



Sample size recommendations for studies on reliability and measurement error: an online application based on simulation studies

Lidwine B. Mokkink^{1,2} · Henrica de Vet^{1,2} · Susanne Diemeer¹ · Iris Eekhout^{1,2,3}

Received: 30 June 2022 / Revised: 7 October 2022 / Accepted: 30 October 2022 /

Published online: 23 November 2022

© The Author(s) 2022

Abstract

Simulation studies were performed to investigate for which conditions of sample size of patients (n) and number of repeated measurements (k) (e.g., raters) the optimal (i.e., balance between precise and efficient) estimations of intraclass correlation coefficients (ICCs) and standard error of measurements (SEMs) can be achieved. Subsequently, we developed an online application that shows the implications for decisions about sample sizes in reliability studies. We simulated scores for repeated measurements of patients, based on different conditions of n , k , the correlation between scores on repeated measurements (r), the variance between patients' test scores (v), and the presence of systematic differences within k . The performance of the reliability parameters (based on one-way and two-way effects models) was determined by the calculation of bias, mean squared error (MSE), and coverage and width of the confidence intervals (CI). We showed that the gain in precision (i.e., largest change in MSE) of the ICC and SEM parameters diminishes at larger values of n or k . Next, we showed that the correlation and the presence of systematic differences have most influence on the MSE values, the coverage and the CI width. This influence differed between the models. As measurements can be expensive and burdensome for patients and professionals, we recommend to use an efficient design, in terms of the sample size and number of repeated measurements to come to precise ICC and SEM estimates. Utilizing the results, a user-friendly online application is developed to decide upon the optimal design, as 'one size fits all' doesn't hold.

Keywords Sample size recommendations · Simulation study · Reliability · Measurement error · Repeated measurements · Outcome measurement instruments

✉ Lidwine B. Mokkink
w.mokkink@amsterdamumc.nl

¹ Department of Epidemiology and Data Science, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

² Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

³ Child Health, Netherlands Organisation for Applied Scientific Research, Leiden, The Netherlands

1 Background

In clinical trials conclusions are drawn based on outcome measurement scores. These scores are measured with measurement instruments, such as clinician-reported outcome measures, imaging modalities, laboratory tests, performance-based tests, or patient-reported outcome measures (PROMs) (Walton et al. 2015). The validity (or reliability) of trial conclusions depend, among other things, on the quality of the outcome measurement instruments. High quality measurement instruments are valid, reliable, and responsive to measure the outcome of interest in the specific patient population.

Reliability and measurement error are two related but distinct measurement properties that can be investigated within the same study design (using the same data). Measurement error refers to how close the results of the repeated measurements are. It refers to the absolute deviation of the scores, or the amount of error, of repeated measurements in stable patients (de Vet et al. 2006), and is expressed in the unit of measurement such as the standard error of measurement (SEM) (de Vet et al. 2006; Streiner and Norman 2008). Reliability relates the measurement error to the variation of the population. Therefore, reliability refers to whether and to what extent an instrument is able to distinguish between patients (de Vet et al. 2006). For continuous scores, reliability is expressed as an intraclass correlation coefficient (ICC), a relative parameter.

In a study on reliability, we are interested in the influence of specific sources of variation such as rater, occasion or equipment, on the score (Mokkink et al. 2022). This specific source of variation of interest (e.g., rater) is varied across the repeated measurements in stable patients. For example, we are interested in the influence of different raters (i.e., the source of variation that is varied across the repeated measurements) on one occasion (inter-rater reliability), or in the influence of different occasions (i.e., the source of variation that is varied across the repeated measurements) by one rater on the score of stable patients (i.e., intra-rater reliability); or in the influence of the occasion on the score when stable patients rate themselves on different occasions with a self-administered questionnaire (i.e., test–retest reliability). In the remainder of this paper, the term ‘repeated measurements’ refers to repeated measurements in stable patients, ‘different raters’ will be used as example of the source of variation of interest, and the term ‘patients’ will be used to refer to the ‘subjects of interest’.

