Title

stata.com

icc — Intraclass correlation coefficients

Description

`icc` estimates intraclass correlations for one-way random-effects models, two-way random-effects models, or two-way mixed-effects models for both individual and average measurements. Intraclass correlations measuring consistency of agreement or absolute agreement of the measurements may be estimated.

Quick start

Individual and average absolute-agreement intraclass correlation coefficients (ICCs) for ratings `y` of targets identified by `tid` in a one-way random-effects model

```
icc y tid
```

As above, but test that the individual and average ICCs are equal to 0.5

```
icc y tid, testvalue(.5)
```

Absolute-agreement ICCs for targets identified by `tid` and raters identified by `rid` in a two-way random-effects model

```
icc y tid rid
```

As above, but estimate consistency-of-agreement ICCs

```
icc y tid rid, consistency
```

Consistency-of-agreement ICCs when estimating random effects for targets and fixed effects for raters in a mixed-effects model

```
icc y tid rid, mixed
```

As above, but estimate absolute-agreement ICCs

```
icc y tid rid, mixed absolute
```

As above, but report 90% confidence intervals and test that ICCs are equal to 0.3

```
icc y tid rid, mixed absolute level(90) testvalue(.3)
```

Menu

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Intraclass correlations

Syntax

Calculate intraclass correlations for one-way random-effects model

```
icc depvar target [if] [in] [, oneway_options]
```

Calculate intraclass correlations for two-way random-effects model

```
icc depvar target rater [if] [in] [, twoway_re_options]
```

Calculate intraclass correlations for two-way mixed-effects model

```
icc depvar target rater [if] [in], mixed [twoway_me_options]
```

oneway_options	Description
Main	
absolute	estimate absolute agreement; the default
testvalue(#)	test whether intraclass correlations equal #; default is testvalue(0)
Reporting	
level(#)	set confidence level; default is level(95)
format(%*fmt*)	display format for statistics and confidence intervals; default is format(%9.0g)

twoway_re_options	Description
Main	
absolute	estimate absolute agreement; the default
consistency	estimate consistency of agreement
testvalue(#)	test whether intraclass correlations equal #; default is testvalue(0)
Reporting	
level(#)	set confidence level; default is level(95)
format(%*fmt*)	display format for statistics and confidence intervals; default is format(%9.0g)

twoway_me_options	Description
Main	
* `mixed`	estimate intraclass correlations for a mixed-effects model
`consistency`	estimate consistency of agreement; the default
`absolute`	estimate absolute agreement
`testvalue(#)`	test whether intraclass correlations equal #; default is `testvalue(0)`
Reporting	
`level(#)`	set confidence level; default is `level(95)`
`format(%`*fmt*`)`	display format for statistics and confidence intervals; default is `format(%9.0g)`

* `mixed` is required.

`bootstrap`, `by`, `jackknife`, and `statsby` are allowed; see [U] **11.1.10 Prefix commands**.

Options for one-way RE model

Main

`absolute` specifies that intraclass correlations measuring absolute agreement of the measurements be estimated. This is the default for random-effects models.

`testvalue(#)` tests whether intraclass correlations equal #. The default is `testvalue(0)`.

Reporting

`level(#)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`; see [R] **level**.

`format(%`*fmt*`)` specifies how the intraclass correlation estimates and confidence intervals are to be formatted. The default is `format(%9.0g)`.

Options for two-way RE and ME models

Main

`mixed` is required to calculate two-way mixed-effects models. `mixed` specifies that intraclass correlations for a mixed-effects model be estimated.

`absolute` specifies that intraclass correlations measuring absolute agreement of the measurements be estimated. This is the default for random-effects models. Only one of `absolute` or `consistency` may be specified.

`consistency` specifies that intraclass correlations measuring consistency of agreement of the measurements be estimated. This is the default for mixed-effects models. Only one of `absolute` or `consistency` may be specified.

`testvalue(#)` tests whether intraclass correlations equal #. The default is `testvalue(0)`.

Reporting

`level(#)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`; see [R] **level**.

`format(%`*fmt*`)` specifies how the intraclass correlation estimates and confidence intervals are to be formatted. The default is `format(%9.0g)`.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Introduction

In some disciplines, such as psychology and sociology, data are often measured with error that can seriously affect statistical interpretation of the results. Thus it is important to assess the amount of measurement error by evaluating the consistency or reliability of measurements. The intraclass correlation coefficient (ICC) is often used to measure the consistency or homogeneity of measurements.

Several versions of ICCs are introduced in the literature depending on the experimental design and goals of the study (see, for example, Shrout and Fleiss [1979] and McGraw and Wong [1996a]). Following Shrout and Fleiss (1979), we describe various forms of ICCs in the context of a reliability study of ratings of different targets (or objects of measurements) by several raters.

Consider n targets (for example, students, patients, athletes) that are randomly sampled from a population of interest. Each target is rated independently by a set of k raters (for example, teachers, doctors, judges). One rating per target and rater is obtained. It is of interest to determine the extent of the agreement of the ratings.

As noted by Shrout and Fleiss (1979) and McGraw and Wong (1996a), you need to answer several questions to decide what version of ICC is appropriate to measure the agreement in your study:

1. Is a one-way or two-way analysis-of-variance model appropriate for your study?
2. Are differences between raters' mean ratings relevant to the reliability of interest?
3. Is the unit of analysis an individual rating or the mean rating over several raters?
4. Is the consistency of agreement or the absolute agreement of ratings of interest?

Three types of analysis-of-variance models are considered for the reliability study: one-way random effects, two-way random effects, and two-way mixed effects. Mixed models contain both fixed effects and random effects. In the one-way random-effects model, each target is rated by a different set of k independent raters, who are randomly drawn from the population of raters. The target is the only random effect in this model; the effects due to raters and possibly due to rater-and-target interaction cannot be separated from random error. In the two-way random-effects model, each target is rated by the same set of k independent raters, who are randomly drawn from the population of raters. The random effects in this model are target and rater and possibly their interaction, although in the absence of repeated measurements for each rater on each target, the effect of an interaction cannot be separated from random error. In the two-way mixed-effects model, each target is rated by the

same set of k independent raters. Because they are the only raters of interest, rater is a fixed effect. The random effects are target and possibly target-and-rater interaction, but again the interaction effect cannot be separated from random error without repeated measurements for each rater and target. The definition of ICC depends on the chosen random-effects model; see *Methods and formulas* for details.

In summary, use a one-way model if there are no systematic differences in measurements due to raters and use a two-way model otherwise. If you want to generalize your results to a population of raters from which the observed raters are sampled, use a two-way random-effects model, treating raters as random. If you are only interested in the effects of the observed k raters, use a two-way mixed-effects model, treating raters as fixed. For example, suppose you compare judges' ratings of targets from different groups. If you use the combined data from k judges to compare the groups, the random-effects model is appropriate. If you compare groups separately for each judge and then pool the differences, the mixed-effects model is appropriate.

