Reliability: Empricial Estimates

PSYC3302: Psychological Measurement and Its Applications

Mark Hurlstone
Univeristy of Western Australia

Weeks 3 & 4

Learning Objectives

- Overview of three methods for estimating reliability from *real* data:
  1. Alternate-Forms Reliability
  2. Test-Retest Reliability
  3. Internal Consistency Reliability
     - 3.1 Split-Halves Reliability
     - 3.2 Cronbach's α
     - 3.3 Standardised Cronbach's α
     - 3.4 KR-20

Empirical Reliability Estimates

- So far, we have focused on the theoretical basis of reliability in terms of CTT
- We will now focus on how observed (empirical) test scores can be used to *estimate* score reliabilities
- We will consider several different methods for generating empirical estimates of reliability
- Each is grounded in the notion of parallel tests—providing a direct link to CTT
- The methods differ in terms of their assumptions and the types of data they lend themselves to

Three Methods For Generating Empirical Reliability Estimates

1. Alternate-Forms Reliability
2. Test-Retest Reliability
3. Internal Consistency Reliability

Three Methods For Generating Empirical Reliability Estimates

1. Alternate-Forms Reliability
2. Test-Retest Reliability
3. Internal Consistency Reliability

Alternate-Forms Reliability

- This involves obtaining scores from two different forms of a test with the same group of people
- An example of the use of this type of reliability would be a "makeup" test
- The correlation between test scores on the two forms is an index of reliability known as the *coefficient of equivalence*

Alternate-Forms Reliability

- To interpret a correlation between alternate forms as an estimate of reliability the two test forms must be parallel—known as *parallel forms*
- Recall from our discussion of parallel tests that this requires that both forms:
  1. measure the same set of true scores
  2. have the same amount of error variance
- Thus, parallel forms of a test exist when, for each form, the observed scored means and variances are the same

Alternate-Forms Reliability: Different Content Problem

- Two forms of a test may ostensibly meet the requirements of CTT, but not measure the same psychological attribute
- This is because different forms will necessarily possess different content
- For example, two versions of a self-esteem questionnaire may tap different components of this construct:
  - socially derived self-esteem vs. nonsocial self-esteem
- Thus, respondents' true scores on one form are not strictly equal to their true scores on the second form—the tests are not "truly" parallel

Alternate-Forms Reliability: Carryover Effects

- According to CTT error scores on one form of a test should be uncorrelated with error scores on a second form of a test
- However, if two forms of a test are completed in close succession there may be *carryover effects*
- For example, a respondent's memory for test content, attitudes, or mood state might similarly affect performance on both forms of a test
- This could cause the error scores on the two forms to be correlated with one another—violating the parallel test assumption

Alternate-Forms Reliability: Carryover Effects

- Let's consider another "all-knowing" example to illustrate the problem of carryover effects
- For sake of demonstration, we must once again pretend that we know people's true scores and error scores

Alternate-Forms Reliability: Carryover Effects

Table: Example of Carryover Effects on Alternate Forms Estimate of Reliability

	Form 1					Form 2				
Respondent	Observed Score (X_{o1})		True Score (X_{t1})		Error Score (X_{e1})	Observed Score (X_{o2})		True Score (X_{t2})		Error Score (X_{e2})
1	14	=	15	+	–1	13	=	15	+	–2
2	17	=	14	+	3	17	=	14	+	3
3	11	=	13	+	–2	12	=	13	+	–1
4	10	=	12	+	–2	11	=	12	+	–1
5	14	=	11	+	3	14	=	11	+	3
6	9	=	10	+	–1	8	=	10	+	–2
Mean	12.5	=	12.5		0	12.5	=	12.5		0
Variance	7.58	=	2.92		4.67	7.58	=	2.92		4.67

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

Table: Example of Carryover Effects on Alternate Forms Estimate of Reliability

	Form 1					Form 2				
Respondent	Observed Score (X_{o1})		True Score (X_{t1})		Error Score (X_{e1})	Observed Score (X_{o2})		True Score (X_{t2})		Error Score (X_{e2})
1	14	=	15	+	–1	13	=	15	+	–2
2	17	=	14	+	3	17	=	14	+	3
3	11	=	13	+	–2	12	=	13	+	–1
4	10	=	12	+	–2	11	=	12	+	–1
5	14	=	11	+	3	14	=	11	+	3
6	9	=	10	+	–1	8	=	10	+	–2
Mean	12.5	=	12.5		0	12.5	=	12.5		0
Variance	7.58	=	2.92		4.67	7.58	=	2.92		4.67