Multiple statistical models can be used to estimate ICCs and SEM. 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 (see Table 1 and “Appendix 1” for model specifications of ICCs and SEMs). Three-way effects models are outside the scope of this paper. The research question together with the corresponding design of the study determine the appropriate statistical model to analyze the data (Mokkink et al. 2022).

Table 1 Overview of one-way and two-way effects models

	Multi level model ^a	ICC	SEM
One-way effects model	$x_{ij} = \beta^0 + a_{0i} + e_{ij}$	$\sigma_j^2 / (\sigma_j^2 + \sigma_e^2)$	$\sqrt{\sigma_e^2}$
Two-way random effects model for agreement	$x_{ij} = \beta_0 + a_{0i} + c_{0j} + e_{ij}$	$\sigma_j^2 / (\sigma_j^2 + \sigma_k^2 + \sigma_e^2)$	$\sqrt{\sigma_r^2 + \sigma_e^2}$
Two-way mixed effects model for consistency	$x_{ij} = \beta_0 + a_{0i} + c_1 + e_{ij}$	$\sigma_j^2 / (\sigma_j^2 + \sigma_e^2)$	$\sqrt{\sigma_e^2}$

^aFull details are available in “Appendix 1”

1.1 Sample size recommendations for studies on reliability and measurement error

In the literature two different approaches are used for calculations of required sample sizes, i.e., an analytical approach (hypothesis testing) and simulation studies. Most previous studies used an analytical approach for a one-way random effects model (Bland 2004; Donner and Eliasziw 1987; Giraudeau and Mary 2001; Walter et al. 1998). In an analytical approach, a null hypothesis is formulated (H_0 is $\rho = \rho_0$), and it is tested whether the observed ICC (i.e., ρ) is similar to a predefined true ρ_0 , against the H_1 that is $\rho > \rho_0$ (Donner and Eliasziw 1987; Giraudeau and Mary 2001; Walter et al. 1998). In this approach, the expected ICC value should be chosen, which is difficult and questionable (Giraudeau and Mary 2001). Furthermore, to use the sample size recommendations derived with an analytical approach, a complete understanding of complex formulas is required. Moreover, currently existing formulas are limited to one-way effects models, and appropriate formulas for two-way effects models are lacking. Recommendations based on a one-way effects model, are generally conservative for situations where a two-way effects model is used (5, 7), because in a two-way model the patient variation is estimated with more precision by disentangling variance from other sources from the scores. Accordingly, efficiency can be gained when sample size recommendations are based on the (chosen) design and appropriate model used for the analysis.

Moreover, most studies focused on sample size recommendation to assess reliability, and only few focused on measurement error. Bland (2004) and Lu et al. (2016) provide sample size recommendations for studies on measurement error using limits of agreement. The SEM calculated from the limits of agreement is similar to the SEM derived from a two-way mixed model for consistency, which ignores the systematic difference of the source of variation that was varied across the measurements (e.g., the raters) (de Vet et al. 2011). While these studies provide useful recommendations for studying measurement errors with a consistency model, they may not apply to other statistical models, such as measurement error with an agreement model.

Another approach to obtain recommendations for sample size is based on simulation studies (Saito et al. 2006; Zou 2012). Simulation studies can show the effects of specific conditions of study designs (such as more raters or more patients) on the estimation of the parameters (i.e., ICC and SEM) in terms of precision and bias. In these studies the choice of conditions is crucial, as the results cannot be generalized beyond the investigated conditions.

In the current study, we focus on the compromise between precision of the ICC and SEM estimations and feasibility in a study to obtain the most efficient recommendations for sample size and repeated measurements using simulation studies. We performed a series of simulation studies based on realistic but artificial data to investigate the precision of various reliability and measurement error parameters under different conditions within different study designs. We aim to study the most efficient combination of sample size of patients (n) and number of repeated measurements (k) (e.g., raters), given the study design at hand. In a freely available online application (i.e., the Sample size decision assistant) we utilize our study findings in order to assist researchers in designing a reliability study, available at <https://iriseekhout.shinyapps.io/ICCPower/>.