The definition of ICC also depends on the unit of analysis in a study—whether the agreement is measured between individual ratings (individual ICC) or between the averages of ratings over several raters (average ICC). The data on individual ratings are more common. The data on average ratings are typically used when individual ratings are deemed unreliable. The average ICC can also be used when teams of raters are used to rate a target. For example, the ratings of teams of physicians may be evaluated in this manner. When the unit of analysis is an average rating, you should remember that the interpretation of ICC pertains to average ratings and not individual ratings.

Finally, depending on whether consistency of agreement or absolute agreement is of interest, two types of ICC are used: consistency-of-agreement ICC (CA-ICC) and absolute-agreement ICC (AA-ICC). Under consistency of agreement, the scores are considered consistent if the scores from any two raters differ by the same constant value for all targets. This implies that raters give the same ranking to all targets. Under absolute agreement, the scores are considered in absolute agreement if the scores from all raters match exactly.

For example, suppose we observe three targets and two raters. The ratings are (2,4), (4,6), and (6,8), with rater 1 giving the scores (2,4,6) and rater 2 giving the scores (4,6,8), two points higher than rater 1. The CA-ICC between individual ratings is 1 because the scores from rater 1 and rater 2 differ by a constant value (two points) for all targets. That rater 1 gives lower scores than rater 2 is deemed irrelevant under the consistency measure of agreement. The raters have the same difference of opinion on every target, and the variation between raters that is caused by this difference is not relevant. On the other hand, the AA-ICC between individual ratings is $8/12 = 0.67$, where 8 is the estimated between-target variance and 12 is the estimated total variance of ratings.

Either CA-ICC or AA-ICC can serve as a useful measure of agreement depending on whether rater variability is relevant for determining the degree of agreement. As McGraw and Wong (1996a) point out, CA-ICC is useful when comparative judgments are made about objects of measurement. The CA-ICC represents correlation when the rater is fixed; the AA-ICC represents correlation when the rater is random.

See Shrout and Fleiss (1979) and McGraw and Wong (1996a) for more detailed guidelines about the choice of appropriate ICC.

Shrout and Fleiss (1979) and McGraw and Wong (1996a) describe 10 versions of ICCs based on the concepts above: individual and average AA-ICCs for a one-way model (consistency of agreement is not defined for this model); individual and average AA-ICCs and CA-ICCs for a two-way random-effects model; and individual and average AA-ICCs and CA-ICCs for a two-way mixed-effects model. Although each of these ICCs has its own definition and interpretation, the estimators for some are identical, leading to the same estimates of those ICCs; see *Relationship between ICCs* and *Methods and formulas* for details.

The icc command calculates ICCs for each of the three analysis-of-variance models. You can use option absolute to compute AA-ICCs or option consistency to compute CA-ICCs. By default, icc computes ICCs corresponding to the correlation between ratings and between average ratings made on the same target: AA-ICC for a random-effects model and CA-ICC for a mixed-effects model. As pointed out by Shrout and Fleiss (1979), although the data on average ratings might be needed for reliability, the generalization of interest might be individuals. For this reason, icc reports ICCs for both units, individual and average, for each model.

In addition to estimates of ICCs, icc provides confidence intervals and one-sided F tests. The F test of H_o: $\rho = 0$ versus H_a: $\rho > 0$ is the same for the individual and average ICCs, so icc reports one test. This is not true, however, for nonzero null hypotheses (see *Tests against nonzero values* for details), so icc reports a separate test in this case.

The icc command requires data in long form; see [D] **reshape** for how to convert data in wide form to long form. The data must also be balanced and contain one observation per target and rater. For unbalanced data, icc omits all targets with fewer than k ratings from computation. Under one-way models, k is determined as the largest number of observed ratings for a target. Under two-way models, k is the number of unique raters. If multiple observations per target and rater are detected, icc issues an error.

We demonstrate the use of icc using datasets from Shrout and Fleiss (1979) and McGraw and Wong (1996a). In the next three sections, we use an example from table 2 of Shrout and Fleiss (1979) with six targets and four judges. For instructional purposes, we analyze these data under each of the three different models: one-way random effects, two-way random effects, and two-way mixed effects.

One-way random effects

In the one-way random-effects model, we assume that the n targets being rated are randomly selected from the population of potential targets. Each is rated by a different set of k raters randomly drawn from the population of potential raters. McGraw and Wong (1996a) describe an example of this setting, where behavioral genetics data are used to assess familial resemblance. Family units can be viewed as "targets", and children can be viewed as "raters". By taking a measurement on a child of the family unit, we obtain the "rating" of the family unit by the "child-rater". In this case, we can use ICC to assess similarity between children within a family or, in other words, assess if there is a family effect in these data.

As we mentioned in the introduction, only AA-ICC is defined for a one-way model. The consistency of agreement is not defined in this case, as each target is evaluated by a different set of raters. Thus there is no between-rater variability in this model.

In a one-way model, the AA-ICC corresponds to the correlation coefficient between ratings within a target. It is also a ratio of the between-target variance of ratings to the total variance of ratings, the sum of the between-target and error variances.

▷ Example 1: One-way random-effects ICCs

Consider data from table 2 of Shrout and Fleiss (1979) stored in judges.dta. The data contain 24 ratings of $n = 6$ targets by $k = 4$ judges. We list the first eight observations:

```
. use http://www.stata-press.com/data/r15/judges
(Ratings of targets by judges)
. list in 1/8, sepby(target)
```

	rating	target	judge
1.	9	1	1
2.	2	1	2
3.	5	1	3
4.	8	1	4
5.	6	2	1
6.	1	2	2
7.	3	2	3
8.	2	2	4

For a moment, let's ignore that targets are rated by the same set of judges. Instead, we assume that a different set of four judges is used to rate each target. In this case, the only systematic variation in the data is due to targets, so the one-way random-effects model is appropriate.

