# Why the ICC changes from one model to another

December 6, 2013

ICC (intra-class correlation) in two-level models is defined as

$$ICC(Y) = \frac{Var(Y_B)}{Var(Y_B) + Var(Y_W)}$$

where

$$Y = \mu + Y_W + Y_B$$

is the two-level decomposition of the observed variable $Y$ as a within level portion $Y_W$ and between level portion $Y_B$. Both $Y_W$ and $Y_B$ are latent variables and their variances need to be estimated using the ML maximization algorithm, i.e., $Var(Y_B)$ and $Var(Y_W)$ are in general parameters that are not directly related to an explicit sample statistic. There is one exception to this rule. If there is no missing data, all clusters are of the same size and no variable is declared on the WITHIN= list, then in that case there are explicit formulas for $Var(Y_B)$ and $Var(Y_W)$: see formulas (36) and (37) in

Asparouhov, T. (2006). General multi-level modeling with sampling weights. Communications in Statistics: Theory and Methods, Volume 35, Number 3, 2006, pp. 439-460(22).

http://statmodel.com/download/asparouhovgmms.pdf

If any of these 3 conditions are violated then $Var(Y_B)$ and $Var(Y_W)$ are estimated iteratively. In addition these parameters are not estimated from a univariate model but they are estimated from the mutivariate H1 unrestricted model. That is, if the model has $p$ variables we estimate the following multivariate model

$$Y_1 = \mu_1 + Y_{1W} + Y_{1B}$$
$$Y_2 = \mu_2 + Y_{2W} + Y_{2B}$$

....

$$Y_p = \mu_p + Y_{pW} + Y_{pB}$$

where $\mu_i$ are the mean parameters and the variance of the $Y_W$ vector is the fully unconstrained matrix $\Theta_W$ and the variance of the $Y_B$ vector is the fully unconstrained matrix $\Theta_B$. The parameters $\Theta_W$, $\Theta_B$, $\mu$ are reported under the heading **sample statistics** although they are actually iteratively estimated parameters. These parameters are used to compute the ICC. Although the ICC is a statistic for one variable it can be affected by the rest of the variables in the model because all the variables and their variance decomposition are estimated together. Again, unless the above three conditions are met the ICC depends not only on the variable itself but also on the other variables. The only case when the ICC will be independent of the other variables is when

1. there is no missing data
2. all clusters are of the same size
3. no variable is declared on the WITHIN= list

If conditions 1-3 are satisfied the ICC will be identical to the ICC computed from the univariate model where the $Y$ variable decomposition is estimated alone.

Failure of conditions 1 and 2 above generally will yield minor changes in the ICC. Failure of condition 3 however can bring about a bigger change, in particular when the WITHIN= list is used inappropriately. Setting a variable on the within list means that the between effect is non-existent and if that is incorrect specification it may lead to large biases in the ICC estimation. If for example the variable $Y_1$ is specified as a within only the H1 model that we now estimate will be

$$Y_1 = \mu_1 + Y_{1W}$$
$$Y_2 = \mu_2 + Y_{2W} + Y_{2B}$$

....

$$Y_p = \mu_p + Y_{pW} + Y_{pB}$$

i.e., the $Y_{1B}$ component is eliminated from the model, i.e., it is fixed to 0 and so are its variance parameters and all covariance parameters $\theta_{1i,B}$. If this

kind of a restriction is grossly inaccurate the ML estimation will increase $\theta_{1i,W}$ on the within level to account for the total correlations between the variables and with that the variances of the other variables on the within level can increase as well, leading to biased ICC estimates for the remaining variables in the model.

Here is a simple montecarlo illustration. Using the following Montecarlo run we get $ICC(Y_1) = 0.844$

```
montecarlo:
names are y1-y2;
nobservations = 10000;
ncsizes = 1;
csizes = 500 (20);
nreps = 1;
save = 1.dat;
ANALYSIS: TYPE = TWOLEVEL;
model population:
%within%
y1*0.2 y2*.01;
%between%
y1-y2*1;
y1 with y2*0.9;
model: %within%
y1*0.2 y2*.01;
%between%
y1-y2*1;
y1 with y2*0.9;
```

If now we incorrectly specify $Y_2$ as within= variable as in the following input we get $ICC(Y_1) = 0.181$. The true value is of course $1/1.2 = 0.83$.

```
variable:
names are y1-y2 c;
cluster=c;
within=y2;
```

data: file = 1.dat;
ANALYSIS: TYPE = TWOLEVEL;

The bias in ICC is again due to misspecifying the variable $Y_2$ as a within only. That kind of misspecification may impact not just the ICC but also model parameters, but in many cases this misspecification will NOT lead to incorrect results. In principle one should carefully check that a variable on the within= list doesn't have any between level variation. That can be done by estimating a basic TWOLEVEL model without the within= command and if a variable has an ICC $\leq 0.05$ then it can be specified as within only.

Also a variable can be specified on the within= list if it is group mean centered first

DEFINE: CENTER $Y_2$ (GROUPMEAN);

Other modeling alternatives are discussed in

Ludtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthen, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. Psychological Methods, 13, 203-229.

http://statmodel.com/download/Ludtkeposted.pdf