2 Methods

2.1 Simulation studies

Artificial data samples were simulated under various conditions. These conditions were seen in various realistic data sets (Dikmans et al. 2017; Mosmuller et al. 2016; Mulder et al. 2018). Samples were generated with different conditions of the sample size of patients (n), number of repeated measurements (k), correlations between the scores on the repeated measurements (r), and variances between patients' test scores (v).

The investigated conditions for n and k represent feasible and realistic conditions in clinical studies (see Table 2 for the chosen conditions). The correlation r is equivalent to the ICC when no systematic difference between the measurements occur. The conditions $r=0.6$, 0.7 and 0.8 were used, because the consensus-based cut-off point for sufficient reliability is at an ICC value of 0.7 (Nunnally and Bernstein 1994). An ICC value of 0.6 refers to insufficient reliability and an ICC value of 0.8 is well above the cut-off point. The variance between patients' test scores (v) indicates to what extent the test scores vary between patients, i.e., it refers to the range of distribution of the scores. The variance between patients' test scores (v) was simulated as 1 , 10 and 100 ; specified as small, medium and large, respectively. Consequently, a vector with a mean of 0 has different ranges between these conditions; $v=1$ ranges from -3 to 3 ; $v=10$ ranges from -9.5 to 9.5 ; and $v=100$ ranges from -30 to 30 . The simulated data were sampled from a multivariate normal distribution with a mean of 0 and the covariance matrix (i.e., r multiplied by v).

To investigate the impact of different statistical models (i.e., one-way or two-way effects models), we introduced systematic differences between the repeated measurements. This way we gradually moved from a one-way design to a two-way design. To incorporate a systematic difference, the scores of one (or two) of the raters was systematically changed by increasing the average score of this rater with 1 standard deviation in score in the respective variance conditions of 1 , 10 and 100 (i.e., standard deviation is 1 , 3 , or 10 points). In the conditions with 4 or more repeated measurements, we additionally investigated the effect of two deviating repeated measures by increasing the average scores for the first two raters with the same amounts (i.e., 1 , 3 , and 10 points).

Table 2 Overview of conditions and their values that will differ in the simulation studies

Condition	Value
Sample size of patients (n)	10, 20, 30, 40, 50, 100, 200
Number of repeated measurements (k)	2, 3, 4, 5, 6
Correlation of scores on repeated measurements (r)	0.6, 0.7, 0.8
Variance between patients' test scores (v)	1, 10, 100
Type of statistical model	ICC one-way random effects model, ICC two-way random effects model, ICC two-way mixed effects model
Systematic difference of 1 standard deviation	No systematic difference or one systematically different rater (when $k=2-6$); two systematically different raters (when $k=4-6$)

The combinations of these conditions, i.e., sample size (n), number of raters (k), variance between patients' test scores (v) and correlation (r), resulted in a total of 360 combination of conditions when no systematic difference was incorporated; 360 combinations when one rater systematically differed (condition $k=2-6$); and in a total of 216 combinations of conditions when two raters systematically differed (i.e., for $k=4-6$). For each combination of conditions 1000 samples were generated using the R package MASS in R statistical software (Venables and Ripley 2002).

2.1.1 Reference values

Additionally, we simulated population data of 100,000 people for each combination of correlation (r ; i.e., 0.6–0.8), variance (v ; 1, 10 and 100) and number of repeated measurements (k ; 2–6) to obtain reference values. We choose 100,000 as this size is sufficient enough to eliminate sampling error effect, while the models would still converge. Also in these populations, we incorporated systematic differences between repeated measurements as described above, i.e., scores of either one rater (when $k=2-6$) or two raters (when $k=4-6$) were systematically changed.