Alternate-Forms Reliability: Carryover Effects

Table: Example of Carryover Effects on Alternate Forms Estimate of Reliability

	Form 1					Form 2				
Respondent	Observed Score (X_{o1})		True Score (X_{t1})		Error Score (X_{e1})	Observed Score (X_{o2})		True Score (X_{t2})		Error Score (X_{e2})
1	14	=	15	+	–1	13	=	15	+	–2
2	17	=	14	+	3	17	=	14	+	3
3	11	=	13	+	–2	12	=	13	+	–1
4	10	=	12	+	–2	11	=	12	+	–1
5	14	=	11	+	3	14	=	11	+	3
6	9	=	10	+	–1	8	=	10	+	–2
Mean	12.5	=	12.5		0	12.5	=	12.5		0
Variance	7.58	=	2.92		4.67	7.58	=	2.92		4.67

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

Table: Example of Carryover Effects on Alternate Forms Estimate of Reliability

	Form 1					Form 2				
Respondent	Observed Score (X_{o1})		True Score (X_{t1})		Error Score (X_{e1})	Observed Score (X_{o2})		True Score (X_{t2})		Error Score (X_{e2})
1	14	=	15	+	–1	13	=	15	+	–2
2	17	=	14	+	3	17	=	14	+	3
3	11	=	13	+	–2	12	=	13	+	–1
4	10	=	12	+	–2	11	=	12	+	–1
5	14	=	11	+	3	14	=	11	+	3
6	9	=	10	+	–1	8	=	10	+	–2
Mean	12.5	=	12.5		0	12.5	=	12.5		0
Variance	7.58	=	2.92		4.67	7.58	=	2.92		4.67

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

Table: Example of Carryover Effects on Alternate Forms Estimate of Reliability

	Form 1					Form 2				
Respondent	Observed Score (X_{o1})		True Score (X_{t1})		Error Score (X_{e1})	Observed Score (X_{o2})		True Score (X_{t2})		Error Score (X_{e2})
1	14	=	15	+	–1	13	=	15	+	–2
2	17	=	14	+	3	17	=	14	+	3
3	11	=	13	+	–2	12	=	13	+	–1
4	10	=	12	+	–2	11	=	12	+	–1
5	14	=	11	+	3	14	=	11	+	3
6	9	=	10	+	–1	8	=	10	+	–2
Mean	12.5	=	12.5		0	12.5	=	12.5		0
Variance	7.58	=	2.92		4.67	7.58	=	2.92		4.67

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

- These hypothetical data meet various assumptions of CTT and parallel tests:
  - the observed scores on each form are the sum of the true scores and error scores
  - the true scores are the same for the two forms
  - the error scores for each form sum to 0 and have the same variance
  - true scores are uncorrelated with error scores
- Accordingly, the means and variances of observed scores are identical for the two forms

Alternate-Forms Reliability: Carryover Effects

Table: Example of Carryover Effects on Alternate Forms Estimate of Reliability

	Form 1					Form 2				
Respondent	Observed Score (X_{o1})		True Score (X_{t1})		Error Score (X_{e1})	Observed Score (X_{o2})		True Score (X_{t2})		Error Score (X_{e2})
1	14	=	15	+	–1	13	=	15	+	–2
2	17	=	14	+	3	17	=	14	+	3
3	11	=	13	+	–2	12	=	13	+	–1
4	10	=	12	+	–2	11	=	12	+	–1
5	14	=	11	+	3	14	=	11	+	3
6	9	=	10	+	–1	8	=	10	+	–2
Mean	12.5	=	12.5		0	12.5	=	12.5		0
Variance	7.58	=	2.92		4.67	7.58	=	2.92		4.67

Alternate-Forms Reliability: Carryover Effects

- From our "all-knowing" vantage point, we can calculate the reliability of the two forms
- We can do this using the ratio of true score variance to observed score variance:
- Reliability for Form 1:

$$R_{xx} = \frac{2.92}{7.58} = .38$$

- Reliability for Form 2:

$$R_{xx} = \frac{2.92}{7.58} = .38$$

Alternate-Forms Reliability: Carryover Effects

- From our "all-knowing" vantage point, we can calculate the reliability of the two forms
- We can do this using the ratio of true score variance to observed score variance:
- Reliability for Form 1:

$$R_{xx} = \frac{2.92}{7.58} = .38$$

- Reliability for Form 2:

$$R_{xx} = \frac{2.92}{7.58} = .38$$

Psychological Measurement
mark.hurlstone@uwa.edu.au
Empirical Estimates
1. Alternate-Forms
2. Test-Retest
3. Internal Consistency
3.1 Split-Half Reliability
3.2 Coefficient α
3.3 Standardised Coefficient α
3.4 KR-20
Factors Affecting Reliability
References

