

Appendix 1. Components of outcome measurement instruments that do not involve biological sampling^{a 1}

Component	Elaboration	Examples
Equipment	All equipment necessary in the preparation, the administration, and the assignment of scores of the outcome measurement instrument	Questionnaire forms, computers, tablet, pen and paper; stair steps of a specific height; device or tools (such as stopwatch, probe, tube); ultrasound machine, ultrasound gels, MRI scanner; software.
Preparatory actions preceding raw data collection by professionals, patients, and others (if applicable)	<p>1. General preparatory actions, such as required expertise or training for professionals to prepare, administer, store or assign the scores</p> <p>2. Specific preparatory actions for each measurement, such as</p> <ul style="list-style-type: none"> • preparations of equipment, environment, storage by professionals^b • preparations of the patient^c by the professional • Preparations undertaken by the patients 	<p>Training, education or experience required, certification.</p> <p>Preparation of equipment: calibration of device/equipment, adjust settings of the machine.</p> <p>Preparation of the environment: light conditions, room temperature, humidity, specific length of a walking track.</p> <p>Preparation for storage: design database and logbook</p> <p>Provide general and preparatory instructions for the patients, such as explaining the tasks/action that need to be performed including time schedule, safety issues and side effects; instructions on diet (e.g. use of caffeine), clothing (e.g. comfortable shoes, no jewelry, glasses or devices), performance during tests (e.g. perform a task as usual; try to walk as fast as you can; lie as calm as possible); set some training or perform a familiarization session.</p> <p>Attaching electrodes to the body, injection with radioactive substance or contrast dye, positioning the patient, applying ultrasound gel.</p> <p>Listen to and understanding the instructions provided; adherence to the preparatory instructions such as fasting, resting, taking medication, bowel preparation, exercising, shaving.</p>

Component	Elaboration	Examples
Collection of raw data	All actions undertaken by patient and professional(s) to collect the data, before any data processing	The patient completing questions at home, or at the hospital; or performing the tasks; the rater observing or timing the performance; switching the imaging device on and off; positioning and moving the ultrasound probe.
Data processing and storage	All actions undertaken on the raw data to store it in a usable (electronic) form for later data manipulation (such as score assignment or statistical analysis)	The digitally converted signal of a specific body MRI scan which is temporarily stored in the K-space, is sent to an image processor where a mathematical formula (i.e. Fourier transformation) is applied, leading to an image which is displayed on a monitor and saved on a computer; Other examples: answers of question items are recorded on e.g. paper forms and stored or Likert scale format response options are converted into a 0-4 score and directly entered in a computer database. Performance of data quality checks e.g. double entry or validation checks on the stored/entered data.
Assignment of the score(s)	Methods used to convert processed data into a score ^d that constitutes the outcome measurement instrument.	A calculation of a mathematical formula or the application of a scorings algorithm (e.g. a set of rules to be followed) to the processed data; a clinician selects the specific images and judges the severity and quantity of e.g. lesions on the set of images or compares it to a reference; scores adjusted for e.g. missing data or patients using devices such as mobility aids.

^a A description of components of outcome measurement instruments that involve biological sampling (i.e. laboratory values) is provided in Mokkink et al. 2020 ²; ^b Professionals are those who are involved in the preparation or the performance of the measurement, in the data processing, or in the assignment of the score; this may be done by one and the same person, or by different persons; ^c In the COSMIN methodology we use the word ‘patient.’ However, sometimes the target population is not patients, but e.g. healthy individuals, caregivers, or clinicians, or a part of the body (e.g. joint, or lesion). In these cases, the word patient should be read as e.g. healthy volunteer, or clinician; ^d The score can be further used or interpreted, by converting a score to another scale, metric or classification. For example, a continuous score is classified into an ordinal score (e.g. mild/moderate/severe), a score is dichotomized into below or above a normal value, patients are classified as responder to the intervention (e.g. when their change is larger than the Minimal Important Change (MIC) value).