2.1.2 Estimation of the reliability parameters

Using the Agree package in R (Eekhout 2022; Eekhout and Mokkink 2022) for each of the $2 \times 360 \times 1000$ and 216×1000 simulated samples we computed: the three types of ICC's [i.e., based on (1) one-way random effects model, (2) two-way random effects model for agreement, and (3) two-way mixed effects model for consistency]; 95% confidence intervals (CI) of each ICC, and the corresponding three types of SEM (see "Appendix 1" for model specifications and for R syntaxes). The same parameters were calculated on the population data.

2.1.3 Evaluation of the performance parameters for the estimations

The performance of the reliability parameters was evaluated by the calculation of the bias, the mean squared error (MSE), and the coverage of the confidence intervals (Burton et al. 2006). First, *sample bias* was defined in each of the 1000 samples per condition as the difference between the sample estimates for each parameter (i.e., ICCs and SEMs) and the reference value for each parameter for that condition (i.e., based on the population data). Next, these 1000 sample biases (per combination of conditions) were averaged, which results in the *bias* for each condition (Burton et al. 2006; Eekhout et al. 2015). The bias is expressed in the 'metric' of the parameter (i.e., ICC or SEM). A negative bias means an underestimation of the true ICC (i.e., the population ICC). Squaring each sample bias (per condition) and averaging these squared sample biases over the 1000 samples give the MSE per combination of conditions (Burton et al. 2006; Eekhout et al. 2015). The MSE provides a measure of the overall precision of the estimated parameters (Burton et al. 2006), and the square root of the MSE value transforms the MSE back into the same 'metric' of the parameter (i.e., ICC or SEM) (Burton et al. 2006). The smallest possible MSE value is zero, meaning that the mean of the estimated parameter in all samples for the specific condition has the same value as the population parameter. Additionally, we expressed the MSE results in terms of the width of the confidence interval per condition, as the width of the CI is often used in analytical approaches for sample size calculations for reliability studies as a

measure of precision. The width was computed from the MSE as follows: $\text{width} = 2 * (1.96 * \sqrt{\text{MSE}})$. The SEM, and thus also the bias, and the width of its confidence interval, is expressed in the unit of measurement. As this unit of measurement changes due to the variance (v) condition in our study, the magnitude of bias and MSE also increase with this variance condition by definition. For that reasons we will only evaluate the $v = 1$ condition for the SEM.

The *coverage of the confidence interval* of the estimated ICCs was calculated as a percentage of the number of times the population value lies within the estimated 95% confidence interval of the ICC parameters for the 1000 samples in each combination of conditions (Burton et al. 2006). By definition, the coverage should be 95% for the 95% confidence interval (Burton et al. 2006).

2.2 Deciding on the sample size of number of patients and repeated measurements in future studies: the online Sample size decision assistant

In an online application (i.e., the Sample size decision assistant), the results of the simulation study are used to inform the choice on sample size (of patients) and number of repeated measurements (e.g., raters). Recommendations for the choice on sample size and number of repeated measurements are based on three different procedures, i.e., the width of the CI, the lower limit of the CI, and the MSE ratio.

The CI width procedure can be used when designing the study, i.e., before the start of the data collection, to determine the precision of the estimations of both reliability parameters (ICC and SEM) in the target design. This procedure uses the results of these simulations studies. In the CI width procedure a pre-specified width of the confidence interval (e.g., 0.3) is set to determine what conditions of sample size and repeated measurements can achieve that specific CI width under the selected design conditions. This way, various designs can be considered to decide on the most efficient target design. This chosen target design is the design as described in the study protocol.