We use `icc` to estimate the intraclass correlations for these data. To compute ICCs for a one-way model, we specify the dependent variable `rating` followed by the target variable `target`:

```
. icc rating target
Intraclass correlations
One-way random-effects model
Absolute agreement
Random effects: target           Number of targets =         6
                                 Number of raters  =         4
```

rating	ICC	[95% Conf.	Interval]
Individual	.1657418	-.1329323	.7225601
Average	.4427971	-.8844422	.9124154

```
F test that
  ICC=0.00: F(5.0, 18.0) = 1.79                 Prob > F = 0.165
Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

`icc` reports the AA-ICCs for both individual and average ratings. The individual AA-ICC corresponds to ICC(1) in McGraw and Wong (1996a) or ICC(1,1) in Shrout and Fleiss (1979). The average AA-ICC corresponds to ICC(k) in McGraw and Wong (1996a) or ICC(1,k) in Shrout and Fleiss (1979).

The estimated correlation between individual ratings is 0.17, indicating little similarity between ratings within a target, low reliability of individual target ratings, or no target effect. The estimated intraclass correlation between ratings averaged over $k = 4$ judges is higher, 0.44. (The average ICC will typically be higher than the individual ICC.) The estimated intraclass correlation measures the similarity or reliability of mean ratings from groups of four judges. We do not have statistical evidence that either ICC is different from zero based on reported confidence intervals and the one-sided F test.

Note that although the estimates of ICCs cannot be negative in this setting, the lower bound of the computed confidence interval may be negative. A common ad-hoc way of handling this is to truncate the lower bound at zero.

The estimates of both the individual and the average AA-ICC are also computed by the `loneway` command (see [R] **loneway**), which performs a one-way analysis of variance.

◁

❑ Technical note

Mean rating is commonly used when individual rating is unreliable because the reliability of a mean rating is always higher than the reliability of the individual rating when the individual reliability is positive.

In the previous example, we estimated low reliability of the individual ratings of a target, 0.17. The reliability increased to 0.44 for the ratings averaged over four judges. What if we had more judges?

We can use the Spearman–Brown formula (Spearman 1910; Brown 1910) to compute the m-average ICC based on the individual ICC:

$$\text{ICC}(m) = \frac{m\text{ICC}(1)}{1 + (m-1)\text{ICC}(1)}$$

Using this formula for the previous example, we find that the mean reliability over, say, 10 judges is $10 \times 0.17/(1 + 9 \times 0.17) = 0.67$.

Alternatively, we can invert the Spearman–Brown formula to determine the number of judges (or the number of ratings of a target) we need to achieve the desired reliability. Suppose we would like an average reliability of 0.9, then

$$m = \frac{\text{ICC}(m)\{(1 - \text{ICC}(1))\}}{\text{ICC}(1)\{1 - \text{ICC}(m)\}} = \frac{0.9(1 - 0.17)}{0.17(1 - 0.9)} = 44$$

See, for example, Bliese (2000) for other examples.

❑

Two-way random effects

As before, we assume that the targets being rated are randomly selected from the population of potential targets. We now also assume that each target is evaluated by the same set of k raters, who have been randomly sampled from the population of raters. In this scenario, we want to generalize our findings to the population of raters from which the observed k raters were sampled. For example, suppose we want to estimate the reliability of doctors' evaluations of patients with a certain condition. Unless the reliability at a specific hospital is of interest, the doctors may be interchanged with others in the relevant population of doctors.

As for a one-way model, the AA-ICC corresponds to the correlation between measurements on the same target and is also a ratio of the between-target variance to the total variance of measurements in a two-way random-effects model. The total variance is now the sum of the between-target, between-rater, and error variances. Unlike a one-way model, the CA-ICC can be computed for a two-way random-effects model when the consistency of agreement is of interest rather than the absolute agreement. The CA-ICC is also the ratio of the between-target variance to the total variance, but the total variance does not include the between-rater variance because the between-rater variability is irrelevant for the consistency of agreement.

Again, the two versions, individual and average, are available for each ICC.

Example 2: Two-way random-effects ICCs

Continuing with example 1, recall that we previously ignored that each target is rated by the same set of four judges and instead assumed different sets of judges. We return to the original data setting. We want to evaluate the agreement between judges' ratings of targets in a population represented by the observed set of four judges.

In a two-way model, we must specify both the target and the rater variables. In `icc`, we now additionally specify the rater variable `judge` following the target variable `target`; the random-effects model is assumed by default.

```
. icc rating target judge
Intraclass correlations
Two-way random-effects model
Absolute agreement
Random effects: target           Number of targets =         6
Random effects: judge            Number of raters  =         4
```

rating	ICC	[95% Conf.	Interval]
Individual	.2897638	.0187865	.7610844
Average	.6200505	.0711368	.927232

```
F test that
  ICC=0.00: F(5.0, 15.0) = 11.03              Prob > F = 0.000
Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

As for a one-way random-effects model, `icc` by default reports AA-ICCs that correspond to the correlation between ratings on a target. Notice that both individual and average ICCs are larger in the two-way random-effects model than in the previous one-way model—0.29 versus 0.17 and 0.62 versus 0.44, respectively. We also have statistical evidence to reject the null hypothesis that neither ICC is zero based on confidence intervals and the F test. If a one-way model is used when a two-way model is appropriate, the true ICC will generally be underestimated.

The individual AA-ICC corresponds to ICC(A,1) in McGraw and Wong (1996a) or ICC(2,1) in Shrout and Fleiss (1979). The average AA-ICC corresponds to ICC(A,k) in McGraw and Wong (1996a) or ICC(2,k) in Shrout and Fleiss (1979).

Instead of the absolute agreement, we can also assess the consistency of agreement. The individual and average CA-ICCs are considered in McGraw and Wong (1996a) and denoted as ICC(C,1) and ICC(C,k), respectively. These ICCs are not considered in Shrout and Fleiss (1979) because they are not correlations in the strict sense. Although CA-ICCs do not estimate correlation, they can provide useful information about the reliability of the raters. McGraw and Wong (1996a) note that the practical value of the individual and average CA-ICCs in the two-way random-effects model setting is well documented in measurement theory, citing Hartmann (1982) and Suen (1988).

To estimate the individual and average CA-ICCs, we specify the `consistency` option:

```
. icc rating target judge, consistency
Intraclass correlations
Two-way random-effects model
Consistency of agreement
Random effects: target           Number of targets =         6
Random effects: judge            Number of raters  =         4
```

rating	ICC	[95% Conf.	Interval]
Individual	.7148407	.3424648	.9458583
Average	.9093155	.6756747	.9858917

```
F test that
  ICC=0.00: F(5.0, 15.0) = 11.03                Prob > F = 0.000
```

We estimate that the consistency of agreement of ratings in the considered population of raters is high, 0.71, based on the individual CA-ICC. On the other hand, the absolute agreement of ratings is low, 0.29, based on the individual AA-ICC from the previous output.

◁

The measure of consistency of agreement among means, the average CA-ICC, is equivalent to Cronbach's alpha (Cronbach 1951); see [MV] **alpha**. The individual CA-ICC can also be equivalent to the Pearson's correlation coefficient between raters when $k = 2$; see McGraw and Wong (1996a) for details.

In the next example, we will see that the actual estimates of the individual and average AA-ICCs and CA-ICCs are the same whether we examine a random-effects model or a mixed-effects model. The differences between these ICCs are in their definitions and interpretations.

