What is reliability?

In psychometric testing, a reliable test is one which will produce consistent results when the same individual is tested on different occasions (often referred to as 'test-retest' reliability). Often, psychologists and statisticians analyse the internal consistency of a test to verify that different parts of a test are all measuring the same quality or trait (in this case, they are measuring 'internal consistency'). When evaluating a test, reliability is generally measured before validity, since the reliability of a test places an upper limit on its validity – in other words, a test cannot measure what it purports to measure unless it is stable and consistent in its measurement.

Test-retest reliability

Belbin is not a psychometric test, since it measures behaviour rather than personality. Whilst personality may remain fairly consistent, we would expect behaviour to change over time, especially for young people. Indeed, it is desirable that behaviours should be adapted to fit new circumstances and as young people try out different roles and activities. Although it is unusual for a Team Role order to become entirely reversed, it is highly likely that preferred and manageable roles may move around during an individual's career.

Internal consistency

Internal consistency is also known as scale homogeneity, in other words, the ability of items in a scale to measure the same construct or trait. Belbin measures nine Team Roles, which are clusters or constellations of behavioural characteristics, rather than individual traits. As a result, Belbin does not repeat items but rather looks to identify strata of different Team Role behaviours displayed by an individual.

Since Belbin differs from psychometric tests in both its underlying theory and its format and scoring mechanism, measuring its reliability via Cronbach's alpha (a) has presented a challenge to researchers over the years. Firstly, the Belbin Get Set inventory is neither fully ipsative nor non-ipsative, so whilst the total score achieved is always the same, the score for each Team Role can vary. Secondly, respondents do not allocate a value to every item in the inventory.

Having recognised these difficulties, Swailes et al. formulated and proposed a new measure of reliability, $IR\alpha$, which offers a weighted mean of average inter-item correlation scores. The results of these calculations for the latest Get Set inventory are shown in the table below:

Team Role	n	$IR\alpha$
CF	92	0.776
CO	49	0.756
IMP	84	0.791
ME	60	0.888
PL	67	0.797
RI	24	0.858
SH	50	0.733
SP	38	0.625
TW	93	0.823

n indicates number of candidates in sample size. A score of 0.6 or above is considered good or acceptable.