Table: Example of Carryover Effects on Alternate Forms Estimate of Reliability

	Form 1					Form 2				
Respondent	Observed Score (X_{o1})		True Score (X_{t1})		Error Score (X_{e1})	Observed Score (X_{o2})		True Score (X_{t2})		Error Score (X_{e2})
1	14	=	15	+	–1	13	=	15	+	–2
2	17	=	14	+	3	17	=	14	+	3
3	11	=	13	+	–2	12	=	13	+	–1
4	10	=	12	+	–2	11	=	12	+	–1
5	14	=	11	+	3	14	=	11	+	3
6	9	=	10	+	–1	8	=	10	+	–2
Mean	12.5	=	12.5		0	12.5	=	12.5		0
Variance	7.58	=	2.92		4.67	7.58	=	2.92		4.67

Alternate-Forms Reliability: Carryover Effects

- From our "all-knowing" vantage point, we can calculate the reliability of the two forms
- We can do this using the ratio of true score variance to observed score variance:
- Reliability for Form 1:

$$R_{xx} = \frac{2.92}{7.58} = .38$$

- Reliability for Form 2:

$$R_{xx} = \frac{2.92}{7.58} = .38$$

Alternate-Forms Reliability: Carryover Effects

- Unfortunately, the data violate an important assumption of CTT
- Error is assumed to occur at random—the error scores on one form should be uncorrelated with error scores on the second form
- The error scores on the two forms are, in fact, very strongly positively correlated: r_{e1e2} = .93
- This correlation could be the result of carryover effects, such as mood state or memory

Alternate-Forms Reliability: Carryover Effects

- The correlation between observed scores on two different forms of a test is a measure of reliability known as *alternate-forms reliability*
- The alternate-forms correlation for the two forms is $r_{o_1 o_2} = .96$
- The reliability is therefore considerably greater than its true value ($R_{xx} = .38$), meaning it is an inaccurate estimate
- The inflated estimate of reliability is brought about due to the strong correlation between error scores on the two forms of the test

Alternate-Forms Reliability: Bottom Line

- It is not enough that two forms of a test have the same observed score means and variances
- We also need to be very confident that the tests are in fact measuring the same psychological attribute
- If both of these conditions are satisfied then we can reasonably use the correlation between two forms as an estimate of reliability
- However, we must also be mindful of carryover effects from one form of a test to another

Three Methods For Generating Empirical Reliability Estimates

1. Alternate-Forms Reliability
2. Test-Retest Reliability
3. Internal Consistency Reliability

Three Methods For Generating Empirical Reliability Estimates

1. Alternate-Forms Reliability
2. Test-Retest Reliability
3. Internal Consistency Reliability

Test-Retest Reliability Estimates

- This involves administering the same test to the same people on two different occasions
- An estimate of reliability is obtained by correlating respondents test-retest scores
- This method overcomes the "different-content" problem associated with the alternative forms method
- It is appropriate when measuring the reliability of a test that purports to measure a relatively stable psychological characteristic—e.g., intelligence, personality traits

Test-Retest Reliability Estimates

- The test-retest method depends on the same assumptions as the parallel forms method:
  1. people's true scores should not change between the two testing occasions
  2. the error variances of the two tests should be identical
- The observed test-retest scores should therefore have the same means and variances

Test-Retest Reliability: Equality of Error Variances

- The "equality of error variances" assumption is not unreasonable if care is taken in the test administration process
- Efforts must be undertaken to control for extraneous variables that might differ from test to retest
- For example, we would want to control:
  - the temperature and noise of the test environment
  - the time of day the testing took place
  - the experimenter administering the test

Test-Retest Reliability: True Score Stability

- The "true score stability" assumption is a harder constraint to meet
- The respondent's levels of a psychological attribute may change between test and retest
- We can identify at least three different threats to this assumption:
  1. construct instability
  2. length of test-retest interval
  3. developmental changes

Test-Retest Reliability: True Score Stability

- The "true score stability" assumption is a harder constraint to meet
- The respondent's levels of a psychological attribute may change between test and retest
- We can identify at least three different threats to this assumption:
  1. construct instability
  2. length of test-retest interval
  3. developmental changes

Test-Retest Reliability: Construct Instability

- Some psychological attributes, like intelligence and personality, are assumed to be relatively stable–known as *psychological traits*
- Other psychological attributes, like state anxiety (anxiety felt at the moment) or mood, are assumed to fluctuate over time—known as *psychological states*
- Test-retest reliability is not appropriate when evaluating the reliability of a test that is assumed to measure psychological states
- In these circumstances, respondents' true scores are likely to change between test and retest

