<table>
<thead>
<tr>
<th>Method for Estimating Reliability</th>
<th>Other Names or Specific formulas</th>
<th>Description</th>
<th>Sources of Error</th>
<th>When is it Used</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| Test-Retest Reliability          | Coefficient of Stability        | Administering same test to same group of examinees on two different occasions then correlate the results | • Time sampling error  
• Random variations in test situation | To indicate the degree of stability of examinee's scores over time | • Attributes that fluctuate over time  
• Attributes that are affected by repeated measurement |
| Alternate (Equivalent, Parallel) Forms Reliability | - Coefficient of Equivalence (when forms administered at the same time)  
- Coefficient of Stability and Equivalence (when a relatively long period of time separates administrations) | Two equivalent forms of the test are administered to the same group of examinees and the two sets are correlated  
Considered the **most thorough method** for estimating reliability | • Content sampling  
(error introduced by an interaction between different examinee's knowledge and the different content assessed  
• Time sampling error | To indicate the consistency of responding to different item samples | • Attributes that fluctuate over time  
• Attributes that are affected by repeated measurement |
| Internal Consistency Reliability | Split-Half Reliability | Test is split into equal halves so that each examinee has two scores and these are correlated (usually underestimates - is corrected using **Spearman-Brown Prophecy Formula**) | • Content sample error  
(resulting from knowledge in one half fitting an examinee better) | When test is designed to measure single characteristic | • Reliability of speed tests |
| Cronbach's Coefficient Alpha (Use **Kuder-Richardson Formula 20 KR-20** when test items are dichotomous) | Whole test is administered once, formula used to determine average degree of inter-item consistency (Average reliability obtained from all possible splits of the test)  
• Tends to be conservative and considered the lower boundary of the tests reliability | • Content sampling  
(differences between individual test items  
• Heterogeneity of content domain (greater the heterogeneity the lower the inter-item correlations) | | | |
| Inter-Rater Reliability | Correlation Coefficient | Uses the **kappa statistic** when scores or rating are nominal or ordinal (coefficient of concordance when there are 3 or more raters and ratings are ranks) | • Factors related to the raters (motivation, characteristics of measuring device)  
• Consensual observer drift – when observers influence each other in idiosyncratic ways | When test scores depend on a rater’s judgment |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Agreement</td>
<td>Calculated by dividing number of items or observations in agreement by total number of items or observations</td>
<td>• Does not take into account level of agreement occurring by chance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>