of 10 elderly, this way: each individual look at 10 different words, memorize them as best s/he can, and s/he recites them back after 20 seconds of study and 10 seconds of break. You repeat the test after two days, when you prepare another list of 10 words related to the same ideas, that you apply in exactly the same manner. At the first test, you register the number of recited words from each participant, the scores ranging from 3 to 7 in a Form A. You repeat the procedure at the second test and register the scores varying from 5 to 8 in a Form B.	Procedure: Correlate the scores from the first form of the test with the scores from the second form of the same test of a similar content, but not the exact same test. In this case, you will correlate the scores from the two versions of the IRMT test. Here, after computing the Pearson correlation coefficient as a measure of the parallel forms reliability of the instrument, you obtained, let's say, a value of 0.13.	Are the two forms of the IRMT test equivalent to one another? Do they have shown parallel forms reliability?

With the support of the Erasmus+ Programme of the European Union

Internal consistency reliability	When you want to know if the items on a test assess one, and only one, dimension, construct, or area of interest. It is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results.	Let's say you developed a test of attitudes towards different types of health care, made of five items where scores ranged from 1 (*strongly disagree*) and 5 (*strongly agree*) on each item (for example, "I don't like spending money on health care."). Now, you explore if individual scores recorded from 10 adults correlate with the total score.	Procedure: Correlate each individual item score with the total score. In this case, when you compute Cronbach's alpha, you correlate the score for each item with the total score for each individual, and then compare that with the variability existent for all individual item scores. In this logic, any individual test taker with a high total test score should have a higher score on each item and conversely, any individual test taker with a lower total test score registered a lower score on each individual item. Let's say that after applying the formula to calculate Cronbach's alpha, you obtained the result 0.24.	Do all the items on the test of attitudes towards different types of health care assess the same construct?
Interrater reliability	When you want to know whether there is consistency in the rating of some outcome. It is useful because human observers/ raters do not necessarily interpret answers of an assessment instrument the same way; they may disagree about how well certain responses	Let's imagine you are interested in the type of social interaction (the supportive attitude) between a student at social work in practice at a town hall and a client who ask for support to complete a form for heating aid during the winter. Two observers will note, within a set of 10-second time frames across 2 minutes (or twelve 10-second periods), whether the student demonstrates one of the three different behaviours he has been trained for in the practice stage – smiling, leaning forward in his chair, or using his hands to make a point. Each time the raters	Procedure: Examine the percentage of agreement between raters/ judges. In this case, for a total of 12 periods (12 possible agreements), let's say there were 7 where both raters agreed that such behaviours took place, 3 where they agreed no such behaviours occurred, and 2 disagreements in what the raters observed. Interrater reliability is computed using a simple formula, by dividing the number of agreements (here 10) to the number of	How much the two raters agree on their judgements related to the supportive attitude of the student in the social interaction with a client?

	demonstrate knowledge of the construct being assessed.	see any of these behaviours, they mark it on a scoring sheet with an X. If they observe none of these behaviours, they mark a dash (–). Part of this process is to find out what the level of agreement is between the two observers related to the occurrence of these behaviours? The more similar the scores are, the higher the level of interrater agreement and interrater reliability.	possible agreements (here 12). So, the resulting interrater reliability coefficient is 0.83.	

Sources: The content of the table is adapted from Salkind (2017) and Phelan and Wren (2005-6)[1].

Notes:

1. If you want to see in more details how the concrete examples given in this table are approached in an empirical manner, to see the individual scores and the concrete way of applying and computing the formula for each reliability measure, you should read Salkind (2017: 163-175).
2. The choice of 10 participants for exemplification at the first three measures of reliability is arbitrary. I used this number for the investigated group as the author had used, for those of you who want to explore in depth the examples used in the cited book.

[1] Exploring reliability in academic assessment, written by Phelan, C. and Wren, J. Graduate Assistants, UNI Office of Academic Assessment (2005-06), https://chfasoa.uni.edu/reliabilityandvalidity.htm, accessed in 20 February 2022.