EDUR 7130
Presentation 6a
Reliability

1. Reliability Defined

Reliability has two meanings, that measured scores are ***consistent***, and/or measured scores **agree**. Consistency is the degree to which sets of scores show similar patterns, and agreement is the extent to which scores are similar, i.e., differences between scores are small in absolute value.

Example 1

Weight scales measure me as

205 205 205

Those weights are very consistent and agree in measurement.

Reliability can vary by ***degree***.

Example 2

Weight scales measure me as

205 206 204

This shows less consistency and agreement than the earlier weights, so less reliability.

Example 3

Weight scales measure me as

112 285 59

Much less consistent and much less agreement, so no reliability.

Reliability is the degree to which we obtain consistent scores that agree from some instrument or measuring device.

2. Classic Test Theory (CTT) and the Reliability Coefficient

Note: You do NOT have to learn these formulas or calculations shown below or shown in any part of this presentation; these are presented to help explain how reliability is derived.

Reliability coefficient is a mathematical index that indicates the degree to which scores are consistent or agree.

Reliability coefficient ranges from 0 to 1, with values closer to 1 indicting greater reliability.

CTT tells us that when we attempt to measure something, like test anxiety, we understand that the score we observe, the observed score X, is made of two parts, a true score (T) and error (E):

$$X = T + E$$

We would like to know how much error, E, is included when we use observed scores, X, because the more error, the worse our measurement and the less confidence we have that X measures what we hope it measures.

Since there will almost always be variability in scores, we can say that the variance for scores will be greater than 0.00. If we use the symbol X for test anxiety scores, we can indicate the variance like this:

VAR(X)

We can also expect variance in both true scores, T, and error in measurement, E, so we can symbolize these variances too:

VAR(T) and VAR(E)

Reliability is defined as the ratio of true score variance to observed score variance:

$$\text{Reliability, rxx} = \frac{\text{VAR(T)}}{\text{VAR(X)}}$$

Since X = T + E, we can show that reliability is the ratio of true score variance to true score variance plus error variance:

$$\text{Reliability, rxx} = \frac{\text{VAR(T)}}{\text{VAR(T)} + \text{VAR(E)}}$$

Reliability is the

- proportion of true score variance to observed score variance;
- should not be less than 0.00;
- should not be greater than 1.00;
- r or rxx or r_{xx} is sample symbol for reliability,
- ρ or ρxx or ρ_{xx} is population symbol for reliability,
- unfortunately, r and ρ are also symbols for Pearson correlation, so easy to confuse the two.

If there were no error in measurement, then VAR(E) would be zero, VAR(E) = 0.00, and reliability would be equal to 1.00:

$$= \frac{\text{VAR(T)}}{\text{VAR(T)} + \text{VAR(E)}}$$

$$= \frac{\text{VAR(T)}}{\text{VAR(T)} + 0}$$

$$= \frac{\text{VAR(T)}}{\text{VAR(T)}} = 1.00$$

A reliability of 1.00 means no measurement error and therefore we have true scores.

Assumptions of CTT:

- Expected value of E = 0.00 (i.e., mean of errors will be 0.00)
- Covariance T and E = 0.00; Cov(T,E) = 0.00 (i.e., correlation of T with E = 0.00)
- Covariance Ej and Ek = 0.00, Cov(Ej,Ek) = 0.00 (i.e., correlation of Ej with Ek = 0.00)

In words, CTT indicates that measurement error, E, is random and therefore correlates with nothing; if E does show a correlation with something, it will likely be small correlation that is random (i.e., varies across samples and due to sampling variation).

Technical note:

VAR(X) = VAR(T) + VAR(E) + 2Cov(T,E)

Since E does not correlate with anything,

VAR(X) = VAR(T) + VAR(E) ~~+ 2Cov(T,E)~~

Question

Which of the following coefficients indicates greater reliability?

.53
.78 ←
.08

3. Methods of Assessing Reliability

There are many formal ways to determine reliability of scores. Below are a few of the more commonly employed approaches.

3.1 Test-retest Reliability with Pearson Correlation Coefficient, r

Test-retest reliability assesses the **stability** of scores across time; the more stable scores over time, the greater the consistency, so the greater reliability.

Test-retest reliability establishment steps:

- Administer the ***same measuring device*** (or some test, scale, inventory, measuring device) to the ***same group*** of people on two different occasions.
- Let appropriate time elapse between administrations (scale dependent).
- Obtain scores from the two occasions.
- Correlate the paired scores using Pearson's correlation coefficient.
- Show that mean scores from the two forms are similar. This last point is often overlooked but is critical.

Paired scores – each person will provide two scores, one score from the first administration and the second score from the second administration. Each set of scores should be paired to the same person like this:

Example 1: True Scores, Test Retest

Student	Test True Score	Re-test True Score
1	95	95
2	90	90
3	85	85
4	80	80
5	75	75
6	70	70
7	65	65
8	60	60

In the above example, true scores = observed scores, so

VAR(T) = 140.00 = VAR(X)

Note, the variance above represents the total variance for both administrations of the test and retest, so 16 observations, not 8.

Reliability of these two sets of scores is

$$rxx = \frac{VAR(T)}{VAR(X)} = \frac{140.00}{140.00} = 1.00$$

The Pearson correlation for these two scores is r = 1.00.

Example 2: True Scores with Error Added

Student	True Score	Error Time 1	Error Time 2	Observed Time 1 (True + Error 1)	Observed Time 2 (True + Error 2)
1	95	3	-3	98	92
2	90	-3	-3	87	87
3	85	3	3	88	88
4	80	-3	3	77	83
5	75	-3	3	72	78
6	70	3	3	73	73
7	65	-3	-3	62	62
8	60	3	-3	63	57

Note: Cov(e1,e2) = 0.00, Cov(e1,T) = 0.00, Cov(e2, T) = 0.00; errors uncorrelated with each other and true scores.

How well does Pearson r work if "random" measurement error is introduced to true scores as shown in Example 2 above?

Variances for true scores and observed scores reported below.

VAR(T) = 140.00 (Variance of 16 true scores to mimic test and retest situation)
VAR(X) = 149.60 (Variance of both Time 1 and Time 2 observed scores combined)

The CTT reliability is

$$rxx = \frac{VAR(T)}{VAR(X)} = \frac{140.00}{149.60} = 0.935$$

which means that 93.5% of variance in observed scores is due to true score variance, or 100(1 - .935) = 6.5% is error variance.

Pearson correlation for these data is r = 0.935

Results show that Pearson r works well to measure reliability when only random measurement error is included, and the means for both sets of scores are the same or similar. In Example 2 above, the means for Observed Time 1 = 77.50 and for Observed Time 2 = 77.50. However, Pearson r can fail when non-random error is included that changes means between the two sets of scores.