Test-Retest Reliability: True Score Stability

- The "true score stability" assumption is a harder constraint to meet
- The respondent's levels of a psychological attribute may change between test and retest
- We can identify at least three different threats to this assumption:
  1. construct instability
  2. length of test-retest interval
  3. developmental changes

Test-Retest Reliability: True Score Stability

- The "true score stability" assumption is a harder constraint to meet
- The respondent's levels of a psychological attribute may change between test and retest
- We can identify at least three different threats to this assumption:
  1. construct instability
  2. length of test-retest interval
  3. developmental changes

Test-Retest Reliability: Length of Test-Retest Interval

- With the passage of time people learn new things, forget some things, and acquire new skills
- The longer the test-retest interval, the more likely that changes in the psychological attribute being measured will occur
- True scores are therefore more likely to change with long (years) compared to short (weeks or days) test-retest intervals
- However, very short test-retest intervals (hours) can yield carryover and contamination effects (see earlier)
- Most test-retest analyses occur over a period of 2-8 weeks

Test-Retest Reliability: True Score Stability

- The "true score stability" assumption is a harder constraint to meet
- The respondent's levels of a psychological attribute may change between test and retest
- We can identify at least three different threats to this assumption:
  1. construct instability
  2. length of test-retest interval
  3. developmental changes

Test-Retest Reliability: True Score Stability

- The "true score stability" assumption is a harder constraint to meet
- The respondent's levels of a psychological attribute may change between test and retest
- We can identify at least three different threats to this assumption:
  1. construct instability
  2. length of test-retest interval
  3. developmental changes

Test-Retest Reliability: Developmental Changes

- The stability assumption can also be compromised if the testing occurs during a period of great developmental change
- This is a particular problem when testing the cognitive skills (e.g., maths, reading) and knowledge of children
- These can develop rapidly, resulting in changes in children's true scores even over relatively brief test-retest intervals
- Such developmental changes prevent the use of a test-retest correlation to measure reliability

Test-Retest Reliability: Bottom Line

- Test-retest reliability depends on the assumption that true scores remain stable across the test-retest interval
- For this reason the test-retest correlation is sometimes known as the *coefficient of stability*
- If the true scores remain stable during the test-retest interval, then the reliability coefficient reflects one thing—the degree to which measurement error affected test scores
- However, the problem is that we we can never be sure if this assumption is satisfied

Test-Retest Reliability: Bottom Line

- If the true scores change during the test-retest interval, then the reliability coefficient will reflect two factors:
  1. the degree of measurement error
  2. the amount of change in true scores
- In this case, an imperfect test-retest correlation indicates the combined effect of measurement error *and* true score instability
- The possibility that true scores might have changed in the test-retest interval renders it difficult to interpret a non-perfect test-retest reliability coefficient

Interim Summary: Alternate-Forms and Test-Retest Reliability

- There are several practical problems associated with both alternate-forms and test-retest reliability
- They require at least two tests to be administered which can be expensive, time consuming, and difficult
- Several assumptions must be made if the correlation between tests is to be interpreted as a measure of reliability
- These assumptions may not be valid in some, or perhaps many cases
- Accordingly, the alternate-forms and test-retest methods are of limited utility

Internal Consistency Reliability

- An estimate of the reliability of a test can be obtained without developing more than one form of a test or testing respondents on more than one occasion
- This type of reliability estimate involves evaluating the internal consistency of test items
- This third approach to reliability is thus known as *internal consistency reliability*
- It is used when items on a scale are summed to produce a composite test score

Internal Consistency Reliability

- There are two factors that determine the internal consistency reliability of test scores:

1. The consistency among parts of a test:
   - if the test items are strongly correlated with each other, the test is likely to be reliable
2. The test's length:
   - all things being equal, a longer test will be more reliable than a shorter test

Internal Consistency Reliability

- We will consider four methods of estimating internal consistency:

  1. Split-Half Reliability
  2. Coefficient α
  3. Standardised Coefficient α
  4. KR-20

Internal Consistency Reliability

- We will consider four methods of estimating internal consistency:

  1. Split-Half Reliability
  2. Coefficient α
  3. Standardised Coefficient α
  4. KR-20

Split-Half Reliability

- This method was developed by Charles Spearman in the 1920s
- It's a measure of reliability obtained by correlating two pairs of scores obtained from equivalent halves of a single test
- There are three steps to computing the split-half reliability:
  1. Divide the test into equal halves
  2. Calculate the correlation between scores on the two halves of the test
  3. Adjust the half-test reliability using the Spearman-Brown formula

Split-Half Reliability: Step 1

- In Spearman's original procedure, odd items on a test are assigned to one sub-test and even items are assigned to the other sub-test
- This is known as *odd-even reliability*
- Here's an example ...