Two-way mixed effects

As in a two-way random-effects model, we assume that the targets are randomly selected from the population of potential targets and that each is evaluated by the same set of k raters. In a mixed-effects model, however, we assume that these raters are the only raters of interest. So as before, the targets are random, but now the raters are fixed.

In the two-way mixed-effects model, the fixed effect of the rater does not contribute to the between-rater random variance component to the total variance. As such, the definitions and interpretations of ICCs are different in a mixed-effects model than in a random-effects model. However, the estimates of ICCs as well as test statistics and confidence intervals are the same. The only exceptions are average AA-ICCs and CA-ICCs. These are not estimable in a two-way mixed-effects model including an interaction term between target and rater; see *Relationship between ICCs* and *Methods and formulas* for details.

In a two-way mixed-effects model, the CA-ICC corresponds to the correlation between measurements on the same target. As pointed out by Shrout and Fleiss (1979), when the rater variance is ignored, the correlation coefficient is interpreted in terms of rater consistency rather than rater absolute agreement. Formally, the CA-ICC is the ratio of the covariance between measurements on the target to the total variance of the measurements. The AA-ICC corresponds to the same ratio, but includes a variance of the fixed factor, rater, in its denominator.

▷ Example 3: Two-way mixed-effects ICCs

Continuing with example 2, suppose that we are now interested in assessing the agreement of ratings from only the observed four judges. The judges are now fixed effects, and the appropriate model is a two-way mixed-effects model.

To estimate ICCs for a two-way mixed-effects model, we specify the `mixed` option with `icc`:

```
. icc rating target judge, mixed

Intraclass correlations
Two-way mixed-effects model
Consistency of agreement

Random effects: target           Number of targets =         6
 Fixed effects: judge            Number of raters  =         4

------------------------------------------------------------------------------
                      rating |        ICC       [95% Conf. Interval]
-----------------------------+------------------------------------------------
                  Individual |   .7148407       .3424648    .9458583
                     Average |   .9093155       .6756747    .9858917
------------------------------------------------------------------------------

F test that
  ICC=0.00: F(5.0, 15.0) = 11.03              Prob > F = 0.000

Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

As we described in the introduction, `icc` by default reports ICCs corresponding to the correlations. So, for a mixed-effects model, `icc` reports CA-ICCs by default. The individual and average CA-ICCs are denoted as ICC(3,1) and ICC(3,k) in Shrout and Fleiss (1979) and ICC(C,1) and ICC(C,k) in McGraw and Wong (1996a).

Our estimates of the individual and average CA-ICCs are identical to the CA-ICC estimates obtained under the two-way random-effects model in example 2, but our interpretation of the results is different. Under a mixed-effects model, 0.71 and 0.91 are the estimates, respectively, of the correlation between individual measurements and the correlation between average measurements made on the same target.

We can also estimate the AA-ICCs in this setting by specifying the `absolute` option:

```
. icc rating target judge, mixed absolute

Intraclass correlations
Two-way mixed-effects model
Absolute agreement

Random effects: target           Number of targets =         6
 Fixed effects: judge            Number of raters  =         4

------------------------------------------------------------------------------
                      rating |        ICC       [95% Conf. Interval]
-----------------------------+------------------------------------------------
                  Individual |   .2897638       .0187865    .7610844
                     Average |   .6200505       .0711368     .927232
------------------------------------------------------------------------------

F test that
  ICC=0.00: F(5.0, 15.0) = 11.03              Prob > F = 0.000
```

The intraclass correlation estimates match the individual and average AA-ICCs obtained under the two-way random-effects model in example 2; but in a mixed-effects model, they do not represent correlations. We demonstrate the use of an individual AA-ICC in a mixed-effects setting in the next example.

The AA-ICCs under a mixed-effects model are not considered by Shrout and Fleiss (1979). They are denoted as ICC(A,1) and ICC(A,k) in McGraw and Wong (1996a).

◁

Adoption study

In this section, we consider the adoption study described in McGraw and Wong (1996a). Adoption studies commonly include two effects of interest. One is the mean difference between the adopted child and its biological parents. It is used to determine if characteristics of adopted children differ on average from those of their biological parents. Another effect of interest is the correlation between genetically paired individuals and genetically unrelated individuals who live together. This effect is used to evaluate the impact of genetic differences on individual differences.

As discussed in McGraw and Wong (1996a), a consistent finding from adoption research using IQ as a trait characteristic is that while adopted children typically have higher IQs than their biological parents, their IQs correlate better with those of their biological parents than with those of their adoptive parents. Both effects are important, and there is additional need to reconcile the two findings. McGraw and Wong (1996a) propose to use the individual AA-ICC for this purpose.

▷ Example 4: Absolute-agreement ICC in a mixed-effects model

The `adoption.dta` dataset contains the data from table 6 of McGraw and Wong (1996a) on IQ scores:

```
. use http://www.stata-press.com/data/r15/adoption
(Biological mother and adopted child IQ scores)
. describe
Contains data from http://www.stata-press.com/data/r15/adoption.dta
  obs:            20                          Biological mother and adopted
                                                child IQ scores
 vars:             5                          15 May 2016 13:50
 size:           160                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
family          byte    %9.0g                 Adoptive family ID
mc              byte    %9.0g      mcvalues   1=Mother, 2=Child
iq3             int     %9.0g                 IQ scores, mother-child
                                                difference of 3 pts
iq9             int     %9.0g                 IQ scores, mother-child
                                                difference of 9 pts
iq15            int     %9.0g                 IQ scores, mother-child
                                                difference of 15 pts
-------------------------------------------------------------------------------
Sorted by:
```

The `family` variable contains adoptive family identifiers, the `mc` variable records a mother or a child, and the `iq3`, `iq9`, and `iq15` variables record IQ scores with differences between mother and child mean IQ scores of 3, 9, and 15 points, respectively.

