

Lecture 10: Reliability

Dewayne E Perry

ENS 623

perry@ece.utexas.edu

Errors in Measurement

- All measurements subject to fluctuations
 - ↳ Affects reliability and validity
- *Reliability* : constancy or stability
- *Validity* : appropriateness or meaningfulness
- *Reliability coefficient* : degree that what is measure is free from measurement fluctuation
- *Observer agreement coefficient* : objectivity and repeatability of rating procedures
- Random vs systematic errors
 - ↳ Random: cancel out on average over repeated measurements
 - ↳ Systematic: do not cancel out
- Systematic errors are known as *Biases*
 - ↳ Main concern of *internal validity*
 - ↳ Can compensate for known biases
 - Eg, in astronomy, known biases of observations

Reliability Criteria

→ Principle criteria of test reliability

- ↳ Test-retest reliability
- ↳ Reliability of test components
 - ie internal consistency

→ Stability (Test-Retest)

- ↳ Temporal stability from one session to the next
- ↳ Problem: distinguishing between real change and the effect of memory
 - Too short an interval between: memory effect possible
 - Too long an interval: real changes may interfere
 - May use changes to test sensitivity of tests

Reliability (of Test Components)

- Internal consistency reliability
- Depends on the average of Intercorrelations among all the single test items
- Coefficients of internal consistency increase as the number of test items goes up (if the new items are positively correlated with the old)
- The more items, the more internally consistent the test; if other relevant factors remain the same
 - ↳ Not always the same for different length tests
 - ↳ Boredom & fatigue can result in attenuation

Spearman Brown Formula

$$\rightarrow R = \frac{n\bar{r}}{1 + (n-1)\bar{r}}$$

↳ R is the reliability coefficient

↳ n is the factor by which the test is lengthened

↳ \bar{r} is the mean correlation among all items

→ Suppose mean correlation is .50, determine reliability of test for twice, thrice:

↳ $2(.50)/[1+(2-1).50] = .667$

↳ $3(.50)/[1+(3-1).50] = .75$ - increase R by half

→ Other Tests

↳ *Kuder-Richardson formula 20 (K-R 20)*

➤ Used to measure internal consistency when items of the test are scored 1 if marked correctly, 0 otherwise

↳ *Cronbach's alpha coefficient*

➤ Employ the use of analysis of variance procedures for estimating reliability of test components

Acceptable Reliability

- Need to evaluate whether low validity is due to low reliability
 - ↳ If so can it be improved by adding items
- What is the acceptable range of reliability?
 - ↳ Depends on situation and nature of variable being measured
 - ↳ For clinical testing $R = .85$ is considered as indicative of dependable psychological tests
 - ↳ In experimental research, accept much lower R
- Problem:
 - ↳ Reliability test reflects both individual differences and measurement fluctuations
 - ↳ If everyone alike, the only differences are in error variations
 - ↳ Hence, lower reliability where fewer differences
 - Eg, IQ at highly selective where students are more similar than at a public university

Acceptable Reliability

→ Reliabilities of major psychological tests

↳ MMPI - MN Multiphasic Personality Inventory

↳ WAIS - Winchester Adult Intelligence Scale

↳ Rorschach inkblot test

→ MMPI and Rorschach most widely used, WAIS used as control

→ Internal consistency - all three acceptable

↳ WAIS $R = .87$, 12 studies with 1759 subjects

↳ MMPI $R = .84$, 33 studies with 3414 subjects

↳ Rorschach $R = .86$, 4 studies with 154 subjects

Acceptable Reliability

→ Stability - respectable scores

↪ Fewer studies available

↪ WAIS as .82 - 4 studies with total $N = 93$

↪ MMPI as .74 - 5 studies with total $N = 171$

↪ Rorschach as .85 - 2 with total $N = 125$

➤ WAIS/Rorschach difference not significant;

➤ MMPI/Rorschach and WAIS/MMPI difference is highly significant

→ Internal consistency usually higher than stability

→ Problem of inter-rater reliability

↪ Use test reliability measures to assess their aggregate internal consistency

↪ Arises in SWE in classifying faults, root causes, evaluating designs, reviewing papers, evaluating developers, etc

Effective Reliability of Judges

- Problem: correlation of .60 between the ratings of two judges tells us only the reliability of either single judge in this situation
- For aggregate or effective reliability, use approach as in “how many test items”
 - ↳ Use *Spearman-Brown* where
 - n is the number of judges and
 - \bar{r} is the mean correlation among them
 - ↳ Aggregate reliability of
 - 2 judges: $2(.60)/[1+(2-1).60] = .75$
 - 3 judges: $3(.60)/[1+(3-1).60] = .82$
 - ↳ The more judges, the higher the reliability
 - ↳ Table 3.3 very useful for planning/analysis

% Agreement & Reliability

- Many use percent agreement as an index of reliability
 - ↳ A agreements and D disagreements
 - %: $[A/(A+D)] \times 100$
 - Net: $[(A-D)/(A+D)] \times 100$
- Misleading - fails to differentiate between accuracy and variability
- Better - use the product moment correlation ϕ
 - ↳ can be computed from the *chi-square*

ANOVA & Reliability

- Sometimes need more than 2-3 judges
- Excellent approach based on analysis of variance
 - ↳ Tedious to do average of large number of correlations of previous approach
 - ↳ Assess how well judges are able to discriminate among sampling units ($MS_{persons}$) minus the judge's disagreements ($MS_{residuals}$) controlling for rating bias or main effect, divided by a standardizing quantity

$$R_{est} = \frac{MS_{persons} - MS_{residuals}}{MS_{persons}}$$

$$\bar{r}_{est} = \frac{MS_{persons} - MS_{residuals}}{MS_{persons} + (n-1)(MS_{residuals})}$$

Replication & Reliability

→ Reliability in research implies generalizability as indicated by replicability (repeatability) of the results

- ↪ Across time (*test-retest reliability*)
- ↪ Across different measurements, observers, or manipulations (*reliability of components*)
- ↪ Note that may not be possible to repeat and authenticate every observation with perfect precision

Replication & Reliability

→ Comparison of literature reviews in behavioral sciences (BS) and physics (P)

- ↳ Compare and assess the consistency of results in the soft and hard sciences
 - BS: variety of hard, middle and soft areas
 - P: properties of elementary particles (one of most elite areas - best physicists work here)

- ↳ Computed a reliability coefficient (*Birge's ratio*)
 - Estimates differ from one another other than randomly
 - Consistency varied as much in P (2.11) as in BS (2.09)

Replication & Reliability

→ Hedges 1987 continued

↳ Physics

- Striking discrepancies in measurements
- 25/64 in error by over 10%, 16 > 30%, 8 > 50%, 2 > 100% and 1 by 245%
- Methodological practice of omitting a large proportion of studies to get a consistent sample

↳ Behavioral Sciences

- Some difficulty due to measurement fluctuations
- Some to limitations of expression or language
 - ✓ Tacit knowledge or "unvoiceable" wisdom
 - ✓ Ie, did not carry out the equipment properly

Replication Factors

→ *Same* experiment can never be repeated

↳ At very least everyone is older

→ 3 important factors affect the utility of a replication as an indicator of reliability:

↳ When the replication is conducted

➤ Earlier better than later; 2nd doubles our info

↳ How the replication is conducted

➤ The more imprecise, the more generalizability

↳ By whom is the replication conducted

➤ Independence is critical - rule out pre-correlations

➤ Selection and training considerations

➤ Correlated observers a critical problem in all fields