Published Examples of Test-Retest Reliability

Example 1
Avanzi, L., Miglioretti, M., Velasco, V., Balducci, C., Vecchio, L., Fraccaroli, F., & Skaalvik, E. M. (2013). Cross-validation of the Norwegian teacher's self-efficacy scale (NTSES). Teaching and Teacher Education, 31, 69-78.

Purpose: Translate into Italian the Norwegian teacher self-efficacy scale.

Note: See presentation video for discussion of this information.

4.1.1. Italian sample page 72

The Italian data were collected in the context of a psychosocial risk assessment conducted in five Italian schools in Trento. We first contacted the principal of each school and explained the aim of our research. Teachers were then contacted in their schools, informed about the aims of the study, and asked for their consent to participate. In three of the five schools, teachers compiled the questionnaire during working hours (approximately 40 min). They were explicitly told to work individually without cooperating. In the two remaining schools, teachers completed the questionnaire at home. They were asked to give the completed questionnaire back within a two-week period by using a sealed envelope (which we provided) to be returned directly to the school.

The questionnaire was administered at two points of time, with an interval of six months. The questionnaire contained a code to identify the participants in the second wave; however, teachers were guaranteed that their responses would remain confidential. A total of 348 teachers (74.7% women) participated in the survey. They were employed in various levels of Italian schools (45.7% primary, 25.6% junior, and 16.1% high schools). The age of the participants varied from 23 to 60 ($M = 41.21$; $SD = 9.70$), while their job tenure ranged from 1 to 45 years ($M = 18.94$; $SD = 9.83$). Most of the teachers had a master's degree (52.3%) and a permanent job contract (52.6%). For 140 teachers, the second (Time 2) measure, which was taken six months after the first administration, was available.

5.3. Reliability of the Italian version of NTSES page 75

We calculated Cronbach's alphas of the Italian NTSES subscales at both Time 1 and Time 2 for the sub-sample ($N = 140$), and they were very good, ranging from .79 to .92. Furthermore, the test-retest reliability for the six subscales ranged from $r = .48$ (for 'cooperate with colleagues and parents') to $r = .64$ (for 'maintain discipline'), while the same index for the overall teacher's self-efficacy measure was $r = .65$. Moreover, we calculated the CR for each dimension of the Italian sample (at Time 1). They were very good, and similar to those obtained using the alphas, indicating that each NTSES subscale was reliable. The values of AVE for all dimensions were greater than .50, ranging from .52 to .73, evidencing that most of the variance in the items was explained by the underlying factor, rather than by the measurement error (for details regarding each dimensions see Table 3).

Test-retest reliability is not in Table 3 but is reported in text above. Table 3 provides other reliability information. Score agreement was not addressed in this study.

Table 3 page 75

Alphas *T1* and *T2*, CR and AVE for the six dimensions of NTSE in Italian sample.

NTSE dimensions	Cronbach alpha *T1* ($N = 348$)	Cronbach alpha *T2* ($N = 140$)	CR ($N = 348$)	AVE ($N = 348$)
1. NTSE instruction	.88	.83	.87	.63
2. NTSE adapt instruction to individual needs	.88	.85	.88	.64
3. NTSE cooperate with colleagues and parents	.81	.79	.81	.52
4. NTSE cope with change	.83	.81	.83	.56
5. NTSE motivate students	.87	.84	.88	.64
6. NTSE maintain discipline	.92	.91	.92	.73

Note. CR = Composite Reliability; AVE = Average Variance Extracted.

Example 2

Vega, E. M., & O'Leary, K. D. (2007). Test–retest reliability of the revised Conflict Tactics Scales (CTS2). Journal of Family Violence, 22(8), 703-708.

Purpose: Learn whether the Conflict Tactics Scale provides stable scores over time because scale "…stability of scores has not been studied." Participants were 82 men mandated to attend "a batterer intervention program" (p. 703).

Note: See presentation video for discussion of this information.

Procedure

page 705

Participants completed questionnaires as part of their participation in a batterer intervention program consisting of 18 sessions, held once per week. They were asked to complete the written questionnaires during their first session in the program, and again approximately nine weeks later. Due to the treatment groups utilizing an "open group" model, participants' first and tenth sessions did not all coincide with each other; participants completed questionnaires apart from the group, typically in a separate room. The data were collected in an anonymous manner; data sets were matched on the basis of identification numbers chosen by the participants themselves. The choice of nine weeks as an interval of assessment for test–retest reliability was chosen for the purpose of minimizing the subjects' recall of their precise questionnaire answers, while also minimizing any possible decrement in memory for the actual events being queried. There was a possibility that being in a batterer intervention program would result in reports of fewer acts of physical aggression following treatment even though the time period in question was before the program began, namely reports of aggression in the year prior to the initial assessment. However, as will be seen in the means on the various scales at the initial and second assessment, reports of physical aggression did not change from the first report of the aggression to the second report.[1]

Table 1 Prevalence and chronicity of self-reported male aggression (CTS2)

Scale	Time 1	Time 2
Negotiation		
Prevalence	96.3%	96.3%
Chronicity		
Mean	76.18	75.81
SD	40.04	38.26
Psychological Aggression		
Prevalence	89.0%	93.9%
Chronicity		
Mean	35.22	33.42
SD	27.26	32.99
Physical Assault		
Prevalence	70.7%	67.1%
Chronicity		
Mean	11.10	9.55
SD	15.34	15.49
Sexual Coercion		
Prevalence	24.4%	24.4%
Chronicity		
Mean	5.25	9.00
SD	6.30	14.08
Injury		
Prevalence	48.8%	46.3%
Chronicity		
Mean	6.73	4.24
SD	9.98	7.24

N=82

Assess group level agreement of scores

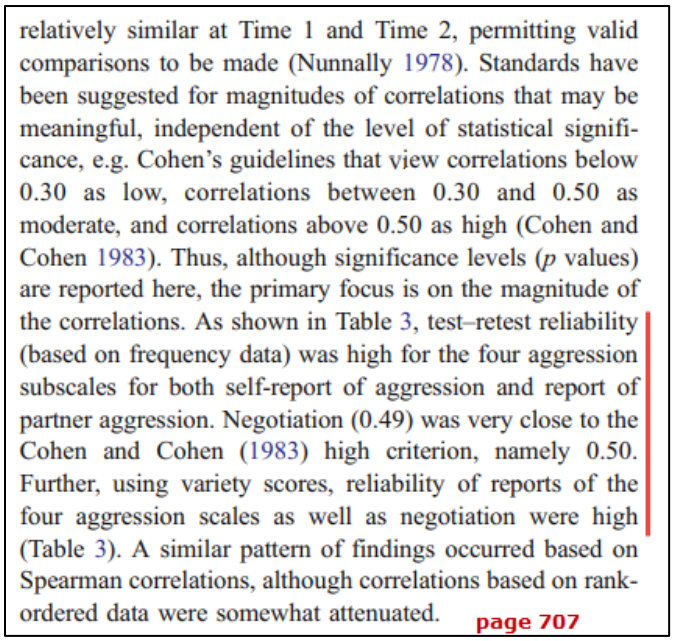
Time span is described above, and comparison of mean or tallied performance scores are presented in Table 1. These comparisons were used to help establish equivalence (agreement) of scores, something Pearson r cannot do. Test-retest Pearson correlations are reported in Table 3.