```
. by mc, sort: summarize iq*
```

```
-> mc = Mother
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         iq3 |         10          97     15.0037        62        116
         iq9 |         10          91     15.0037        56        110
        iq15 |         10          85     15.0037        50        104
```

```
-> mc = Child
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         iq3 |         10         100     15.0037        65        119
         iq9 |         10         100     15.0037        65        119
        iq15 |         10         100     15.0037        65        119
```

The variances of the mother and child IQ scores are the same.

Children are fixed effects, so the mixed-effects model is appropriate for these data. We want to compare individual CA-ICC with individual AA-ICC for each of the three IQ variables. We could issue a separate `icc` command for each of the three IQ variables to obtain the intraclass correlations. Instead, we use `reshape` to convert our data to long form with one `iq` variable and the new `diff` variable recording mean differences:

```
. reshape long iq, i(family mc) j(diff)
(note: j = 3 9 15)
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       20   ->      60
Number of variables                   5   ->       4
j variable (3 values)                     ->   diff
xij variables:
                           iq3 iq9 iq15   ->   iq
-----------------------------------------------------------------------------
```

We can now use the `by` prefix with `icc` to estimate intraclass correlations for the three groups of interest:

```
. by diff, sort: icc iq family mc, mixed
```

```
-> diff = 3

Intraclass correlations
Two-way mixed-effects model
Consistency of agreement

Random effects: family           Number of targets =        10
 Fixed effects: mc               Number of raters  =         2

------------------------------------------------------------------------------
                  iq |        ICC       [95% Conf. Interval]
---------------------+--------------------------------------------------------
          Individual |   .7142152       .1967504     .920474
             Average |   .8332853       .3288078    .9585904
------------------------------------------------------------------------------

F test that
  ICC=0.00: F(9.0, 9.0) = 6.00                 Prob > F = 0.007

Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

```
-> diff = 9

Intraclass correlations
Two-way mixed-effects model
Consistency of agreement

Random effects: family           Number of targets =        10
 Fixed effects: mc               Number of raters  =         2

------------------------------------------------------------------------------
                  iq |        ICC       [95% Conf. Interval]
---------------------+--------------------------------------------------------
          Individual |   .7142152       .1967504     .920474
             Average |   .8332853       .3288078    .9585904
------------------------------------------------------------------------------

F test that
  ICC=0.00: F(9.0, 9.0) = 6.00                 Prob > F = 0.007

Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

```
-> diff = 15
```

(*output omitted*)

The estimated CA-ICCs are the same in all three groups and are equal to the corresponding estimates of the Pearson's correlation coefficients because mothers' and childrens' IQ scores have the same variability. The scores differ only in means, and mean differences are irrelevant when measuring the consistency of agreement.

The AA-ICCs, however, differ across the three groups:

```
. by diff, sort: icc iq family mc, mixed absolute

-------------------------------------------------------------------------------
-> diff = 3

Intraclass correlations
Two-way mixed-effects model
Absolute agreement

Random effects: family           Number of targets =        10
 Fixed effects: mc               Number of raters  =         2

------------------------------------------------------------------------------
                  iq |        ICC       [95% Conf. Interval]
---------------------+--------------------------------------------------------
          Individual |    .7204023       .2275148    .9217029
             Average |    .8374812       .3706917    .9592564
------------------------------------------------------------------------------

F test that
  ICC=0.00: F(9.0, 9.0) = 6.00                 Prob > F = 0.007

-------------------------------------------------------------------------------
-> diff = 9

Intraclass correlations
Two-way mixed-effects model
Absolute agreement

Random effects: family           Number of targets =        10
 Fixed effects: mc               Number of raters  =         2

------------------------------------------------------------------------------
                  iq |        ICC       [95% Conf. Interval]
---------------------+--------------------------------------------------------
          Individual |    .6203378       .0293932    .8905025
             Average |    .7656895       .0571077    .9420802
------------------------------------------------------------------------------

F test that
  ICC=0.00: F(9.0, 9.0) = 6.00                 Prob > F = 0.007

-------------------------------------------------------------------------------
-> diff = 15

Intraclass correlations
Two-way mixed-effects model
Absolute agreement

Random effects: family           Number of targets =        10
 Fixed effects: mc               Number of raters  =         2

------------------------------------------------------------------------------
                  iq |        ICC       [95% Conf. Interval]
---------------------+--------------------------------------------------------
          Individual |    .4854727      -.1194157    .8466905
             Average |    .6536272      -.2712191    .9169815
------------------------------------------------------------------------------

F test that
  ICC=0.00: F(9.0, 9.0) = 6.00                 Prob > F = 0.007
```

As the mean differences increase, the AA-ICCs decrease. Their attenuation reflects the difference in means between biological mother and child IQs while still measuring their agreement. Notice that for small mean differences, the estimates of AA-ICCs and CA-ICCs are very similar.

Note that our estimates match those given in McGraw and Wong (1996b), who correct the original table 6 of McGraw and Wong (1996a).

◁

Relationship between ICCs

In examples 2 and 3, we saw that the estimates of AA-ICCs and CA-ICCs are the same for two-way random-effects and two-way mixed-effects models. In this section, we consider the relationship between various forms of ICCs in more detail; also see *Methods and formulas*.

There are 10 different versions of ICCs, but only six different estimators are needed to compute them. These estimators include the two estimators for the individual and average AA-ICCs in a one-way model, the two estimators for the individual and average AA-ICCs in two-way models, and the two estimators for the individual and average CA-ICCs in two-way models.

Only individual and average AA-ICCs are defined for the one-way model. The estimates of AA-ICCs based on the one-way model will typically be smaller than individual and average estimates of AA-ICCs and CA-ICCs based on two-way models. The estimates of individual and average CA-ICCs will typically be larger than the estimates of individual and average AA-ICCs.

Although AA-ICCs and CA-ICCs have the same respective estimators in two-way random-effects and mixed-effects models, their definitions and interpretations are different. The AA-ICCs based on a random-effects model contain the between-rater variance component in the denominator of the variance ratio. The AA-ICCs based on a mixed-effects model contain the variance of the fixed-factor rater instead of the random between-rater variability. The AA-ICCs in a random-effects model represent correlations between any two measurements made on a target. The AA-ICCs in a mixed-effects model measure absolute agreement of measurements treating raters as fixed. The CA-ICCs for random-effects and mixed-effects models have the same definition but different interpretations. The CA-ICCs represent correlations between any two measurements made on a target in a mixed-effects model but estimate the degree of consistency among measurements treating raters as random in a random-effects model. The difference in the definitions of AA-ICCs and CA-ICCs is that CA-ICCs do not contain the between-rater variance in the denominator of the variance ratio.

For two-way models, the definitions and interpretations (but not the estimators) of ICCs also depend on whether the model contains an interaction between target and rater. For two-way models with interaction, ICCs include an additional variance component for the target-rater interaction in the denominator of the variance ratio. This component cannot be separated from random error because there is only one observation per target and rater.

Also, under a two-way mixed-effects model including interaction, the interaction components are not mutually independent, as they are in a two-way random-effects model. The considered version of the mixed-effects model places a constraint on the interaction effects—the sum of the interaction effects over levels of the fixed factor is zero; see, for example, chapter 7 in Kuehl (2000) for an introductory discussion of mixed models. In this version of the model, there is a correlation between the interaction effects. Specifically, the two interaction effects for the same target and two different raters are negatively correlated. As a result, the estimated intraclass correlation can be negative under a two-way mixed-effects model with interaction. Also, average AA-ICC and average CA-ICC cannot be estimated in a two-way mixed-effects model including interaction; see *Methods and formulas* and McGraw and Wong (1996a) for details.

Tests against nonzero values

It may be of interest to test whether the intraclass correlation is equal to a value other than zero. `icc` supports testing against positive values through the use of the `testvalue()` option. Specifying `testvalue(#)` provides a one-sided hypothesis test of H_o: $\rho = \#$ versus H_a: $\rho > \#$. The test is provided separately for both individual and average ICCs.

▷ Example 5: Testing ICC against a nonzero value

We return to the two-way random-effects model for the judge and target data from Shrout and Fleiss (1979). Suppose we want to test whether the individual and average AA-ICCs are each equal to 0.2. We specify the `testvalue(0.2)` option with `icc`:

```
. use http://www.stata-press.com/data/r15/judges, clear
(Ratings of targets by judges)
. icc rating target judge, testvalue(0.2)
Intraclass correlations
Two-way random-effects model
Absolute agreement
Random effects: target           Number of targets =         6
Random effects: judge            Number of raters  =         4
```

rating	ICC	[95% Conf.	Interval]
Individual	.2897638	.0187865	.7610844
Average	.6200505	.0711368	.927232

```
F test that
  ICC(1)=0.20: F(5.0, 5.3) = 1.54              Prob > F = 0.317
  ICC(k)=0.20: F(5.0, 9.4) = 4.35              Prob > F = 0.026
Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

We reject the null hypothesis that the average AA-ICC, labeled as ICC(k) in the output, is equal to 0.2, but we do not have statistical evidence to reject the null hypothesis that the individual AA-ICC, labeled as ICC(1), is equal to 0.2.

◁

Stored results

`icc` stores the following in `r()`:

Scalars

`r(N_target)`	number of targets
`r(N_rater)`	number of raters
`r(icc_i)`	intraclass correlation for individual measurements
`r(icc_i_F)`	F test statistic for individual ICC
`r(icc_i_df1)`	numerator degrees of freedom for `r(icc_i_F)`
`r(icc_i_df2)`	denominator degrees of freedom for `r(icc_i_F)`
`r(icc_i_p)`	p-value for F test of individual ICC
`r(icc_i_lb)`	lower endpoint for confidence intervals of individual ICC
`r(icc_i_ub)`	upper endpoint for confidence intervals of individual ICC
`r(icc_avg)`	intraclass correlation for average measurements
`r(icc_avg_F)`	F test statistic for average ICC
`r(icc_avg_df1)`	numerator degrees of freedom for `r(icc_avg_F)`
`r(icc_avg_df2)`	denominator degrees of freedom for `r(icc_avg_F)`
`r(icc_avg_p)`	p-value for F test of average ICC
`r(icc_avg_lb)`	lower endpoint for confidence intervals of average ICC
`r(icc_avg_ub)`	upper endpoint for confidence intervals of average ICC
`r(testvalue)`	null hypothesis value
`r(level)`	confidence level

Macros

`r(model)`	analysis-of-variance model
`r(depvar)`	name of dependent variable
`r(target)`	target variable
`r(rater)`	rater variable
`r(type)`	type of ICC estimated (`absolute` or `consistency`)

Methods and formulas

We observe y_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, k$. y_{ij} is the jth rating on the ith target. Let $\alpha = 1 - l/100$, where l is the significance level specified by the user.

Methods and formulas are presented under the following headings:

Mean squares

The mean squares within targets are

$$\text{WMS} = \sum_i \sum_j \frac{(y_{ij} - \overline{y}_{i\cdot})^2}{n(k-1)}$$

where $\overline{y}_{i\cdot} = \sum_j y_{ij}/k$.

The mean squares between targets are

$$\text{BMS} = \sum_i \frac{(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2}{n-1}$$

where $\overline{y}_{\cdot\cdot} = \sum_i \overline{y}_{i\cdot}/n$.

These are the only mean squares needed to estimate ICC in the one-way random-effects model. For the two-way models, we need two additional mean squares.

The mean squares between raters are

$$\text{JMS} = \sum_j \frac{(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})^2}{k-1}$$

where $\overline{y}_{\cdot j} = \sum_i y_{ij}/n$ and $\overline{y}_{\cdot\cdot} = \sum_j \overline{y}_{\cdot j}/k$.

The residual or error mean square is

$$\text{EMS} = \frac{\sum_i \sum_j (y_{ij} - \overline{y})^2 - (k-1)\text{JMS} - (n-1)\text{BMS}}{(n-1)(k-1)}$$

One-way random effects

Under the one-way random-effects model, we observe

$$y_{ij} = \mu + r_i + \epsilon_{ij} \tag{M1}$$

where μ is the mean rating, r_i is the target random effect, and ϵ_{ij} is random error. The r_is are i.i.d. $N(0, \sigma_r^2)$; ϵ_{ij}s are i.i.d. $N(0, \sigma_\epsilon^2)$ and are independent of r_is. There is no rater effect separate from the residual error because each target is evaluated by a different set of raters.

The individual AA-ICC is the correlation between individual measurements on the same target:

$$\rho_1 = \text{ICC}(1) = \text{Corr}(y_{ij}, y_{ij'}) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_\epsilon^2}$$

The average AA-ICC is the correlation between average measurements of size k made on the same target:

$$\rho_k = \text{ICC}(k) = \text{Corr}(\overline{y}_{i\cdot}, \overline{y}'_{i\cdot}) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_\epsilon^2/k}$$

They are estimated by

$$\widehat{\rho}_1 = \widehat{\text{ICC}(1)} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (k-1)\text{WMS}}$$

$$\widehat{\rho}_k = \widehat{\text{ICC}(k)} = \frac{\text{BMS} - \text{WMS}}{\text{BMS}}$$

Confidence intervals. Let $F_{\text{obs}} = \text{BMS}/\text{WMS}$, let F_l be the $(1-\alpha/2) \times 100$th percentile of the $F_{n-1,n(k-1)}$ distribution, and let F_u be the $(1-\alpha/2) \times 100$th percentile of the $F_{n(k-1),n-1}$ distribution. Let $F_L = F_{\text{obs}}/F_l$ and $F_U = F_{\text{obs}}F_u$.

A $(1-\alpha) \times 100\%$ confidence interval for ρ_1 is

$$\left(\frac{F_L - 1}{F_l + k - 1}, \frac{F_U - 1}{F_U + k - 1} \right) \tag{1}$$

A $(1-\alpha) \times 100\%$ confidence interval for ρ_k is

$$\left(1 - \frac{1}{F_L}, 1 - \frac{1}{F_U} \right) \tag{2}$$

Hypothesis tests. Consider a one-sided hypothesis test of H_o: ICC $= \rho_0$ versus H_a: ICC $> \rho_0$.