The CI lower limit procedure is based on an analytical approach (Zou 2012), and can be used to do recommendations for $\text{ICC}_{\text{one-way}}$ only. It can be used when designing the study, i.e., before the start of the data collection, to determine the precision of the estimations of the reliability parameter in the target design. The CI lower limit procedure is a known method in the literature and uses a formula for the confidence intervals presented in Zou (2012) to estimate the sample size required given the assumed $\text{ICC}_{\text{one-way}}$, lower limit of the $\text{ICC}_{\text{one-way}}$ and the number of raters that will be involved. The advantage of this method is that it can be used beyond the specified conditions that are used in our simulation studies. However, as this method is based on the $\text{ICC}_{\text{one-way}}$, results cannot be generalized to the other types of ICCs. We used this formula-based method during our analyses to compare the results and recommendations based on CI width procedure and the MSE ratio procedure.

The MSE ratio procedure can be used when the data collection has started and there is a need to change or reconsider the target design of the study, e.g., the patient recruitment is slow or one of the raters drops out. The MSE ratio procedure uses the results of these simulations studies and focus on the precision of the estimations; it can be used to do recommendations for ICC and SEM. The MSE ratio can be calculated as $\text{MSE}_{\text{target}}/\text{MSE}_{\text{adapted}}$ (Eekhout et al. 2015). These conditions differ on one variable, either the sample size (n) or the number of repeated measurements (k). The MSE ratio procedure can be used in two ways: (1) it informs how much the precision of the adapted design (i.e., a new design,

such as the data collected so far) deviates from the precision of an target design chosen at the start of the study (i.e., as described in the protocol of the study). Or (2) it represent the proportional increase of the sample size (n) or number of repeated measurements (k) in the adapted design that is required to achieve the same level of precision as in the target design.

3 Results

In this result section, the effect of study design conditions on the estimation of the ICC and SEM are shown, in terms of bias and MSE (for ICC and SEM), the coverage of the CI of ICC, and the influence of various conditions on the width of the 95% CI of ICCs and SEMs, averaged over various conditions. For tailored results and recommendations about the sample size and number of repeated measurements for specific conditions, we developed an online application (<https://iriseekhout.shinyapps.io/ICCPower/>) that shows the implications for decisions. Subsequently, we describe the online application, and how to use this to come to tailored recommendations in future studies.

3.1 Bias and MSE values in ICC estimations

Results for bias showed a slight underestimation of the estimated ICCs, especially with small sample sizes. Overall the bias was so small that it was negligible, i.e., maximum bias for the ICCs found in any of the conditions was -0.05 (in case of a sample size of 10 with only 2 raters of which 1 deviated in an ICC one-way random effects model with a v of 1 and r of 0.7).

In Fig. 1 we plotted the MSE of the ICC estimates for the number of raters per the conditions of sample size (shown in different colors), shown for the situation that one rater systematically deviates and for each of the three statistical models separately. Here, we see that the steepness of the curve declines most between $k=2$ and $k=3$, especially for a sample size up to $n=50$. So the gain in precision (i.e., the largest change in MSE) is highest going from 2 to 3 raters. Moreover, we see the distance between the curves decreases when n increases, especially in the curves up to $n=40$. So the gain in precision is relatively smaller above $n=40$. In other words, the gain in precision diminishes at larger values of n and k . The MSE values for condition $n=40$ and $k=4$ is very similarly compared to the condition $n=50$ and $k=3$. As this pattern was seen for all conditions of r , and v , we averaged over these conditions in Fig. 1.

The presence of a systematic difference between raters increased the MSE values for ICC_{one-way}, but not for ICC_{agreement} and ICC_{consistency} (see online tool). This means that the required sample size for the one-way effects models increases when a systematic difference between raters occurs, while the required sample sizes for the two-way effects models remains the same.

Next, we noticed an influence of the correlation between scores on repeated measurements (r) on the MSE values for all types of ICCs, specifically when no rater deviated (Fig. 2 shows the MSE per correlation condition for ICC_{agreement}). That is, increasing correlation (i.e., 0.8 instead of 0.6) leads to decreasing MSE values. When one rater deviates, r continues to affect the MSE for ICC_{consistency} to the same extent, but to a lesser extent for ICC_{agreement} and ICC_{one-way} (“Appendix 2”).