Split-Half Reliability: Step 1

Table: Example of Internal Consistency Method of Estimating Reliability

	Items					Split-Half 1		Split-Half 2	
Person	1	2	3	4	Total	"Odd"	"Even"	1 and 4	2 and 4
1	4	4	5	4	17	9	8	8	9
2	5	2	4	2	13	9	4	7	6
3	5	4	2	2	13	7	6	7	6
4	2	3	1	2	8	3	5	4	4
Mean	4	3.25	3	2.5	12.75	7	5.75	6.5	6.25
Variance	1.5	0.6875	2.5	.75	10.1875	6	2.1875	2.25	3.1875

Split-Half Reliability: Step 1

Table: Example of Internal Consistency Method of Estimating Reliability

	Items					Split-Half 1		Split-Half 2	
Person	1	2	3	4	Total	"Odd"	"Even"	1 and 4	2 and 4
1	4	4	5	4	17	9	8	8	9
2	5	2	4	2	13	9	4	7	6
3	5	4	2	2	13	7	6	7	6
4	2	3	1	2	8	3	5	4	4
Mean	4	3.25	3	2.5	12.75	7	5.75	6.5	6.25
Variance	1.5	0.6875	2.5	.75	10.1875	6	2.1875	2.25	3.1875

Split-Half Reliability: Step 1

Table: Example of Internal Consistency Method of Estimating Reliability

	Items					Split-Half 1		Split-Half 2	
Person	1	2	3	4	Total	"Odd"	"Even"	1 and 4	2 and 4
1	4	4	5	4	17	9	8	8	9
2	5	2	4	2	13	9	4	7	6
3	5	4	2	2	13	7	6	7	6
4	2	3	1	2	8	3	5	4	4
Mean	4	3.25	3	2.5	12.75	7	5.75	6.5	6.25
Variance	1.5	0.6875	2.5	.75	10.1875	6	2.1875	2.25	3.1875

Split-Half Reliability: Step 1

Table: Example of Internal Consistency Method of Estimating Reliability

	Items					Split-Half 1		Split-Half 2	
Person	1	2	3	4	Total	"Odd"	"Even"	1 and 4	2 and 4
1	4	4	5	4	17	9	8	8	9
2	5	2	4	2	13	9	4	7	6
3	5	4	2	2	13	7	6	7	6
4	2	3	1	2	8	3	5	4	4
Mean	4	3.25	3	2.5	12.75	7	5.75	6.5	6.25
Variance	1.5	0.6875	2.5	.75	10.1875	6	2.1875	2.25	3.1875

Split-Half Reliability: Step 2

- In the second step, we calculate the split-half correlation between scores on the two halves of the test
- The odd-even split-half correlation for these data is r_{hh} = .276
- However, we can't use this as an estimate of reliability
- This is because it is an estimate of the reliability of a test that has been halved in length
- We want to know the reliability of the full test

Split-Half Reliability: Step 2

- In the second step, we calculate the split-half correlation between scores on the two halves of the test
- The odd-even split-half correlation for these data is $r_{hh} = .276$
- However, we can't use this as an estimate of reliability
- This is because it is an estimate of the reliability of a test that has been halved in length
- We want to know the reliability of the full test

Split-Half Reliability: Step 1

Table: Example of Internal Consistency Method of Estimating Reliability

	Items					Split-Half 1		Split-Half 2	
Person	1	2	3	4	Total	"Odd"	"Even"	1 and 4	2 and 4
1	4	4	5	4	17	9	8	8	9
2	5	2	4	2	13	9	4	7	6
3	5	4	2	2	13	7	6	7	6
4	2	3	1	2	8	3	5	4	4
Mean	4	3.25	3	2.5	12.75	7	5.75	6.5	6.25
Variance	1.5	0.6875	2.5	.75	10.1875	6	2.1875	2.25	3.1875

Split-Half Reliability: Step 2

- In the second step, we calculate the split-half correlation between scores on the two halves of the test
- The "odd-even" split-half correlation is $r_{hh} = .276$
- However, we can’t use this as an estimate of reliability
- This is because it is an estimate of the reliability of a test that has been halved in length
- As noted earlier, the reliability of a test is affected by its length, so the split-half correlation will underestimate the reliability of the complete test