1. Mokkink LB, Boers M, van der Vleuten CPM, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Medical Research Methodology*. 2020;20(293)doi:<https://doi.org/10.1186/s12874-020-01179-5>

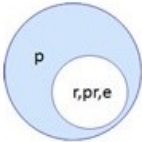
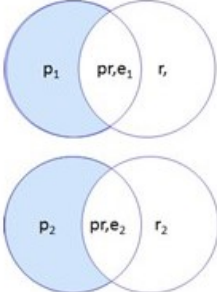
Appendix 2. Designs for nested studies.

In a nested design the object of measurement (patients) is not measured by all elements of the facet of generalization of interest (e.g. raters). For example we could do a nested inter-rater reliability study where three pairs of raters (i.e. three different measurement conditions) each measure one-third of the patients. Another example is a nested intra-rater reliability study, where half of the patients is measured twice by rater A, and the other half twice by rater B. Both situations are written as 'r : p' ('rater nested in patient'). These nested designs can be very efficient because the raters have to perform fewer assessments. In addition, this is also efficient from a practical point of view, because logistically this can often be better arranged. However, more complex statistics are required to take these different conditions into account.

Two-way nested design

In a two-way nested design, multiple measurement conditions are used for one facet of generalization. For example, in a nested inter-rater reliability study, some of the patients are measured under one measurement condition of the facet of generalization (e.g. by raters A and B) while other patients are measured under another measurement condition (e.g. by raters C and D), or another combination of two raters (Supplementary Figure 1 Appendix 2 data collection scheme). In a nested design there are two ways to estimate the agreement parameters. Results of the two methods may slightly differ as different effects models are used, resulting in a slight different estimation of the variance components. Either the design can be considered as nested, and analyzed as a one-way random effects model (Supplementary Figure 1 Appendix 2 method 1) ¹, or the design can be considered as a series of crossed studies. That is, per measurement condition the variance components are estimated, and next the variance components per source of variation are pooled

(i.e. adding them up and dividing by the number of measurement conditions (e.g. 2). We assume here equal sample size for each facet across the series of studies. Variance components with unequal sample sizes are differently pooled as explained here. The pooled variance components are subsequently used in the calculation of the ICC and SEM agreement parameters (Supplementary Figure 1 Appendix 2 method 2 where we use 2 measurement conditions).

Design (r : p)		Venn diagram	Statistical formulas	
Patient	Rater	Method 1: consider it as fully nested:	Method 1, using one-way random effects model:	
1	a b		ICC = $\frac{\sigma_p^2}{\sigma_p^2 + \sigma_{r,pr,e}^2}$	
...	a b		SEM = $\sqrt{\sigma_{r,pr,e}^2}$	
10	a b		Method 2: consider it as two separate crossed studies 	Method 2, using two two-way random effects models*, with pooled variance components:
11	c d			ICC _{agreement} =
...	c d			$\frac{(\sigma_{p1}^2 + \sigma_{p2}^2)}{2}$ $\frac{(\sigma_{p1}^2 + \sigma_{p2}^2)}{2} + \frac{(\sigma_{r1}^2 + \sigma_{r2}^2)}{2} + \frac{(\sigma_{pr,e1}^2 + \sigma_{pr,e2}^2)}{2}$
n	c d	SEM _{agreement} =		
			$\sqrt{\left[\frac{(\sigma_{r1}^2 + \sigma_{r2}^2)}{2} + \frac{(\sigma_{pr,e1}^2 + \sigma_{pr,e2}^2)}{2} \right]}$	

Appendix 2 Figure 1. Two-way random effects model for agreement (ICC (2.1)) in a nested design.

n = total number of included patients; A, B, C, D = refer to a specific rater; Blue surface = variation of interest, white surface = measurement error of interest; red surface = variation that will be ignored; p = patient, r = rater, pr,e = residual error; ICC = intraclass correlation coefficient, SEM = standard error of measurement, σ^2 = variance component. Subscript 1 refers to measurement condition 1 (e.g. measured by raters A and B); subscript 2 refers to measurement condition 2 (e.g. measured by raters C and D) (number of measurement conditions can be extended); * assuming equal sample sizes for each facet across the studies.