Table 3 Test–retest reliability of men's self-report of aggression, and report of partner aggression (Pearson product–moment), based on frequency scores and variety scores

Scale	Frequency scores		Variety scores	
	Self-report	Partner-report	Self-report	Partner-report
Negotiation	0.486**	0.617**	0.602**	0.672**
Psychological aggression	0.716**	0.650**	0.693**	0.708**
Physical assault	0.677**	0.863**	0.762**	0.776**
Sexual coercion	0.666**	0.798**	0.303*	0.538**
Injury	0.794**	0.531**	0.699**	0.718**

N=82
**$p<0.001$
*$p<0.005$

Unfortunately, their interpretation of the level of test-retest agreement appears to be incorrect. They cite work by Cohen and Cohen (1983) who I think discussed how to interpret correlation coefficient sizes which are distinct from reliability coefficients. A test-retest correlation of .486 is very low and indicates this sub-scale (negotiation) produces scores that are not stable across time.


relatively similar at Time 1 and Time 2, permitting valid comparisons to be made (Nunnally 1978). Standards have been suggested for magnitudes of correlations that may be meaningful, independent of the level of statistical significance, e.g. Cohen's guidelines that view correlations below 0.30 as low, correlations between 0.30 and 0.50 as moderate, and correlations above 0.50 as high (Cohen and Cohen 1983). Thus, although significance levels (*p* values) are reported here, the primary focus is on the magnitude of the correlations. As shown in Table 3, test–retest reliability (based on frequency data) was high for the four aggression subscales for both self-report of aggression and report of partner aggression. Negotiation (0.49) was very close to the Cohen and Cohen (1983) high criterion, namely 0.50. Further, using variety scores, reliability of reports of the four aggression scales as well as negotiation were high (Table 3). A similar pattern of findings occurred based on Spearman correlations, although correlations based on rank-ordered data were somewhat attenuated. page 707

Pretest and Posttest vs Test-Retest

With experimental studies one will often see use of a pretest and posttest design. An instrument, such as an achievement test or scale of self-efficacy, will be administered to an experimental and control group before any treatments (e.g., novel instruction vs. traditional instruction) are introduced. This is a **pretest** and it is designed to obtain a baseline measure of whatever is being examined (e.g., achievement, self-efficacy, heart rate, etc.). After the pretest the treatments are introduced to both groups, and when that is complete the scale, test, or whatever measure is being used will be administered again to both groups and this is called a **posttest**. The goal of such studies is to learn whether the treatments cause changes in scores from pretest to posttest.

Studies that include pretests and posttest should not be confused with test-retest studies. With test and retest, the goal is to learn whether scores are stable over time. In such studies one does not wish to introduce any treatments that might cause scores to change between test and retest. Any factors that do cause scores to change will alter score stability over time and therefore reduce test-retest reliability.

If one wishes to use the same instrument in a pretest and posttest design, it is wise to first assess test-retest score stability with that instrument well before any pretest-posttest study is implemented. Only after ensuring the scale produces stable scores, via test-retest, should such scale be used in a pretest-posttest study.

If you read a pretest and posttest study that does not include test-retest information, know that you cannot use the pretest and posttest scores to assess test-retest reliability due to the treatment introduction and the assumption that at least one treatment is likely to produce changes in pretest to posttest scores, therefore assessment of score stability will be compromised.

3.2 Consistency vs Agreement

As noted above, it is critical that one (a) shows scores are consistent, and (b) shows that scores across time agree. Unfortunately, Pearson r alone cannot demonstrate agreement for test-retest reliability, as illustrated below.

Consistency refers to the relative position, rank order, of scores across two sets of scores. Consistency is an assessment of whether two sets of scores tend to rank order something in similar positions.

Agreement refers to the degree to which two sets of scores agree or show little difference in actual scores; the lower the absolute difference, the greater the agreement between scores.

Pearson r is designed to provide a measure of consistency. Loosely described, this means Pearson r helps assess whether relative rank appears to be replicated from one set of scores to another.

Pearson r does not assess magnitude of absolute differences and can therefore present a misleading assessment of reliability when scores from test-retest or parallel forms show large differences.

As Example 3 below demonstrates, Pearson r shows a value of .91 for the Relative Reliability scores but note that the actual scores are very different (Mean for Test 1 = 77.50, mean for Test 2 = 16.62).

Example 3: Relative vs. Absolute Reliability

	Relative Reliability, Consistency				Absolute Reliability, Agreement		
Student	Test 1	Rank 1	Test 2	Rank 2	Test 1	Test 2	Difference
1	95	1	44	1	95	92	3
2	90	2	22	2	90	91	-1
3	85	3	20	3	85	83	2
4	80	4	19	4	80	79	1
5	75	5	10	5	75	78	-3
6	70	6	9	6	70	72	-2
7	65	7	8	7	65	64	1
8	60	8	1	8	60	61	-1
	Test 1 and 2 Pearson r = .91				Test 1 and 2 Pearson r = .98		

Example 4 helps to solidify the problem with using Pearson r to assess test-retest and parallel forms reliability.

In Example 4, note that Time 2 scores have error, but also has a growth component of 20 points from Time 1. The two sets of observed scores, Time 1 and Time 2, are no longer equivalent, so scores are no longer stable over time.

Example 4: True Scores with Error and Systematic Difference Added

Student	True Score	Error Time 1	Error Time 2	Time 2 Change	Observed Time 1 (True + Error 1)	Observed Time 2 (True + Error 2 + Change)
1	95	3	-3	+20	98	112
2	90	-3	-3	+20	87	107
3	85	3	3	+20	88	108
4	80	-3	3	+20	77	103
5	75	-3	3	+20	72	98
6	70	3	3	+20	73	93
7	65	-3	-3	+20	62	82
8	60	3	-3	+20	63	77

Variances for true scores and observed scores:

VAR(T) = 140.00 (Variance of 16 true scores to mimic test and retest situation)
VAR(X) = 256.26 (Variance of both Time 1 and Time 2 observed scores combined)

The CTT reliability is

$$r_{xx} = \frac{VAR(T)}{VAR(X)} = \frac{140.00}{256.26} = 0.546$$

which means that 54.6% of variance in observed scores is due to true score variance.