Split-Half Reliability: Step 2

- In the second step, we calculate the split-half correlation between scores on the two halves of the test
- The "odd-even" split-half correlation is r_{hh} = .276
- However, we can't use this as an estimate of reliability
- This is because it is an estimate of the reliability of a test that has been halved in length
- As noted earlier, the reliability of a test is affected by its length, so the split-half correlation will underestimate the reliability of the complete test

Split-Half Reliability: Step 3

- To determine the reliability of the full test, we can use the Spearman-Brown Prophecy formula:

$$R_{xx} = \frac{2\, r_{hh}}{1 + r_{hh}}. \tag{17}$$

- For our "odd-even" split-half example:

$$R_{xx} = \frac{2(.276)}{1 + .276} = \frac{.552}{1.276} = .433\,.$$

Split-Half Reliability: Assumptions

- Like the alternate-forms and test-retest reliability methods, the legitimacy of the split-half approach rests on the assumption that the two halves are parallel tests
- The two halves should therefore have equal means and variances
- However, in our example, the two halves do not meet the criteria for being parallel
- This means our split-half estimate of reliability may be inaccurate

Split-Half Reliability: Assumptions

- Like the alternate-forms and test-retest reliability methods, the legitimacy of the split-half approach rests on the assumption that the two halves are parallel tests
- The two halves should therefore have equal means and variances
- However, in our example, the two halves do not meet the criteria for being parallel
- This means our split-half estimate of reliability may be inaccurate

Split-Half Reliability: Step 1

Table: Example of Internal Consistency Method of Estimating Reliability

	Items					Split-Half 1		Split-Half 2	
Person	1	2	3	4	Total	"Odd"	"Even"	1 and 4	2 and 4
1	4	4	5	4	17	9	8	8	9
2	5	2	4	2	13	9	4	7	6
3	5	4	2	2	13	7	6	7	6
4	2	3	1	2	8	3	5	4	4
Mean	4	3.25	3	2.5	12.75	7	5.75	6.5	6.25
Variance	1.5	0.6875	2.5	.75	10.1875	6	2.1875	2.25	3.1875

Split-Half Reliability: Assumptions

- Like the alternate-forms and test-retest reliability methods, the legitimacy of the split-half approach rests on the assumption that the two halves are parallel tests
- The two halves should therefore have equal means and variances
- However, the two halves do not meet the criteria for being parallel
- This means our split-half estimate of reliability may be inaccurate

Split-Half Reliability: Problem of Multiple Splits

- A serious problem with the split-half method is that there are multiple ways of randomly splitting a test into two halves
- The results can therefore be a product of the way the data were split
- For example, suppose we split the data so items 1 and 4 appeared in one half of a test, and items 2 and 3 appeared in the other half
- This yields a split-half correlation of r_{hh} = .89
- With the Spearman-Brown adjustment, R_{xx} = .94:

$$R_{xx} = \frac{2(.89)}{1 + .89} = .94\,.$$

Internal Consistency Reliability

- We will now consider methods for estimating internal consistency reliability based on *inter-item consistency*
- These so-called "item-level" approaches assume that each item on a test is itself a sub-test (like the split halves are considered sub-tests in the split-half method)
- Item-level methods examine the degree of correlation among all items on a scale to provide an estimate of reliability
- This overcomes the "multiple-split problem" of split-half reliability

Internal Consistency Reliability

- We will consider four methods of estimating internal consistency:

1. Split-Half Reliability
2. Coefficient α
3. Standardised Coefficient α
4. KR-20

Internal Consistency Reliability

- We will consider four methods of estimating internal consistency:

  1. Split-Half Reliability
  2. Coefficient α
  3. Standardised Coefficient α
  4. KR-20

Coefficient α

- This is the most widely used method for estimating reliability
- It is usually referred to as Cronbach's α after its developer—Lee Cronbach (1951)
- There are many ways of calculating α
- The book reports two of these methods
- I will illustrate the first method, which to me is the most intuitive

Lee Cronbach (1916–2001)

Coefficient α

- We will calculate α for the four-item scale example from before
- The first thing we need to do is construct the variance–covariance matrix
- It sounds horrible—but don't feel threatened!
- Remember, we covered the concepts of variance and covariance in our Week 2 lecture

Split-Half Reliability: Step 1

Table: Variance–Covariance Matrix For the Four-Item Scale Example

	Item 1	Item 2	Item 3	Item 4
Item 1	1.500	0.000	1.000	0.000
Item 2	0.000	0.686	0.000	0.375
Item 3	1.000	0.000	2.500	1.000
Item 4	0.000	0.375	1.000	0.750