When we want to estimate the consistency parameters, and thus choose to ignore the systematic difference between the raters (σ_r^2), we still need to estimate the main effect of raters. Therefore, we have to disentangle the influence of the facet of generalization from the residual error. This is not possible in a fully nested design (see method 1 in Supplementary Figure 1 Appendix 2), because in such a design, the main effect of the raters is part of the residual error ($\sigma_{r,pr,e}^2$). Therefore, there is only one method we can use, i.e. consider the design as a series of separate crossed studies (per measurement condition, e.g. the rater pair) and pool the estimated variance components (see method 2 in Supplementary Figure 1 Appendix 2), ignoring the variance due to main effect of the raters (see Supplementary Figure 2 Appendix 2).

Data collection scheme Design (r : p)			Venn diagram	Statistical formulas Using two two-way random effects model*, and subsequently average de variance components:
Patient	Rater			
1	a	b		$ICC_{\text{consistency}} = \frac{\frac{(\sigma_{p1}^2 + \sigma_{p2}^2)}{2}}{\frac{(\sigma_{p1}^2 + \sigma_{p2}^2)}{2} + \frac{(\sigma_{pr,e1}^2 + \sigma_{pr,e2}^2)}{2}}$ $SEM_{\text{consistency}} = \sqrt{\left[\frac{(\sigma_{pr,e1}^2 + \sigma_{pr,e2}^2)}{2} \right]}$
...	a	b		
10	a	b		
11	c	d		
...	c	d		
n	c	d		

Figure 2 Appendix 2. Two-way mixed effects model for consistency (ICC (3.1)) in a nested design. n = total number of included patients; A, B, C, D = refer to a specific rater; Blue surface = variation of interest, white surface = measurement error of interest; red surface = variation that will be ignored; p = patient, r = rater, p,e = residual error; ICC = intraclass correlation coefficient, SEM = standard error of measurement, σ^2 = variance component. Subscript 1 refers to measurement condition 1 (e.g. measured by raters A and B); subscript 2 refers to measurement condition 2 (e.g. measured by raters C and D) (number of measurement conditions can be extended); * assuming equal sample sizes for each facet across the studies.

Three-way designs

Nested design with patients nested in technicians crossed with raters

Suppose now that we have two technicians per hospital who acquire the images, so four technicians all together. These technicians only measure the patients from their own hospital, so the technicians are nested in the patients. All sets of images are subsequently scored by the two raters involved in the study (Supplementary Figure 3 Appendix 2 data collection scheme). The raters are crossed with the patients. As we are interested in all sources of error, we are building the ICC_{agreement} and SEM_{agreement}. There are two methods to build the appropriate ICC or SEM (Supplementary Figure 3 Appendix 2 statistical method). Again, both methods will provide slightly different results, as the variance components are estimated with different effects models.

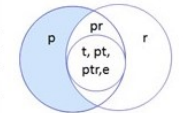
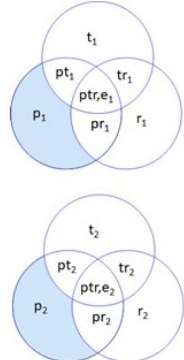
Data collection scheme					Venn diagram	Statistical formulas	
Design ((p : t) x r)							
	Technician 1 or 3		Technician 2 or 4				
Patient	Rater A	Rater b	Rater a	Rater b			
1	1a	1b	2a	2b	<p>Method 1:</p>  <p>Methods 2:</p> 		
...	1a	1b	2a	2b			
10	1a	1b	2a	2b			
11	3a	3b	4a	4b			
...	3a	3b	4a	4b			
n	3a	3b	4a	4b			

Figure 3 Appendix 2. Three-way random effects model for agreement in a nested design.

n = total number of included patients; A, B, C, D = refer to a specific rater; Blue surface = variation of interest, white surface = measurement error of interest; p = patient t = technician, r = rater, pt = interaction between patient and technician, pr = interaction between patient and rater, tr = interaction between technician and rater, ptr,e = residual error; ICC = intraclass correlation coefficient, SEM = standard error of measurement, σ^2 = variance component; Subscript 1 refers to

measurement condition 1 (e.g. measured by technician 1 and 2); subscript 2 refers to measurement condition 2 (e.g. measured by technician 3 and 4) (number of measurement conditions can be extended); * assuming equal sample sizes for each facet across the studies.