The test statistic for ρ_1 is

$$F_{\rho_1} = \frac{\text{BMS}}{\text{WMS}} \frac{1-\rho_0}{1+(k-1)\rho_0} \tag{3}$$

The test statistic for ρ_k is

$$F_{\rho_k} = \frac{\text{BMS}}{\text{WMS}}(1-\rho_0) \tag{4}$$

Under the null hypothesis, both F_{ρ_1} and F_{ρ_k} have the $F_{n-1,n(k-1)}$ distribution. When $\rho_0 = 0$, the two test statistics coincide.

Two-way random effects

In this setting, the target is evaluated by the same set of raters, who are randomly drawn from the population of raters. The underlying models with and without interaction are

$$y_{ij} = \mu + r_i + c_j + (rc)_{ij} + \epsilon_{ij} \tag{M2}$$

$$y_{ij} = \mu + r_i + c_j + \epsilon_{ij} \tag{M2A}$$

where y_{ij} is the rating of the ith target by the jth rater, μ is the mean rating, r_i is the target random effect, c_j is the rater random effect, $(rc)_{ij}$ is the target-rater random effect, and ϵ_{ij} is random error. The r_is are i.i.d. $N(0, \sigma_r^2)$, c_js are i.i.d. $N(0, \sigma_c^2)$, $(rc)_{ij}$s are i.i.d. $N(0, \sigma_{rc}^2)$, and ϵ_{ij}s are i.i.d. $N(0, \sigma_\epsilon^2)$. Each effect is mutually independent of the others.

Below we provide formulas for ICCs for model (M2). The corresponding ICCs for model (M2A) can be obtained by setting $\sigma_{rc}^2 = 0$.

The individual AA-ICC is the correlation between individual measurements on the same target:

$$\rho_{A,1} = \text{ICC}(A,1) = \text{Corr}(y_{ij}, y_{ij'}) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + (\sigma_{rc}^2 + \sigma_\epsilon^2)}$$

The average AA-ICC is the correlation between average measurements of size k made on the same target:

$$\rho_{A,k} = \text{ICC}(A,k) = \text{Corr}(\overline{y}_{i\cdot}, \overline{y}'_{i\cdot}) = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_c^2 + \sigma_{rc}^2 + \sigma_\epsilon^2)/k}$$

The consistency-of-agreement intraclass correlation for individual measurements, individual CA-ICC, is

$$\rho_{C,1} = \text{ICC}(C,1) = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_\epsilon^2)}$$

The consistency-of-agreement intraclass correlation for average measurements of size k, average CA-ICC, is

$$\rho_{C,k} = \text{ICC}(C,k) = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_\epsilon^2)/k}$$

With one observation per target and rater, σ_{rc}^2 and σ_ϵ^2 cannot be estimated separately.

The estimators of intraclass correlations, confidence intervals, and test statistics are the same for models (M2) and (M2A). The estimators of ICCs are

$$\widehat{\rho}_{A,1} = \widehat{\text{ICC}(A,1)} = \frac{\text{BMS} - \text{EMS}}{\text{BMS} + (k-1)\text{EMS} + \frac{k}{n}(\text{JMS} - \text{EMS})}$$
$$\widehat{\rho}_{A,k} = \widehat{\text{ICC}(A,k)} = \frac{\text{BMS} - \text{EMS}}{\text{BMS} + \frac{1}{n}(\text{JMS} - \text{EMS})}$$
$$\widehat{\rho}_{C,1} = \widehat{\text{ICC}(C,1)} = \frac{\text{BMS} - \text{EMS}}{\text{BMS} + (k-1)\text{EMS}}$$
$$\widehat{\rho}_{C,k} = \widehat{\text{ICC}(C,\text{k})} = \frac{\text{BMS} - \text{EMS}}{\text{BMS}}$$

Confidence intervals. Let $a = k\widehat{\rho}_{A,1}/\{n(1-\widehat{\rho}_{A,1})\}$, $b = 1 + k\widehat{\rho}_{A,1}(n-1)/\{n(1-\widehat{\rho}_{A,1})\}$, and

$$v = \frac{(a\text{JMS} + b\text{EMS})^2}{\frac{a^2\text{JMS}^2}{k-1} + \frac{b^2\text{EMS}^2}{(n-1)(k-1)}} \tag{5}$$

Let F_l be the $(1-\alpha/2)\times 100$th percentile of the $F_{n-1,v}$ distribution and F_u be the $(1-\alpha/2)\times 100$th percentile of the $F_{v,n-1}$ distribution.

A $(1-\alpha)\times 100\%$ confidence interval for $\rho_{A,1}$ is given by (L,U), where

$$L = \frac{n(\text{BMS} - F_l\text{EMS})}{F_l\{k\text{JMS} + (kn-k-n)\text{EMS}\} + n\text{BMS}}$$
$$U = \frac{n(F_u\text{BMS} - \text{EMS})}{k\text{JMS} + (kn-k-n)\text{EMS} + nF_u\text{BMS}} \tag{6}$$

A $(1-\alpha)\times 100\%$ confidence intervals for $\rho_{A,k}$ is a special case of (6) with $k=1$, where $a = \widehat{\rho}_{A,k}/\{n(1-\widehat{\rho}_{A,k})\}$, $b = 1 + \widehat{\rho}_{A,k}(n-1)/\{n(1-\widehat{\rho}_{A,k})\}$, and v is defined in (5).

To define confidence intervals for $\rho_{C,1}$ and $\rho_{C,k}$, let $F_{\text{obs}} = \text{BMS}/\text{EMS}$, F_l be the $(1-\alpha/2)\times 100$th percentile of the $F_{n-1,(n-1)(k-1)}$ distribution, and F_u be the $(1-\alpha/2)\times 100$th percentile of the $F_{(n-1)(k-1),n-1}$ distribution. Let $F_L = F_{\text{obs}}/F_l$ and $F_L = F_{\text{obs}}F_u$.

A $(1-\alpha)\times 100\%$ confidence intervals for $\rho_{C,1}$ and $\rho_{C,k}$ are then as given by (1) and (2) for model (M1).

Hypothesis tests. Consider a one-sided hypothesis test of H_o: ICC $= \rho_0$ versus H_a: ICC $> \rho_0$. Let $a = k\rho_0/\{n(1-\rho_0)\}$ and $b = 1 + k\rho_0(n-1)/\{n(1-\rho_0)\}$.

The test statistic for $\rho_{A,1}$ is

$$F_{\rho_{A,1}} = \frac{\text{BMS}}{a\text{JMS} + b\text{EMS}}$$

Under the null hypothesis, $F_{\rho_{A,1}}$ has the $F_{n-1,v}$ distribution, where v is defined in (5).

