Inter-rater Reliability

challenges affecting LAC assessment projects

Inter-rater Reliability

Purposes

Adequate levels ensure

- **accuracy**
- **consistency**

in the assessment.

Inadequate levels indicate

- **scale inadequacy**
- **need for additional rater training**

Inter-rater Reliability

A numerical estimate/measure of the degree of agreement among raters

The basic model for calculating inter-rater reliability is percent agreement in the two-rater model.

ARTIFACTS	RATER 1	RATER 2	AGREEMENT
1	2	2	1
2	3	4	0
3	2	2	1
4	3	3	1
5	3	4	0
			3/5

Inter-rater Reliability

A numerical estimate/measure of the degree of agreement among raters

The basic model for calculating inter-rater reliability is percent agreement in the two-rater model.

1. Calculate the number of ratings that are in agreement

ARTIFACTS	RATER 1	RATER 2	AGREEMENT
1	2	2	1
2	3	4	0
3	2	2	1
4	3	3	1
5	3	4	0
			3/5

Inter-rater Reliability

A numerical estimate/measure of the degree of agreement among raters

The basic model for calculating inter-rater reliability is percent agreement in the two-rater model.

1. Calculate the number/rate of ratings that are in agreement
2. Calculate the total number of ratings

ARTIFACTS	RATER 1	RATER 2	AGREEMENT
1	2	2	1
2	3	4	0
3	2	2	1
4	3	3	1
5	3	4	0
			3/5

Inter-rater Reliability

A numerical estimate/measure of the degree of agreement among raters

The basic model for calculating inter-rater reliability is percent agreement in the two-rater model.

1. Calculate the number/rate of ratings that are in agreement
2. Calculate the total number of ratings
3. Convert the fraction to a percentage

ARTIFACTS	RATER 1	RATER 2	AGREEMENT
1	2	2	1
2	3	4	0
3	2	2	1
4	3	3	1
5	3	4	0
			3/5

Percent Agreement = 60%

Inter-rater Reliability

A numerical estimate/measure of the degree of agreement among raters

The basic model for calculating inter-rater reliability is percent agreement in the two-rater model.

1. Calculate the number/rate of ratings that are in agreement
2. Calculate the total number of ratings
3. Convert the fraction to a percentage

ARTIFACTS	RATER 1	RATER 2	AGREEMENT
1	2	2	1
2	3	4	0
3	2	2	1
4	3	3	1
5	3	4	0
			3/5

Percent Agreement = 60%

What does this mean? How do we interpret the numbers?

Inter-rater Reliability

benchmarking inter-rater reliability

Rules-of-Thumb for Percent Agreement			
Number of Ratings	High Agreement	Minimal Agreement	Qualifications
4 or fewer categories	90%	75%	No ratings more than one level apart
5-7 categories	75%		Approximately 90% of ratings identical or adjacent

Inter-rater Reliability

benchmarking inter-rater reliability

Rules-of-Thumb for Percent Agreement			
Number of Ratings	High Agreement	Minimal Agreement	Qualifications
4 or fewer categories	90%	75%	No ratings more than one level apart
5-7 categories	75%		Approximately 90% of ratings identical or adjacent

Percent Agreement = 60%

What does this mean? Since 60% is lower than the minimal benchmark, inter-rater reliability is unacceptable.

Inter-rater Reliability

generalizing the percent agreement calculation

ARTIFACTS	RATER 1	RATER 2	RATER 3	AGREEMENT
1	2	2	1	?

Calculating the generalized (more than two raters) percent agreement statistic is less intuitive than for the two-rater case.

Inter-rater Reliability

generalizing the percent agreement calculation

ARTIFACTS	RATER 1	RATER 2	RATER 3	AGREEMENT
1	2	2	1	?

- Many assume that since 2 of 3 ratings are identical, the percent agreement for this artifact is 2/3 or 66%.