References

1. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996;1:30-46.
2. Govaerts MJ, van der Vleuten CP, Schuwirth LW. Optimising the reproducibility of a performance-based assessment test in midwifery education. *Adv Health Sci Educ Theory Pract*. 2002;7(2):133-45. doi:10.1023/a:1015720302925

Appendix 3. Model specifications of ICCs and SEMs and the Agree package for R

The intraclass correlation coefficient (ICC) and standard error of measurement (SEM) can be calculated for continuous scores. Multiple statistical models can be used to analyze reliability and measurement error. Often used models are the one-way random effects model, the two-way random effects model for agreement and the two-way mixed effects model for consistency. Also three-way effects models are possible. The research question together with the corresponding design of the study determine the best statistical model to analyze the data ¹. The Agree package is developed to calculate the reliability and measurement error for the scores of multiple raters or repeated measurements in stable patients ². The `varcomp()` function from this package uses the linear mixed effects model approach, which can deal with the missing data. This model is estimated with the `lmer()` function from the `lme4` package ³. In this Appendix, we will first explain these statistical models, and subsequently, we show the R functions from the Agree package, that can be used to obtain the variance components from each of these statistical models to compute the ICC and SEM.

Statistical models

In the design of the one-way random effects model the raters are unknown, so the effect of raters is not present in the model. This model is specified in Equation 1:

$$x_{ij} = \beta_0 + a_{0i} + e_{ij} \quad (1)$$

$$a_{0i} \sim N(0, \sigma_{a0}^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where x_{ij} denotes the score for measurement j of patient i , β_0 denotes the overall population mean of the measurements, a_{0i} denotes the random patient effect with a mean of 0 and a variance of σ_{a0}^2

and e_{ij} denotes the residual variance in the observed scores with a mean of 0 and a variance of σ_e^2 .

In this model the observed scores are only explained by the differences between patients.

In the design of the two-way random effects model of agreement and the two-way mixed effects model of consistency the raters are known, so these effects are present in the models. In the two-way random effects model of agreement an additional random effect is added for the raters, as presented in Equation 2.

$$x_{ij} = \beta_0 + a_{0i} + c_{0j} + e_{ij} \quad (2)$$

$$a_{0i} \sim N(0, \sigma_{a0}^2)$$

$$c_{0j} \sim N(0, \sigma_{c0}^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where x_{ij} denotes the score for measurement j of patient i , β_0 denotes the overall population mean of the measurements, a_{0i} denotes the random patient effect with a mean of 0 and a variance of σ_{a0}^2 , c_{0j} denotes the random rater effect with a mean of 0 and a variance of σ_{c0}^2 and e_{ij} denotes the residual variance in the observed scores with a mean of 0 and a variance of σ_e^2 . This model accounts for systematic differences between raters represented in the random effect of the raters.

In the design of the two-way mixed effects model of consistency the effect for raters is considered as fixed, so the systematic differences between raters are not taken into account. The two-way mixed effects model is presented in Equation 3:

$$x_{ij} = \beta_0 + a_{0i} + c_1 + e_{ij} \quad (3)$$

$$a_{0i} \sim N(0, \sigma_{0a}^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where x_{ij} denotes the score for measurement j of patient i , β_0 denotes the overall population mean of the measurements, a_{0i} denotes the random patient effect with a mean of 0 and a variance of $\sigma_{a_0}^2$, c_1 denotes the fixed rater effect (so this effect does not vary between raters as opposed to the rater effect (c_{0j}) specified in Equation 2) and e_{ij} denotes the residual variance in the observed scores with a mean of 0 and a variance of σ_e^2 .

The three-way effects models are an extension of the two-way effects models with an extra random (d_{0k}) or fixed effect (d_1) (e.g. for technician). In case of a random effect, it has a mean of 0 and a variance of $\sigma_{d_0}^2$.