The Pearson correlation, however, between Observed scores at Time 1 and 2, is

r = 0.935

The Pearson r of .935 suggests the scores are stable over time and therefore provides a misleading assessment of stability.

In some situations, one desires a measure of consistency. For example, when comparing student performance on the ACT and SAT, a measure of consistency would be helpful to know whether the general ranking, or relative position of students, remains similar despite the ACT and SAT having different scoring scales. If raters are asked to independent rate something, such as observed anti-social behavior, and if the raters develop and use different rating scales, Pearson r could assess whether scores obtained from the two rating scales and raters provided similar relative ratings of those observed for anti-social behavior.

When comparing parallel scales or tests, or when assessing stability of scores from a scale or test, a preferred measure is one that considers both relative performance (consistency) and absolute performance (score agreement). Pearson r does not provide a measure that addresses both conditions.

Question

Are there any types of measuring devices (scales, inventories, etc.) that may not be appropriate for test-retest reliability?

Answer

Easy recall tests in which participants could easily remember their responses from one administration to the next. Multiple choice tests are subject to recall over short periods of time.

Why is this problematic?

If participants can recall their answers, they may choose to use the same answers not because those answers represent their current thoughts or understanding, but because it is easier to answer similarly than to think through a question again. Consistent scores from one administration to the next is therefore the result of recall rather than similar thoughts or achievement.

1st administration responses based upon thoughtful effort
2nd administration responses based upon memory

This process of using memory rather than thought and consideration to answer each item can artificially inflate reliability estimates for test-retest reliability.

Scores from studies in which a pre-test and post-test are used are also inappropriate for test-retest types of reliability assessments.

Question

Why is it inappropriate to assess test-retest reliability from a pretest to posttest study?

Answer

With pre-test to post-test types of studies, test-retest reliability is not ideal because test-retest assumes scores should remain stable, consistent, over time. However, pre-test and a post-test studies usually implement a treatment between the pre-test and the post-test, and usually this treatment is designed to change scores from pre to post. If one expects, as a result of the treatment, that scores will change between pre and post, then test-retest is not appropriate since scores should change and therefore not exhibit ***stability***.

Note the term ***stability*** used above. Test-retest reliability estimates are also known as ***coefficient of stability*** since test-retest scores should be stable (unchanging) over time.

Question

Suppose an instrument measures something that is highly variable over short time periods (such as test anxiety), would test-retest be appropriate for that variable?

Answer

No, test-retest reliability is designed to determine the stability of scores over time, so something that fluctuates over time is inappropriate for test-retest reliability. In short, test-retest reliability is designed to measure the STABILITY of scores over time, and the correlation coefficient obtained from correlating two sets of scores is sometimes called the coefficient of stability. Instruments designed to measure highly variable constructs, such as test anxiety, may not be suitable for test-retest reliability since test-retest focuses upon STABLE traits, such as IQ.

One type of test suitable for test-retest reliability would be IQ tests since IQ is supposedly a consistent trait that does not vary much over time (which makes it suitable for test-retest) and also people are unlikely to recall their answers over a period of several years thus reliability is not likely to be inflated due to recall.

Note – Mean scores must also be similar in test-retest, and same is true for equivalent-forms reliability. If mean scores between sets of scores differ, this suggests the scores from the two sets do not agree.

3.3 Intraclass Correlation Coefficient, ICC

When scores do not agree as illustrated in Example 4 above, the Pearson correlation coefficient, r, provides a poor assessment of reliability. A better assessment is known as the ICC which provides an index much like the Pearson r, but can account for level of agreement in addition to level of consistency.

The formula for calculating various types of ICC are too complex for this course, but note that the ICC works well. For Example 4 data, the Pearson correlation reported a test-retest reliability of 0.935, which is very high and suggest strong stability of scores over time. However, the actual reliability was calculated to be much lower, as shown below.

$$rxx = \frac{VAR(T)}{VAR(X)} = \frac{140.00}{256.26} = 0.546$$

So, can the ICC do a better job of assessing agreement over time – stability – than the Pearson r? Yes, the ICC for absolute agreement for a single test form is:

$$ICC_{Agreement} = \frac{VAR(R)}{VAR(R)+VAR(T)+VAR(E)} = \frac{150}{150+198.714+10.286} = \frac{150}{359} = .417$$

so the ICC of .417 provides a more realistic view of reliability than does the Pearson correlation of .935 provided above.

Another benefit of ICC is that is can easily incorporate more than two time periods for retesting, e.g., period 1, period 2, period 3, period 4, etc.

Unfortunately, the ICC is not frequently used as a measure of test-retest or parallel form reliability likely due to researchers' and educators' unfamiliarity with it.

One other point about ICC, it is flexible and can be used to assess agreement or consistency. When it measures consistency for two sets of scores, i.e., time 1 and time 2, it produces the same estimate as Pearson r. In most cases one should use ICC for agreement rather than consistency. ICC would be an excellent choice for consistency if one wanted to know whether three or more separate scales or raters produced scores that ranked-ordered participants in a similar away but using different criteria, rating scales, etc. This is not an assessment of agreement, rather it answers the question of whether any two scales or rating systems will agree on the order of rank for whatever is assessed.

Published Example of ICC Use

Levy, S., Sherritt, L., Harris, S. K., Gates, E. C., Holder, D. W., Kulig, J. W., & Knight, J. R. (2004). Test-retest reliability of adolescents' self-report of substance use. Alcoholism: Clinical and Experimental Research, 28(8), 1236-1241.
Purpose of study was to translate into Italian the Norwegian teacher self-efficacy scale.

Purpose: Learn whether a substance abuse scale provides stable scores over time for adolescent participants. A one-week interval was used for time 1 and 2.

Note: See presentation video for discussion of this information.