Coefficient α

- The diagonal elements in the matrix are the "item variances"
  - the variances of the distribution of scores for item 1 through to item 4
- The off-diagonal elements in the matrix are the "inter-item covariances"
  - the associations between each item and every other item, as measured by covariance

Coefficient α

- The diagonal elements in the matrix are the "item variances"
  - the variances of the distribution of scores for item 1 through to item 4
- The off-diagonal elements in the matrix are the "inter-item covariances"
  - the associations between each item and every other item, as measured by covariance

Coefficient α

Table: Variance–Covariance Matrix For the Four-Item Scale Example

	Item 1	Item 2	Item 3	Item 4
Item 1	1.500	0.000	1.000	0.000
Item 2	0.000	0.686	0.000	0.375
Item 3	1.000	0.000	2.500	1.000
Item 4	0.000	0.375	1.000	0.750

Coefficient α

Table: Example of Internal Consistency Method of Estimating Reliability

	Items					Split-Half 1		Split-Half 2	
Person	1	2	3	4	Total	"Odd"	"Even"	1 and 4	2 and 4
1	4	4	5	4	17	9	8	8	9
2	5	2	4	2	13	9	4	7	6
3	5	4	2	2	13	7	6	7	6
4	2	3	1	2	8	3	5	4	4
Mean	4	3.25	3	2.5	12.75	7	5.75	6.5	6.25
Variance	1.5	0.6875	2.5	.75	10.1875	6	2.1875	2.25	3.1875

Coefficient α

- The diagonal elements in the matrix are the "item variances"
  - the variances of the distribution of scores for item 1 through to item 4
- The off-diagonal elements in the matrix are the "inter-item covariances"
  - the associations between each item and every other item, as measured by covariance

Coefficient α

- The diagonal elements in the matrix are the "item variances"
  - the variances of the distribution of scores for item 1 through to item 4
- The off-diagonal elements in the matrix are the "inter-item covariances"
  - the associations between each item and every other item, as measured by covariance

Coefficient α

Psychological Measurement

mark.hurlstone@uwa.edu.au

Table: Variance–Covariance Matrix For the Four-Item Scale Example

	Item 1	Item 2	Item 3	Item 4
Item 1	1.500	0.000	1.000	0.000
Item 2	0.000	0.686	0.000	0.375
Item 3	1.000	0.000	2.500	1.000
Item 4	0.000	0.375	1.000	0.750

Coefficient α

- The formula for coefficient α can be expressed as:

$$\alpha = \left(\frac{k}{k-1}\right)\left(\frac{\sum c_{ij}}{s_x^2}\right) \tag{18}$$

- Where k is the number of items
- $\sum c_{ij}$ is the sum of covariances between any particular item (denoted i) and any other item (denoted as j)
- s_x^2 is the variance of the total scores (the sum of all variances and covariances in the matrix)

Coefficient α

- Note that the second term $\left(\frac{\sum c_{ij}}{s_x^2}\right)$ may be thought of as the mean of all possible inter-item correlations
- It provides an overall index of the degree to which all the items on a scale are associated with one another
- The first term $\left(\frac{k}{k-1}\right)$ is the Spearman–Brown correction introduced previously
- It "scales" the reliability estimate derived from the second term according to the length of the test

Coefficient α

- For our example data:

$$\alpha = \text{estimated } R_{xx} = \left(\frac{4}{4-1}\right)\left(\frac{4.75}{10.1875}\right) = (1.333)(0.4663) = .62$$

- The numerator in the second term (4.75) is the sum of <u>covariances</u>
- The denominator in the second term (10.1875) is the sum of <u>variances</u> and <u>covariances</u>

Coefficient α

- For our example data:

$$\alpha = \text{estimated } R_{xx} = \left(\frac{4}{4-1}\right)\left(\frac{4.75}{10.1875}\right) = (1.333)(0.4663) = .62$$

- The numerator in the second term (4.75) is the sum of <u>covariances</u>
- The denominator in the second term (10.1875) is the sum of <u>variances</u> and <u>covariances</u>

Coefficient α

- For our example data:

$$\alpha = \text{estimated } R_{xx} = \left(\frac{4}{4-1}\right)\left(\frac{4.75}{10.1875}\right) = (1.333)(0.4663) = .62$$

- The numerator in the second term (4.75) is the sum of <u>covariances</u>
- The denominator in the second term (10.1875) is the sum of <u>variances</u> and <u>covariances</u>