The R code to estimate ICC and SEM

The package can be installed directly from GitHub by:

```
remotes::install_github(repo = 'iriseekhout/Agree')
```

ICC and SEM one-way effect model (Figure 8A):

The ICC and SEM can be directly obtained from a data set in a wide format using:

```
icc_oneway(data)
```

This function returns the ICC with the 95% confidence interval calculated with the exact F method ⁴, the SEM as well as the estimated variance components. Another possibility is to use the `varcomp()` function to obtain the variance components for the one-way model from the data set in long format:

```
varcomp(score ~ (1|id), data)
```

The `varcomp()` function returns the variance components in a data frame and these can be used to calculate the ICC and SEM respectively:

```
vc["id","vcov"] / (vc["id","vcov"] + vc["Residual","vcov"])  
sqrt(vc["Residual","vcov"])
```

ICC and SEM two-way random effects model for agreement (Figure 8B):

The ICC and SEM can be directly obtained from a data set in a wide format using:

```
icc_agreement(data)
```

This function returns the ICC with the 95% confidence interval approximated with the F method ^{5,6}, the SEM as well as the estimated variance components. Another possibility is to use the varcomp() function to obtain the variance components for the two-way model from the data set in long format:

```
varcomp(score ~ (1|id)+ (1|rater), data)
```

The varcomp() function returns the variance components in a data frame and these can be used to calculate the ICC and SEM respectively:

```
vc["id","vcov"]/ (vc["id","vcov"]+ vc["rater","vcov"]+  
vc["Residual","vcov"])  
sqrt(vc["Residual","vcov"])
```

ICC and SEM two-way mixed effects model for consistency (Figure 8C):

The ICC and SEM can be directly obtained from a data set in a wide format using:

```
icc_consistency(data)
```

This function returns the ICC with the 95% confidence interval calculated with the exact F method ⁴, the SEM as well as the estimated variance components. Another possibility is to use the varcomp() function to obtain the variance components for the mixed two-way model from the data set in long format with a fixed effect for rater:

```
varcomp(score ~ rater + (1|id), data)
```

The varcomp() function returns the variance components in a data.frame and these can be used to calculate the ICC and SEM respectively:

```
vc["id","vcov"]/ (vc["id","vcov"]+ vc["Residual","vcov"])  
sqrt(vc["Residual","vcov"])
```

ICC and SEM three-way random effects model for agreement (Figure 8E):

The `varcomp()` function can be used to obtain the variance components for the three-way model from the data set in long format:

```
varcomp(score ~ (1|id)+ (1|rater)+ (1|technician), data)
```

The `varcomp()` function returns the variance components in a data frame and these can be used to calculate the ICC and SEM respectively:

```
vc["id","vcov"]/ (vc["id","vcov"]+ vc["rater","vcov"]+  
vc["technician","vcov"]+ vc["Residual","vcov"])  
sqrt(vc["Residual","vcov"])
```

ICC and SEM three-way mixed effects model for consistency (Figure 8F):

The `varcomp()` function can be used to obtain the variance components for the three-way mixed effect model from the data set in long format with a fixed effects for rater and technician:

```
varcomp(score ~ technician + rater +(1|id), data)
```

The `varcomp()` function returns the variance components in a data frame and these can be used to calculate the ICC and SEM respectively:

```
vc["id","vcov"]/ (vc["id","vcov"]+ vc["Residual","vcov"])  
sqrt(vc["Residual","vcov"])
```

ICC and SEM three-way mixed effects model (Figure 8G):

The `varcomp()` function can be used to obtain the variance components for the three-way mixed effect model from the data set in long format with a fixed effect for technician:

```
varcomp(score ~ technician +(1|id)+ (1|rater), data)
```

The `varcomp()` function returns the variance components in a data frame and these can be used to calculate the ICC and SEM respectively:

```
vc["id","vcov"]/ (vc["id","vcov"]+ vc["rater","vcov"]+  
vc["Residual","vcov"])  
sqrt(vc["Residual","vcov"])
```

References

1. Mokkink LB, Eekhout I, Boers M, van der Vleuten CP, De Vet HC. Studies on reliability and measurement error in medicine – from design to statistics explained for medical researchers. *Patient Related Outcome Measures*. 2023.
2. Eekhout I. Agree: Agreement and reliability between multiple raters. R package version 0.1.8. Accessed 8 March 2022, Accessed 8 March 2022 <https://github.com/iriseekhout/Agree/>
3. Bates D, Machler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 2015;67:1-48. doi:<http://dx.doi.org/10.18637/jss.v067.i01>
4. Shrout PE, Fleiss JL. Intraclass Correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979;86:420-428.
5. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics*. 1946;2:110-114.
6. Fleiss JL, Shrout PE. Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika*. 1978;43(2):259-262. doi:<https://doi.org/10.1007/BF02293867>

Appendix 4. SPSS syntax to estimate the variance components for a three-way effects model

SPSS can be used to estimate the variance components in crossed and nested designs.