RELIABILITY OF ADOLESCENTS' SELF-REPORT OF SUBSTANCE USE 1239

Table 2. Means With 95% Confidence Intervals at Time 1 and Time 2 and Intraclass Correlation Coefficients for Lifetime and Past-Year CRAFFT Scores Among Adolescent Medical Clinic Patients (n = 93)

	CRAFFT lifetime			CRAFFT past year		
Variable	Time 1, mean ± SD (median, range)	Time 2, mean ± SD (median, range)	ICC (95% CI)	Time 1, mean ± SD (median, range)	Time 2, mean ± SD (median, range)	ICC (95% CI)
All	0.7 ± 1.2 (0, 0–6)	0.5 ± 1.1** (0, 0–5)	0.93 (0.90–0.95)	0.4 ± 0.9 (0, 0–6)	0.4 ± 0.9 (0, 0–5)	0.91 (0.83–0.95)
Females	0.5 ± 1.0 (0, 0–5)	0.4 ± 0.8** (0, 0–4)	0.87 (0.80–0.92)	0.3 ± 0.7 (0, 0–4)	0.3 ± 0.7 (0, 0–4)	0.86 (0.79–0.91)
Males	1.0 ± 1.5 (0, 0–6)	0.9 ± 1.6 (0, 0–6)	0.97 (0.94–0.99)	0.7 ± 1.3 (0, 0–5)	0.6 ± 1.3 (0, 0–5)	0.94 (0.88–0.97)
12–15 years	0.3 ± 0.6 (0, 0–3)	0.2 ± 0.5* (0, 0–2)	0.88 (0.80–0.93)	0.2 ± 0.4 (0, 0–2)	0.2 ± 0.4 (0, 0–1)	0.76 (0.62–0.86)
16–19 years	1.0 ± 1.5 (0, 0–6)	0.9 ± 1.5* (0, 0–6)	0.94 (0.89–0.97)	0.7 ± 1.2 (0, 0–5)	0.6 ± 1.2 (0, 0–5)	0.92 (0.86–0.96)

CI, confidence interval.
Wilcoxon signed rank test; * $p = 0.05$; ** $p < 0.01$.

3.4 Parallel-forms, or Equivalent-forms, Reliability

Equivalent-forms (also called parallel-forms) reliability is designed to assess whether two forms of the same scale or test provide similar scores. The two forms of scales or tests should measure the same thing, but have differently worded items.

Examples of Similar Test Items

Example 1: Assessing Basic Multiplication Understanding

Form A

What is 5 × 7?

Form B

What is 4 × 8?

Example 2: Assessing Reliability Coefficient Understanding

Form A

Which reliability estimate is strongest?

.35

.56

.81

Form B

Which reliability estimate is weakest?

.21

.49

.75

Example 3

If you took the ACT, SAT, or GRE more than once, you would not take the same form of the test, but all forms supposedly provide similar scores for an individual. Thus, anyone who took the SAT twice is likely to get similar scores for each subtest (e.g., 50 verbal first time, 52 verbal the next time).

The procedure for establishing equivalent-forms reliability is nearly identical to the procedure for test-retest reliability.

Equivalent-forms reliability establishment steps:

- Have two (or more) forms of an instrument (scale, test, etc.)
- Administer the ***both forms*** to the ***same group*** of people at roughly the same time (e.g., within a few hours or days).
- Obtain scores from the two forms.
- Correlate the paired scores using Pearson's correlation coefficient or use the ICC.
- Show that mean scores from the two forms are similar.

The Pearson correlation between the two sets of scores will produce the equivalent-forms reliability coefficient, or use the ICC to assess both consistency and agreement.

Paired scores – each person will provide two scores, one score from the first form and a second score from the second form. Each set of scores must be paired to the same person as shown below. If scores are not paired, then reliability estimates will be incorrect.

Person	*Form 1 Score*	*Form 2 Score*
Bryan	81	83
Marijke	93	91
Gunther	88	88
Marlynn	90	92

Question

Why is it important to administer both forms to the same group of people?

Answer

True score assessment works by assessing the true score for the same individuals on both forms – true scores should not change within a person. Also, correlation requires paired scores, so scores must be paired by participants others any deviation between pairs of scores will be due to differences in participants and not due to differences in forms. Also, paired scores eliminate a large source of variability – individual differences across participants in performance.

Question

Two forms of a test are administered to the same group of people. Scores from both forms are correlated. Suppose the correlation is very high (r = .95), does this mean the two forms are equivalent?

Answer

Correlation coefficient of .95 is very high and would indicate high level of equivalence between forms. However, one must examine means from the two forms to ensure the mean scores are also similar. It is possible for scores to be highly correlated, yet means be very different – if that is the case, then the forms are not equivalent.

Example

Listed below are five people and the scores they received on form A (first score) and form B (second score). These scores show a similar pattern, yet there is little to no agreement, however the Pearson r = .98.

Participant	*Form 1*	*Form 2*
Person A	90	50
Person B	80	40
Person C	70	30
Person D	50	20
Person E	30	10

The equivalent forms reliability would be .98. Does this mean the two forms are equivalent?

No, the means on the two forms are very different, so these forms cannot be equivalent despite having a similar pattern of scores (as shown by the correlation coefficient of .98).

As you can see, a high reliability coefficient is not enough with equivalent forms -- these forms must also produce similar mean scores too. The ICC discussed above would provide an assessment of both consistency and agreement.

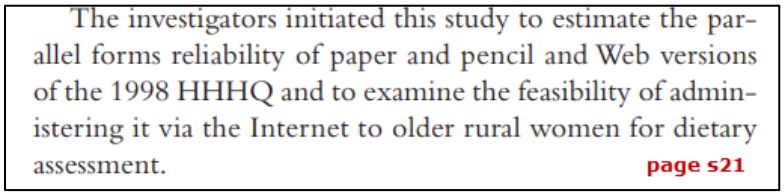
In sum, to show equivalent forms reliability, one must show that the pattern of scores from two tests are very similar (have a high correlation), and one must also show that scores from the two tests are similar, that is, they must have similar levels of difficulty (have similar mean scores) or agreement.

Published Example of Parallel-Forms Reliability

Boeckner, L. S., Pullen, C. H., Walker, S. N., Abbott, G. W., & Block, T. (2002). Use and reliability of the World Wide Web version of the Block Health Habits and History Questionnaire with older rural women. Journal of Nutrition Education and Behavior, 34, S20-S24.

Purpose: Learn whether the Block Health Habits questionnaire provided similar scores between a paper version and an internet version.

Note: See presentation video for discussion of this information.


The investigators initiated this study to estimate the parallel forms reliability of paper and pencil and Web versions of the 1998 HHHQ and to examine the feasibility of administering it via the Internet to older rural women for dietary assessment. page s21

They used Pearson r and t-tests, but they could have opted for ICC to assess agreement.

FINDINGS AND DISCUSSION page s22

Parallel Forms Reliability Table 1 provides Pearson product moment correlations between the Web version of the HHHQ and the paper and pencil version. Pearson correlation coefficients are shown for selected measures of energy, macronutrients, minerals, vitamins, and daily food group servings. Coefficients ranged between .54 and .86 for all dietary variables (median r = .80). These correlations between the paper and pencil and Web-based versions of the HHHQ for most of the energy and nutrient measures approximated acceptable reliability coefficients (≥ .70) per Nunnally[32] and were statistically significant at < .05.

Paired t test values for measures of energy, macronutrients, minerals, vitamins, and daily food group servings are shown in Table 2. There were no significant differences in means between the paper and pencil and Web versions, further supporting the parallel forms reliability of the Web-based version.

