Interrater Reliability

Martyn Shuttleworth84K reads

For any research program that requires qualitative rating by different researchers, it is important to establish a good level of interrater reliability, also known as interobserver reliability.

This ensures that the generated results meet the accepted criteria defining reliability, by quantitatively defining the degree of agreement between two or more observers.

Interrater Reliability and the Olympics

Interrater reliability is the most easily understood form of reliability [1], because everybody has encountered it.

For example, watching any sport using judges, such as Olympics ice skating or a dog show, relies upon human observers maintaining a great degree of consistency between observers. If even one of the judges is erratic in their scoring system, this can jeopardize the entire system and deny a participant their rightful prize.

Outside the world of sport and hobbies, inter-rater reliability has some far more important connotations and can directly influence your life.

Examiners marking school and university exams are assessed on a regular basis, to ensure that they all adhere to the same standards. This is the most important example of

interobserver reliability - it would be extremely unfair to fail an exam because the observer was having a bad day.

For most examination boards, appeals are usually rare, showing that the interrater reliability [2] process is fairly robust.

An Example From Experience

I used to work for a bird protection charity and, every morning, we went down to the seashore and used to estimate the number individuals for each bird species.

Obviously, you cannot count thousands of birds individually; apart from the huge numbers, they constantly move, leaving and rejoining the group. Using experience, we estimated the numbers and then compared our estimate.

If one person estimated 1000 dunlin, one 4000 and the other 12000, then there was something wrong with our estimation and it was highly unreliable.

If, however, we independently came up with figures of 4000, 5000 and 6000, then that was accurate enough for our purposes, and we knew that we could use the average with a good degree of confidence.

Qualitative Assessments and Interrater Reliability

Any qualitative [3] assessment using two or more researchers must establish interrater reliability to ensure that the results generated will be useful.

One good example is Bandura's Bobo Doll experiment [4], which used a scale to rate the levels of displayed aggression in young children. Apart from extensive pre-testing, the observers constantly compared and calibrated their ratings, adjusting their scales to ensure that they were as similar as possible.

Guidelines and Experience

Interobserver reliability is strengthened by establishing clear guidelines and thorough experience. If the observers are given clear and concise instructions about how to rate or estimate behavior, this increases the interobserver reliability.

Experience is also a great teacher; researchers who have worked together for a long time will be fully aware of each other's strengths, and will be surprisingly similar in their observations.

Bibliography

Auerbach, C., La Porte, H.H. & Caputo, R.K. (2004). Statistical Methods for Estimates of Interrater Reliability. In Roberts, A.R. & Yeager, K.R. *Evidence Based Practice Manual: Reasearch and Outcome Measures in Health and Human Services,* pp 444-448, New York, NY: Oxford University Press

Jackson, S.L. (2011). *Research Methods and Statistics: A Critical Thinking Approach* (2nd

Ed.). Belmont, CA: Wadsworth Cengage Learning

Rubin, A., & Babbie, E.R. (2007). *Essential Research Methods for Social Work,* Belmont, CA: Wadsworth Cengage Learning

---

Links
[1] https://explorable.com/validity-and-reliability
[2] http://en.wikipedia.org/wiki/Inter-rater_reliability
[3] https://explorable.com/qualitative-research-design
[4] https://explorable.com/bobo-doll-experiment