In this appendix we explain how to estimate the variance components for three-way effect models. With the variance components you can manually calculate the ICC or SEM. SPSS syntaxes can be found [here](#)¹.

For the one-way and two-way effects models another method, which also provides the 95% confidence intervals can be used. An explanation can be found [here](#)²:

Data structure for crossed design (three-way effects models)

Data are structured in a 'long' datafile:

Patients	Technician	Rater	Scores
1	1	1	
1	1	2	
1	2	1	
1	2	2	
...			

Commands in SPSS for three-way random effects models

Analyze: 'General Linear Model': option 'varcomps'

Dependent variable: include 'Scores'

Random factors: include the facet of differentiation (patients)
and all facets of generalization: e.g. technician, rater

If one chooses option 'fixed' for any of the facets of generalization (raters) the main effects for this facet σ_r^2 will not be calculated. We recommend to consider all facets to be random, and use the required variance components in the formula

Go to 'model': choose 'build terms' under 'specify model'

Include all main effects: 'patients' and 'rater' as *main effects* in a two-way design

'patients', 'technician' and 'rater' as *main effects* in a three-way design

Include all interactions: 'patients x rater' as *interaction* in a two-way design

'patient x technician', 'patient x rater', and 'technician x rater' as *interactions* in a three-way design

The interaction between 'patient x technician x rater' need not to be appointed, as the random error, which is included in this interaction will be estimated by default.

Choose for the option 'include the intercept'.

Press: Continue

Go to 'options': select 'restricted maximum likelihood' under 'method'

Commands in SPSS for three-way mixed effects models

If you would consider one of the facets of generalization to be fixed, than in the first step this facets is considered to be fixed. The main effect of this facet is not estimated, while all interactions with this facets are.

Data structure for nested three-way design (technician nested in patient, raters crossed with patients). See also design Appendix 2 Figure 3.

Data are structured in a 'long' datafile:

Patients	Technician	Rater	Scores
1	1	1	
1	1	2	
1	2	1	
1	2	2	
1	-	-	
1	-	-	
1	-	-	
1	-	-	
...			
20	-	-	
20	-	-	
20	-	-	
20	-	-	
20	3	1	
20	3	2	
20	4	1	
20	4	2	
...			

Commands in SPSS

Method 1, technician considered in the residual error

Analyze: 'General Linear Model': option 'varcomps'

Dependent variable: include 'Scores'

Random factors: include the facet of differentiation (patients)
the crossed facets of generalization: here: rater

The facet that is nested (here: technician) is actually ignored.

Go to 'model': choose 'build terms' under 'specify model'

Include all main effects: 'patients' and 'rater' as *main effects* in a two-way design

Include the interactions: 'patients x rater' as *interaction* in a two-way design

Choose for the option 'include the intercept'.

Press: Continue

Go to 'options': select 'restricted maximum likelihood' under 'method'

Commands in SPSS

Method 2, variance components are estimated per measurement condition

This can be done when the sample size of each measurement condition is equal.

Go to 'data' and to 'select cases'

Click on 'Condition is satisfied if'

Describe the first measurement condition, e.g. 'technician = 1 or technician = 2'

And next the same random effects model can be conducted as for a crossed three-way random effects model.

This step is repeated for each measurement condition

Last, the variance components for each facet from each measurement condition are manually pooled. In case of equal sample sizes, this is by taking the average across the variance components.

References

1. Mokkink LB, de Vet HCW, Eekhout I. SPSS syntax for estimation of variance components in crossed and nested designs. <https://www.cosmin.nl/wp-content/uploads/SPSS-syntax-for-estimation-of-variance-components-in-crossed-and-nested-designs-1.pdf>
2. Vet HCW. Guide for the calculation of ICC in SPSS. http://www.clinimetrics.nl/images/upload/files/Chapter%205/chapter%205_5_Calculation%20of%20ICC%20in%20SPSS.pdf