1. **Reliability – scoring consistency for NRT – relative decisions**

Observed score = product of true test + measurement error.

1. First choose the reliability target level
2. Types of reliability:
3. Issued twice - correlation between two tests
4. Test-retest reliability (the same test)
5. Parallel forms reliability (two versions of the same test)
6. Issued once
7. Internal consistency reliability (the extend to which each part of the test measures the same thing or the degree to which they measure it in the same way)
8. Split-half reliability (splitting the test in half. REMEMBER: each part will be shorter than the test itself so reliability will be lower which affects correlation.)

* Spearman-Brown prophecy formula – also answers how many items should be added to increase the reliability (ex. M = 1.5 then the length needs to be increased by 50%)
* Cronbach’s alpha.

Needs applying **correction for attenuation** (formula)and **SEM** (standard error of measurement).

To find the confidence interval:

68% confidence that true score is within 1 SEM of the observed score

95% confidence that it is within 2 SEM

99% confidence that it is within 3 SEM

If the questions on the test all measuring the same thing, then they should be highly correlated with each other and with the total test score.

**RULL OF THUMB:** it is easier to achieve an adequate reliability level on item-based test when there are at least 20-30 well-written items.

1. **Dependability – classification consistency for CRT – absolute decisions**
2. Consistency scoring (similar to reliability for NRT)
3. *Generalizability theory – too advanced. Check R or Python packages*
4. Ф (CRT dependability index) using Cronbach’s alpha (only for dichotomous data) using proportion-correct metric.

CIcrt also using proportion-correct metric. Similar interpretation as SEM.

1. Consistency for classification (classifications of students as masters/non-masters)

Ф(cut score) also using Cronbach’s alpha (only for dichotomous data).