Calculating the generalized (more than two raters) percent agreement statistic is less intuitive than for the two-rater case.

Inter-rater Reliability

generalizing the percent agreement calculation

ARTIFACTS	RATER 1	RATER 2	RATER 3	AGREEMENT
1	2	2	1	1/3

- Many assume that since 2 of 3 ratings are identical, the percent agreement for this artifact is 2/3 or 66%.
- This assumption is in error: inter-rater reliability is based on agreement between pairs of raters.
  - R1-R2: 1/1
  - R1-R3: 0/1
  - R2-R3: 0/1

Inter-rater Reliability

generalizing the percent agreement calculation

ARTIFACTS	RATER 1	RATER 2	RATER 3	AGREEMENT
1	2	2	1	1/3
2	3	4	4	1/3
3	2	2	2	3/3
4	3	3	3	3/3
5	3	4	3	1/3
6	4	4	4	3/3
				2/3

Percent Agreement = 66% - even though only “3/18 ratings differ.”

Inter-rater Reliability

generalizing the percent agreement calculation

ARTIFACTS	R 1	R2	R 3	AGREE
1	2	2	1	1/3
2	3	4	4	1/3
3	2	2	2	3/3
4	3	3	3	3/3
5	3	4	3	1/3
6	4	4	4	3/3
				2/3

This is an inadequate level of agreement.

Rules-of-Thumb for Percent Agreement			
Number of Ratings	High Agreement	Minimal Agreement	Qualifications
4 or fewer categories	90%	75%	No ratings more than one level apart
5-7 categories	75%		Approximately 90% of ratings identical or adjacent

Percent Agreement = 66% - even though only "3/18 ratings differ."

Inter-rater Reliability

problems with the percent agreement statistic

- **unintuitive and more difficult to hand calculate with multiple raters**
- **absolute agreement is an unforgiving standard**
- **does not take chance agreement into account**

Inter-rater Reliability

problems with the percent agreement statistic

Absolute agreement is an unforgiving standard – a common solution is to count adjacent ratings as being in-agreement.

Inter-rater Reliability

problems with the percent agreement statistic

Absolute agreement is an unforgiving standard – a common solution is to count adjacent ratings as being in-agreement.

ARTIFACTS	RATER 1	RATER 2	AGREE
1	3	2	0
2	3	2	0
3	2	3	0
4	3	2	0
5	2	3	0
6	3	2	0
7	2	3	0
			0/7

Counting adjacent ratings as in-agreement turns this percent agreement = 0

Inter-rater Reliability

problems with the percent agreement statistic

Absolute agreement is an unforgiving standard – a common solution is to count adjacent ratings as being in-agreement.

ARTIFACTS	RATER 1	RATER 2	AGREE
1	3	2	1
2	3	2	1
3	2	3	1
4	3	2	1
5	2	3	1
6	3	2	1
7	2	3	1
			7/7

Counting adjacent ratings as in-agreement turns this percent agreement = 0 into a percent agreement = 100%

Inter-rater Reliability

problems with the percent agreement statistic

This adjustment can be extremely problematic when benchmarks (the just-barely-passing standard) have been identified. As in this case: complete disagreement about each artifact's pass/fail status results in a determination of 'perfect agreement'.

ARTIFACTS	RATER 1	RATER 2	AGREE
1	3	2	1
2	3	2	1
3	2	3	1
4	3	2	1
5	2	3	1
6	3	2	1
7	2	3	1
			7/7

Counting adjacent ratings as in-agreement turns this percent agreement = 0 into a percent agreement = 100%

Inter-rater Reliability

problems with the percent agreement statistic

The percent agreement statistic does not take chance agreement into account – over-estimating the inter-rater reliability estimate.