Coefficient α

- Unlike a correlation coefficient, which ranges in value from –1 to +1, coefficient α typically ranges in value from 0 to 1
- This is because coefficient α—like other coefficients of reliability—is calculated to help answer questions about how *similar* sets of data are
- Here similarity is gauged on a scale from 0 (absolutely no similarity) to 1 (perfectly identical)
- It is possible, however, to conceive of data sets that would yield a negative α value
- Under such rare circumstances the α should be reported as 0

Coefficient α: Assumption 1

- Coefficient α is built on more liberal assumptions than the other reliability methods

1. The α method assumes that test items are *essentially tau equivalent*
    - each item is an equally strong indicator of the true score scores, but they may differ in their precision by a constant
    - in other words, the items can have different means

- This assumption is not made clear in the textbook

Coefficient α: Assumption 2

2. Items can have possibly different error variances

Coefficient α: Assumption 3

3. Error scores should be uncorrelated with true scores—error should be random
   - This assumption has been stated previously in the context of the other methods
   - It is an assumption of all forms of reliability

Coefficient α: Assumption 4

4. Coefficient α assumes that all items used to generate a composite score measure the same attribute or construct

Coefficient α: Some Caveats

- The value of α depends upon the number of items on your scale
- As the number of items increases, so too does the α level
- Thus, "bigger is not always better"—it is possible to get a large α level merely because you have a lot of items on your scale, rather than because your scale is reliable
- Thus, an α level of .9 or greater may be "too high" and indicate redundancy in the items

Coefficient α: Some Caveats

- Coefficient α does not measure "unidimensionality", or the extent to which the scale measures one undelrying factor or construct—this is a common misconception
- Data sets with the same α level can nevertheless have different factor structures
- α should not therefore be used as a measure of unidimensionality
- Cronbach (1951) suggests that if a scale consists of sub-scales, α should be calculated separately for each sub-scale—this follows from assumption 4 (see earlier)

Standardised Coefficient α

- All you have to know about standardised coefficient α is that you apply it to scores that have been converted from a raw score to a standardised score
- For example, if you had z scores and you wanted to calculate the level of internal consistency associated with a composite which consisted of a sum of two or more z scores, you would use the standardised version of coefficient alpha
- In practice, it is not often that you find yourself analysing standardised scores, but it does happen from time to time

KR-20

- Before Cronbach (1951) introduced Coefficient α, Kuder and Richardson (1937) developed a set of formulas for estimating reliability
- The most widely-known of these is the Kuder–Richardson formula 20, or KR–20
- The KR–20 is used for determining the internal consistency reliability of composite scores based on dichotomously scored items
- The formula is shown on p.142 of the textbook (equation 6.5)
- Compare this formula with the second formula for calculating coefficient α on p.138 of the textbook (equation 6.3)

KR-20

- You will notice that the formulas are remarkably similar
- This is because coefficient α is a translation of KR–20
- Coefficient α can be applied to dichotomously scored items and it will produce the exact same result as KR–20
- You don't need to know anything more than the above about KR–20

Factors Affecting Reliability

- Earlier, I mentioned that there are two factors that determine the internal consistency reliability of test scores:

1. The consistency among parts of a test:
   - if the test items are strongly correlated with each other, the test is likely to be reliable
2. The test's length:
   - all things being equal, a longer test will be more reliable than a shorter test

Factors Affecting Reliability: Part Consistency

- The consistency among the parts of a test has a direct effect on reliability estimates
- All things being equal, a test with greater internal consistency will have a greater estimated reliability
- For example, a greater average inter-item covariance will yield a larger value of coefficient α

Factors Affecting Reliability: Test Length

- All things being equal, a long test is more reliable than a short test
- To understand why, know that one way to define reliability under CTT is:

$$R_{xx} = \frac{s_t^2}{s_t^2 + s_e^2}$$

- Where s_t^2 is the true score variance and s_e^2 is the error score variance

Factors Affecting Reliability: Test Length

- Increasing the length of a test—by adding new items that measure the same construct as the original items—will increase the true score variance more than the error variance
- This, in turn, will increase the reliability
- For example, suppose the true score variance for a 10-, 20-, and 30-item test is 300, 450, and 600, respectively
- Further, suppose that the error variance is constant for all three test lengths and is equal to 250

Factors Affecting Reliability: Test Length

- For the 10-item test:

$$R_{xx} = \frac{300}{300 + 250} = \frac{300}{550} = 0.55$$

- For the 20-item test:

$$R_{xx} = \frac{450}{450 + 250} = \frac{450}{700} = 0.64$$

- For the 30-item test:

$$R_{xx} = \frac{600}{600 + 250} = \frac{600}{850} = 0.71$$

References

Furr, M. R., & Bacharach, V. R. (2014; Chapter 6). Psychometrics: An Introduction (second edition). Sage.