The test statistic for $\rho_{A,k}$ is defined similarly, except $a = \rho_0/\{n(1-\rho_0)\}$ and $b = 1 + \rho_0(n-1)/\{n(1-\rho_0)\}$. Under the null hypothesis, $F_{\rho_{A,k}}$ has the $F_{n-1,v}$ distribution, where v is defined in (5). When $\rho_0 = 0$, then $a = 0$, $b = 1$, and the two test statistics coincide.

The test statistics for $\rho_{C,1}$ and $\rho_{C,k}$ are defined by (3) and (4), respectively, with WMS replaced by EMS. Under the null hypothesis, both $F_{\rho_{C,1}}$ and $F_{\rho_{C,k}}$ have the $F_{n-1,(n-1)(k-1)}$ distribution. They also both have the same value when $\rho_0 = 0$.

Two-way mixed effects

In this setting, every target is evaluated by the same set of judges, who are the only judges of interest. The underlying models with and without interaction are

$$y_{ij} = \mu + r_i + c_j + (rc)_{ij} + \epsilon_{ij} \tag{M3}$$

$$y_{ij} = \mu + r_i + c_j + \epsilon_{ij} \tag{M3A}$$

where y_{ij} is the rating of the ith target by the jth rater, μ is the mean rating, r_i is the target random effect, c_j is the rater random effect, $(rc)_{ij}$ is an interaction effect between target and rater, and ϵ_{ij} is random error. The r_is are i.i.d. $N(0, \sigma_r^2)$, $(rc)_{ij}$s are $N(0, \sigma_{rc}^2)$, and ϵ_{ij}s are i.i.d. $N(0, \sigma_\epsilon^2)$. Each random effect is mutually independent of the others. The c_js are fixed such that $\sum_j c_j = 0$. The variance of c_js is $\theta_c^2 = \sum c_j^2/(k-1)$.

In the presence of an interaction, two versions of a mixed-effects model may be considered. One assumes that $(rc)_{ij}$s are i.i.d. $N(0, \sigma_{rc}^2)$. Another assumes that $(rc)_{ij}$s are $N(0, \sigma_{rc}^2)$ with an additional constraint that $\sum_j (rc)_{ij} = 0$ (for example, Kuehl [2000]), so only interaction terms involving different targets are independent. The latter model is considered here.

We now define the intraclass correlations for individual measurements for model (M3).

The individual CA-ICC, the correlation between individual measurements on the same target, is

$$\rho_{C,1} = \text{ICC}(C,1) = \text{Corr}(y_{ij}, y_{ij'}) = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_\epsilon^2)}$$

The absolute-agreement intraclass correlation for individual measurements, individual AA-ICC, is

$$\rho_{A,1} = \text{ICC}(A,1) = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + \theta_c^2 + (\sigma_{rc}^2 + \sigma_\epsilon^2)}$$

Shrout and Fleiss (1979) show that the individual ICC could be negative in this case—a phenomenon first pointed out by Sitgreaves (1960). This can happen when the interaction term has a high variance relative to the targets and there are not many raters.

The individual intraclass correlations for model (M3A) have similar definitions with $\sigma_{rc}^2 = 0$. The individual CA-ICC is the correlation between individual measurements on the same target, $\text{Corr}(y_{ij}, y_{ij'})$.

We now discuss the intraclass correlations that correspond to average measurements. Neither average AA-ICC, $\rho_{A,k}$, nor average CA-ICC, $\rho_{C,k}$, can be estimated under model (M3) (Shrout and Fleiss 1979; McGraw and Wong 1996a). The problem is that in this model, σ_r^2, which is the covariance between two means based on k raters, cannot be estimated.

Specifically, the parameter σ_r^2 appears only in the expectation of the between-target mean squares BMS. Under the restriction $\sum_j (rc)_{ij} = 0$,

$$E(\text{BMS}) = k\sigma_r^2 + \sigma_\epsilon^2$$

Note that σ_{rc}^2 does not appear in the expectation of between-target mean squares. With one observation per target and rater, σ_{rc}^2 and σ_ϵ^2 cannot be estimated separately (only their sum $\sigma_{rc}^2 + \sigma_\epsilon^2$ can be estimated), so BMS alone cannot be used to estimate σ_r^2.

Under model (M3A), however, there is no interaction (and thus no interaction variance component σ^2_{rc}), so $\rho_{A,k}$ or $\rho_{C,k}$ can be estimated.

The average AA-ICC, the absolute-agreement intraclass correlation for average measurements of size k, is

$$\rho_{A,k} = \text{ICC}(A,k) = \frac{\sigma^2_r}{\sigma^2_r + (\theta^2_c + \sigma^2_\epsilon)/k}$$

The average CA-ICC, the correlation between average measurements of size k made on the same target, is

$$\rho_{C,k} = \text{ICC}(C,k) = \text{Corr}(\overline{y}_{i\cdot}, \overline{y}'_{i\cdot}) = \frac{\sigma^2_r}{\sigma^2_r + \sigma^2_\epsilon/k}$$

The estimators of ICCs, their confidence intervals, and hypothesis tests are as described for two-way random-effects models, except $\rho_{A,k}$ and $\rho_{C,k}$ are not defined under model (M3).

References

Bliese, P. D. 2000. Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*, ed. K. J. Klein and S. W. J. Kozlowski, 349–381. San Francisco: Jossey-Bass.

Brown, W. 1910. Some experimental results in the correlation of mental abilities. *British Journal of Psychology* 3: 296–322.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.

Hartmann, D. P. 1982. Assessing the dependability of observational data. In *Using Observers to Study Behavior*, 51–65. San Francisco: Jossey-Bass.

Kuehl, R. O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis*. 2nd ed. Belmont, CA: Duxbury.

McGraw, K. O., and S. P. Wong. 1996a. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1: 30–46.

——. 1996b. Forming inferences about some intraclass correlation coefficients: Correction. *Psychological Methods* 1: 390.

Shrout, P. E., and J. L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86: 420–428.

Sitgreaves, R. 1960. Book reviews: Intraclass Correlation and the Analysis of Variance, Ernest A. Haggard. *Journal of the American Statistical Association* 55: 384–385.

Spearman, C. E. 1910. Correlation calculated from faulty data. *British Journal of Psychology* 3: 271–295.

Suen, H. K. 1988. Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment* 10: 343–366.

Also see

[R] **anova** — Analysis of variance and covariance

[R] **correlate** — Correlations (covariances) of variables or coefficients

[R] **loneway** — Large one-way ANOVA, random effects, and reliability

[MV] **alpha** — Compute interitem correlations (covariances) and Cronbach's alpha