Ease of Computer Use Twenty of the 31 women (64%) who were recruited for the study had a computer at home, 26 (84%) had access to a computer, and 27 (87%) had previously used a computer. At study baseline, 20 women (64%) reported being comfortable or very comfortable with the computer, whereas 11 (36%) were not comfortable. Following their experience with completing the HHHQ on the Internet, only 3 women (10%) indicated that they would not use the computer at another time to complete a survey. This experience increased 8 women's interest (26%) in using computers, whereas interest stayed the same for 21 women

sion of the HHHQ were not statistically different from that obtained using the paper and pencil version. The validity

Table 1. Pearson Correlations for 1998 Health Habits and History Questionnaire: Paper and Pencil versus Web Version (n = 29)

	Pearson Correlation Coefficient	*P* Value
Energy and Nutrients		
Energy, kcal	.79	.0001
Protein, g	.84	.0001
Carbohydrates, g	.81	.0001
Fat, g	.76	.0001
Saturated fat, g	.81	.0001
Calcium, mg	.84	.0001
Iron, mg*	.65	.0001
Vitamin A, IU*	.86	.0001
Thiamine, mg*	.75	.0001
Riboflavin, mg	.73	.0001
Niacin, mg*	.74	.0001
Vitamin C, mg	.54	.003
Food Group Servings		
Grains	.80	.0001
Fruits	.69	.0001
Vegetables*	.68	.0001
Meats*	.84	.0001
Dairy*	.84	.0001
Fat/sweets	.76	.0001

*Square root transformation was performed.

Table 2. Paired *t* Tests for Dietary Factors from Paper and Pencil and Web Versions of the 1998 Health Habits and History Questionnaire (n = 29)

	Paper and Pencil Mean ± SD	Web Mean ± SD	t*	Sig (2-Tailed)
Energy and Nutrients				
Energy, kcal	1572 ± 635	1605 ± 633	–0.440	0.66
Protein, g	62.6 ± 25.7	63.8 ± 25.5	–0.460	0.649
Carbohydrates, g	204.6 ± 94.4	207.3 ± 96.8	–0.249	0.805
Fat, g	59.2 ± 28.7	61.5 ± 25.0	–0.631	0.533
Saturated fat, g	16.9 ± 8.6	17.8 ± 8.0	–0.908	0.372
Calcium, mg	716 ± 392	723 ± 408	–0.153	0.880
Iron, mg†	14.5 ± 8.3	14.8 ± 8.9	–0.231	0.819
Vitamin A, IU†	14894 ± 12450	13195 ± 9669	1.073	0.292
Thiamin, mg†	1.3 ± 0.6	1.4 ± 0.7	–0.522	0.606
Riboflavin, mg	1.7 ± 0.9	1.8 ± 0.9	–0.588	0.561
Niacin, mg†	19.7 ± 8.3	20.8 ± 10.5	–0.656	0.517
Vitamin C, mg	108.2 ± 58.9	106.4 ± 52.8	0.181	0.857
Food Group Servings				
Grains	4.3 ± 2.4	4.1 ± 2.1	0.564	0.577
Fruits	1.6 ± 0.9	1.8 ± 1.1	–1.226	0.230
Vegetables†	3.7 ± 2.4	3.4 ± 1.7	0.872	0.391
Meats†	1.7 ± 1.0	1.8 ± 0.9	–1.148	0.261
Dairy†	1.3 ± 1.1	1.2 ± 1.1	0.014	0.989
Fat/sweets	2.3 ± 1.5	2.4 ± 1.4	–0.965	0.343

*t statistic for paired differences between means (paper and pencil, Web).
†Square root transformation was performed.

3.5 Internal Consistency Reliability

One of the most practical and employed forms of reliability estimation is ***internal consistency*** which refers to the level of consistency in responses to items that are designed to measure the same construct.

Construct refers to a conceptualized variable that is measured using responses to several items designed to measure that construct. Thus, one constructs scores for the conceptualized variable by forming a ***composite score*** from several items.

Examples of constructs include IQ, mathematics self-efficacy, and science motivation. Each of these conceptualized variables are formed by taking a composite score from responses to many items (or indicators – items designed to measure a construct or latent variable are known as indicators).

Example 1

Test Anxiety Items appear below. **Logical Consistency Check –** To assess whether items are likely to generate internally consistency responses, do the following:

- Assume you have high levels of test anxiety
- Answer each item
- Determine whether responses are similar for each item

Instructions: Please indicate, on the scale provided, how true each statement is for you ***immediately before taking an important test***.

	Not True of Me					*Very True of Me*
1. I have an uneasy, upset feeling.	1	2	3	4	5	6
2. I'm concerned about doing poorly.	1	2	3	4	5	6
3. I'm thinking about the consequences of failing.	1	2	3	4	5	6

If each of the items generates similar responses, then these items may produce high levels of internal consistency.

The construct of test anxiety could be formed by taking the overall mean response to the three items.

So, for example, student Beth G. responded to items 1, 2, and 3 with these scores: 5, 4, 5.

Here test anxiety composite score would be (5+4+5)/3 = 4.66

Example 2

The Course Satisfaction Survey **(Logical Consistency Check)**

- Assume you are very dissatisfied with this course
- Answer each item
- Determine whether responses are similar for each item

Are there any items that appear to produce inconsistent responses?

1. Do your ever feel like skipping this class?	never 1	rarely 2	sometimes 3	often 4	always 5
2. Do you like this class?	very much 1	quite 2	fairly 3	not too much 4	not at all 5
3. Do you like the way this class is taught?	very much 1	quite 2	fairly 3	not too much 4	not at all 5
4. Are you glad you chose or were assigned to be in this class?	very glad 1	most of the time 2	sometimes 3	not too often 4	not at all 5
5. How much do you feel you have learned in this class?	a great deal 1	quite a bit 2	a fair amount 3	not much 4	nothing 5
6. Do you like your other courses?	very much 1	quite a bit 2	a fair amount 3	not much 4	not at all 5
7. Does the teacher give you help when needed?	always 1	most of the time 2	usually 3	sometime 4	never 5

(Course Satisfaction Survey items adapted from B. W. Tuckman (1988). Conducting Ed. Res. (3rd). HBJ, p. 236.)

Answer

Item 6 seems to be different and will likely produce responses that differ (be inconsistent) from the other items.

While we may be able to predict responses to the other items given responses to previous items, item 6 would be difficult to predict therefore it produces responses that lower consistency.

Internal Consistency establishment steps:

- Have items, must be more than one item, designed to measure a single construct.
- Administer items a group of participants.
- Calculate internal consistency on responses to all items designed to measure that single construct from each participant. Note: One should NOT calculate internal consistency for a total instrument if not all items are designed to measure the same construct.
- There are several ways to calculate internal consistency, Cronbach's alpha most common.