Inter-rater Reliability

problems with the percent agreement statistic

ARTIFACTS	RATER 1	RATER 2	RATER 3	AGREEMENT
1	1	2	1	1/3

To illustrate the agreement inflating effect of chance, imagine that rater 1 and rater 2 disagree in principle on all ratings; and, that rater 3 uses mercurial, arbitrary rationales for ratings. In this case, R1 and R2 will never agree and R3 will 'agree' with one of them at a rate that depends on the number of levels in the rubric.

In this case, there is no real inter-rater agreement.

Inter-rater Reliability

problems with the percent agreement statistic

The Percent Agreement Statistic

- **unintuitive and more difficult to hand calculate with multiple raters**
- **absolute agreement is an unforgiving standard – and the equating of adjacent ratings with agreement – can result in meaningless reliability estimates**
- **does not take chance agreement into account**

Inter-rater Reliability

optimizing the estimate of agreement for assessment

Characteristics of a more optimal agreement coefficient

- **chance-corrected statistic**
- **resistant to prevalence and marginal probability errors**
- **important benchmark attainment cut-offs can be taken into account**

Inter-rater Reliability

optimizing the estimate of agreement for assessment

Characteristics of an optimal agreement coefficient

- **chance-corrected statistic**
- **resistant to prevalence and marginal probability errors**

Gwet's AC_2 best satisfies these characteristics

Inter-rater Reliability

optimizing the estimate of agreement for assessment

- **Gwet's AC_2 best satisfies these characteristics**
- **in addition, custom weightings can be applied to the calculation of this coefficient that can correctly take into account crucial benchmarks**

Inter-rater Reliability

optimizing the estimate of agreement for assessment

Example 1

	R1	R2
1	1	2
2	1	2
3	3	4
4	2	1
5	3	4
6	4	3
7	3	4

No identical ratings - with complete agreement on pass/fail performance:

Percent Agreement = 0

Gwet's AC_2 (with custom weighting) = .838

Custom weights

	1	2	3	4
1	1.000	0.800	0.000	0.000
2	0.800	1.000	0.000	0.000
3	0.000	0.000	1.000	1.000
4	0.000	0.000	1.000	1.000

Inter-rater Reliability

optimizing the estimate of agreement for assessment

Example 2

	R1	R2
1	1	2
2	1	2
3	3	4
4	2	1
5	3	4
6	4	3
7	3	4

No identical ratings - with complete disagreement on pass/fail performance:

Percent Agreement = 0

Gwet's AC_2 (with custom weighting) = -1.00

Custom weights

	1	2	3	4
1	1.000	0.800	0.000	0.000
2	0.800	1.000	0.000	0.000
3	0.000	0.000	1.000	1.000
4	0.000	0.000	1.000	1.000

Inter-rater Reliability

optimizing the estimate of agreement for assessment

Example 3

	R1	R2
1	2	1
2	2	3
3	3	2
4	3	3
5	4	3
6	3	4
7	4	4

Ratings with limited agreement:

Percent Agreement = .286
Cohen's Kappa (with st. ordinal weighting) = .435
Gwet's AC_2 (with custom weighting) = .446

Custom weights

	1	2	3	4
1	1.000	0.800	0.000	0.000
2	0.800	1.000	0.000	0.000
3	0.000	0.000	1.000	1.000
4	0.000	0.000	1.000	1.000

Inter-rater Reliability

benchmarking inter-rater reliability

Benchmark scales for Kappa's value, as proposed by different investigators

Landis and Koch	Altman	Fleiss
<.0 Poor		
.00 to .20; Slight	<.20 ;Poor	<.40; Poor
.21 to .40; Fair	.21 to .40; Fair	.40 to .75; Intermediate to Good
.41 to .60; Moderate	.41 to .60; Moderate	
.61 to .80; Substantial	.61 to .80; Good	More than .75; Excellent
.81 to 1.00; Almost Perfect	.81 to 1.00; Very Good	

Wongpakaran *et al.*

Wongpakaran *et al. BMC Medical Research Methodology* 2013 **13**:61 doi:10.1186/1471-2288-13-61