Cronbach's alpha is a measure of internal consistency, it is not someone's car.

Cronbach's Alpha (α)

- an advanced form of the split-half reliability method (briefly explained video presentation)
- ranges from 0.00 to 1.00
- closer to 1.00 the more consistent are responses from items
- closer to 0.00 the less consistent are responses from items

Question

What is the cut-off level one usually expects for scores to be judged internally consistent?

Answer

For most research purposes Cronbach's alpha should be .70 or larger, and for professionally developed instruments (such as SAT, GRE, CRCT), one expects alpha to be .90 or better. This is true for any measures of reliability considered – test-retest, ICC, Cronbach's alpha.

Published Example of Cronbach's Alpha

Avanzi, L., Miglioretti, M., Velasco, V., Balducci, C., Vecchio, L., Fraccaroli, F., & Skaalvik, E. M. (2013). Cross-validation of the Norwegian teacher's self-efficacy scale (NTSES). Teaching and Teacher Education, 31, 69-78.

Purpose: Translate into Italian the Norwegian teacher self-efficacy scale.

Note: See presentation video for discussion of this information.

5.3. Reliability of the Italian version of NTSES page 75

We calculated Cronbach's alphas of the Italian NTSES subscales at both Time 1 and Time 2 for the sub-sample ($N = 140$), and they were very good, ranging from .79 to .92. Furthermore, the test-retest reliability for the six subscales ranged from $r = .48$ (for 'cooperate with colleagues and parents') to $r = .64$ (for 'maintain discipline'), while the same index for the overall teacher's self-efficacy measure was $r = .65$. Moreover, we calculated the CR for each dimension of the Italian sample (at Time 1). They were very good, and similar to those obtained using the alphas, indicating that each NTSES subscale was reliable. The values of AVE for all dimensions were greater than .50, ranging from .52 to .73, evidencing that most of the variance in the items was explained by the underlying factor, rather than by the measurement error (for details regarding each dimensions see Table 3).

Table 3 page 75

Alphas *T1* and *T2*, CR and AVE for the six dimensions of NTSE in Italian sample.

NTSE dimensions	Cronbach alpha *T1* ($N = 348$)	Cronbach alpha *T2* ($N = 140$)	CR ($N = 348$)	AVE ($N = 348$)
1. NTSE instruction	.88	.83	.87	.63
2. NTSE adapt instruction to individual needs	.88	.85	.88	.64
3. NTSE cooperate with colleagues and parents	.81	.79	.81	.52
4. NTSE cope with change	.83	.81	.83	.56
5. NTSE motivate students	.87	.84	.88	.64
6. NTSE maintain discipline	.92	.91	.92	.73

Note. CR = Composite Reliability; AVE = Average Variance Extracted.

Often researchers report Cronbach's alpha within text, usually in the Instrumentation section when describing instruments. One can also find alpha reported in tables. In this research the authors reported Cronbach's alpha in both.

Intra-class Correlation Coefficient (ICC)

The ICC can also be used as a measure of internal consistency. Recall earlier discussion of ICC – it can be used to assess agreement or consistency. If calculated for consistency, it provides the same value of Cronbach's alpha, thus ICC and alpha show the same results. If ICC is calculated for agreement, the ICC will likely differ from Cronbach's alpha. If the items examined for consistency use the same scale (e.g., 1 = not like me, 5 = very much like me; or 1 = strongly disagree and 5 = strongly agree), then using ICC in agreement mode can be helpful for detecting whether items show both consistency and agreement (i.e., provide similar scores). If the items do not follow the same scale, one should focus on consistency and not on agreement. Despite the flexibility of ICC, Cronbach's is the most common measure of internal consistency.

3.6 Scorer/Rater Reliability (or, more accurately termed Agreement)

Often judges/raters must evaluate something using standards or scoring rubrics (e.g., essays evaluated by teachers, Olympic divers rated by judges, boxing matches rated by ring judges, doctors rating severity of infection, researchers coding transcriptions of interviews, etc.). In such situations it can be critical to determine the level of agreement among judges/raters. If that agreement level is low, it suggests judges/raters are not employing scoring rubrics or standards, or not employing them to the same degree, or assigning scores based upon personal opinions/preferences/biases.

Intra-judge, one judge vs inter-judge, multiple judges.

Question

What is the difference between intra-judge and inter-judge reliability?

Answer

Intra-judge reliability refers to the consistency with which one judge assigns scores, ***inter-judge (inter-rater) reliability*** refers to the consistency with which 2 or more judges assign scores to the same event.

Question

For intra-judge, how is consistency of scores established?

Answer

One judge ***must score something more than once*** to establish intra-judge reliability. To establish intra-judge, ***two or more scores*** of the same assessment are needed to show consistency. If only one score is obtained, it is impossible to assess intra-judge reliability.

Question

A teacher reads an essay and scores it based upon predefined scoring criteria as defined within a scoring rubric. For example, using an essay grading rubric, the teacher reads Bryan's essay and scores it an 8 out of 10.

Does this establish intra-judge reliability? Explain why or why not.

Answer

To show reliability, one must show that scores are similar, so ***more than one score*** on that essay is needed to assess intra-judge reliability. The teacher must read Bryan's essay once and score it, then let time pass and read Bryan's essay again and score it again. If the two essay scores were similar, say an 8 and 9 out of 10 for the two readings, that would show consistency. If however, the scores are dissimilar, such as 8 and 4, then that demonstrates no intra-judge reliability and this suggests problems with scoring of essays.

Question

How is inter-judge (inter-rater) consistency determined?

Answer

One compare scores from 2 or more judges to learn if the scores are similar -- if the scores are similar, then there is evidence for inter-judge reliability. Key is that scores must be on the same observations (e.g., same essay, diving event, etc.).

Example

Olympic judges for high dive competition rates one dive from one diver and we examine whether those ratings are similar.

Question

What statistics are used to measure intra/inter-judge reliability?

Answer

For quantitative rating scales, both Pearson r and ICC can be used exactly like described above for test-retest and parallel forms reliability. If judges use a rating format that categorical/qualitative (e.g., Would you classify this behavior as aggressive, indifferent, defensive, excited, etc.), then Pearson r and ICC won't work. There are measures of agreement available for qualitative data such as Cohen's kappa, Scott's pi, and Krippendorff's alpha, but we won't cover those in this course. They function much like ICC and Pearson r, but interpretations can be different such that 0.70 may be too high a cut level for judging agreement. See the example below, especially the footnote in Table 2 on the next page.

Published Example of Intra and Inter-rater Reliability

McCullough, G. H., Wertz, R. T., Rosenbek, J. C., Mills, R. H., Webb, W. G., & Ross, K. B. (2001). Inter-and intrajudge reliability for videofluoroscopic swallowing evaluation measures. Dysphagia, 16(2), 110-118.

Purpose: "Interjudge reliability for videofiuoroscopic (VFS) swallowing evaluations has been investigated, and results have, for the most part, indicated that reliability is poor." Purpose was to add intrajudge and test frame by frame assessments.

Note: See presentation video for discussion of this information.

Interjudge reliability was determined by comparing the responses made on the data sheets by all three participants (clinical judges). At least one week after the original viewing, the primary study clinician reanalyzed each of the VFS examinations and recorded all measurements on a new data sheet. Intrajudge reliability was determined by comparing his original ratings with his ratings from the second viewing.

All data were entered on spreadsheets and analyzed using SPSS 10.0 for Windows. The following analyses determined intrajudge reliability: for all binary ratings, Cohen's kappa; for all duration measures, Pearson's product moment correlations; and for the 8-point penetration-aspiration scale, Kendall's tau correlations. To determine interjudge reliability, the following analyses were performed: for all binary ratings, group kappas; and for all duration measures and the 8-point penetration-aspiration scale, intraclass correlation coefficients. The intraclass correlation coefficient (ICC) is based on a two-way random effects analysis of variance (ANOVA) model, as defined by Shrout and Fleiss [15]. A two-way random effects analysis for single-measure ICCs was used to determine the applicability of the results to other clinical judges who would be independently rating the measures.

Table 2. Reliability for videofiuoroscopic measures of penetration-aspiration rated on two scales. Intrajudge reliability results from comparisons of judge 1 ratings made at two different times. Interjudge reliability results from comparisons among all three judges

	Present-absent ratings		8-point scale ratings[b]	
Consistency	Intra[a]	Inter	Intra	Inter
1. Thin liquid (5 cc)	K = 0.843	K = 0.400	r = 0.467	ICC = 0.114
2. Thin liquid (5 cc)	K = 0.757	K = 0.274	'C = 0.750	ICC = 0.380
3. Thin liquid (10 cc)	K = 0.693	K = 0.415	r = 0.473	ICC = 0.591
4. Thin liquid (10 cc)	K = 0.530	K = -0.138	r = 0.355	ICC = 0.085
5. Thick liquid (5 cc)	K = 0.755	K = 0.067	r = 0.332	ICC = 0.647
6. Thick liquid (5 cc)	K = 0.724	K = 0.190	r = 0.284	ICC = 0.628
7. Thick liquid (10 cc)	K = 0.577	K = 0.352	r = 0.172	ICC = 0.623
8. Thick liquid (10 cc)	K = 0.886	K = -0.200	100%	ICC = 0.008
9. Puree (5 cc)	K = 90%	K = 0.194	83%	ICC = 0.224
10. Puree (5 cc)	K = 90%	K = 0.194	83%	ICC = 0.224
11. Solid (1/4 cookie)	K = 0.893	K = -0.081	100%	ICC = 0.322
12. Solid (1/4 cookie)	K = 1.000	K = -0.269	100%	ICC = 0.166

[a]Intrajudge reliability for present-absent was analyzed with Cohen's kappa, and interjudge reliability was analyzed with group kappas. Scale for kappa values: below 0.00 = poor agreement; 0.00-0.20 = slight agreement; 0.21-0.40 = fair agreement; 0.41-0.60 = moderate agreement; 0.61-0.80 = substantial agreement; 0.81-1.00 = almost perfect agreement. Bold type indicates that reliability for the measure was moderate" to almost perfect" in agreement or that percent agreement is 90% or better. Percent agreement is reported when kappa could not be computed.

[b]Intrajudge reliability for the 8-point scale was analyzed using Kendall's tau, and interjudge reliability was analyzed using an intraclass correlation coefficient. Bold type indicates that the value is significant at $p < 0.01$ or percent agreement is 90% or better. Percent agreement is reported when tau or ICC could not be computed.

4. Summary Reliability Chart

A former student developed this chart and you may find it useful.

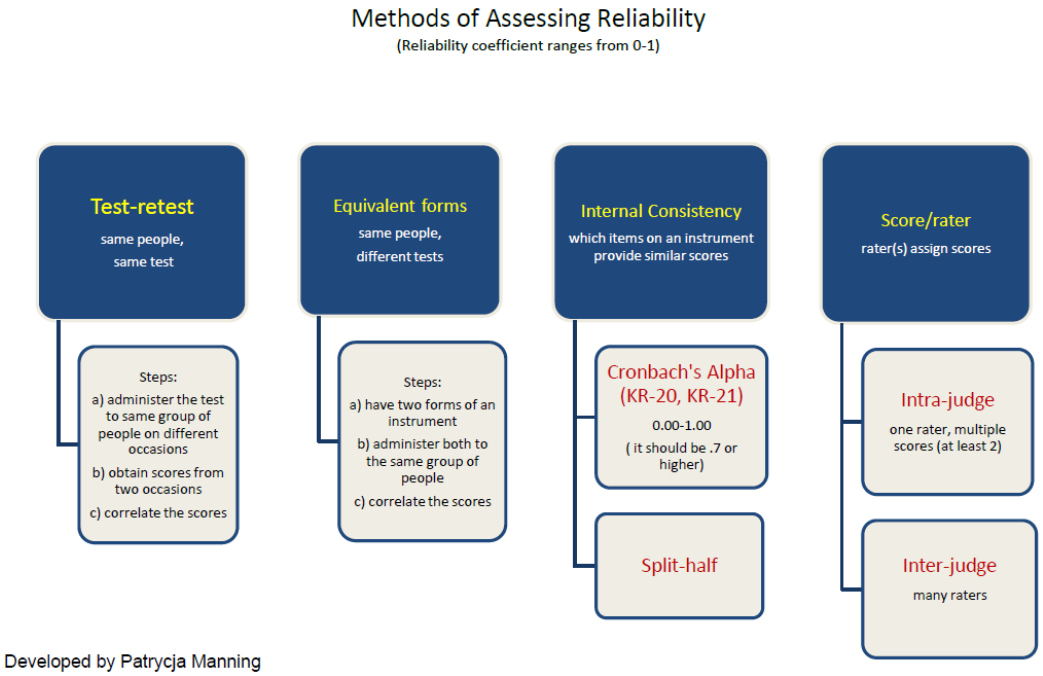


Figure update: Note that the Intraclass Correlation Coefficient (ICC) is better measure of Test-retest and Equivalent Forms than correlation using Pearson r, and ICC can also be used for Intra and Inter judge reliability if ratings are quantitative rather than qualitative scores.