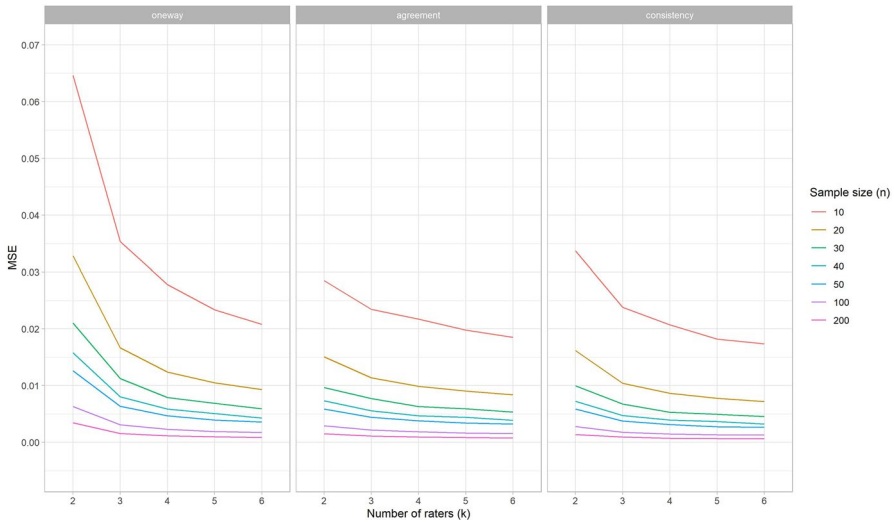


Fig. 1 MSE values of ICC estimations with different sample sizes, plotted against k per type of ICC model (one rater systematically deviates; averaged over all conditions of r and v)

3.2 Bias and MSE in SEM estimations

Overall the bias for the SEM was very small and thus negligible. All results for bias can be found in the online application.

In Fig. 3 we plotted the MSE values of the $SEM_{agreement}$ estimations for the number of raters per condition of sample size (shown in different colors), for one rater with a systematic difference and for each of the three conditions of r. Similar as we saw above for the MSE curves for ICCs, the steepness of the curves declines most between $k=2$ and $k=3$,

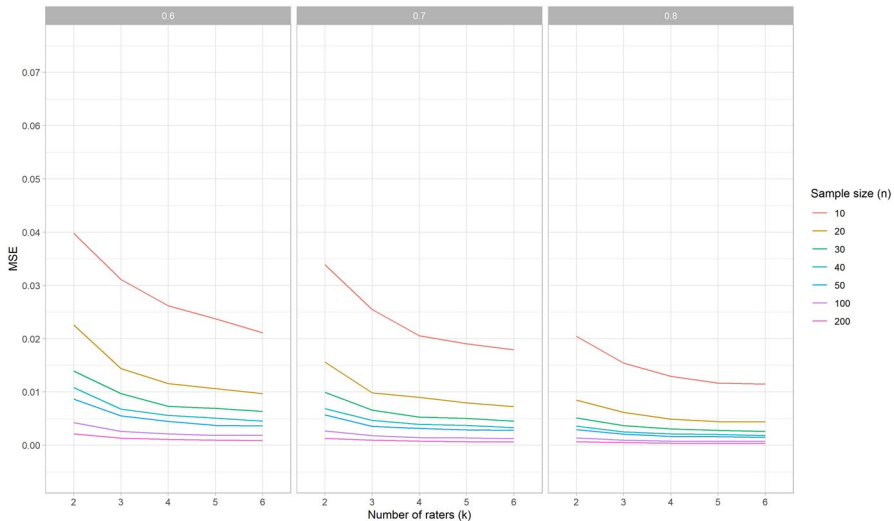


Fig. 2 MSE values of $ICC_{agreement}$ estimations plotted against k per condition r (no rater systematically deviated; averaged over all conditions of v)

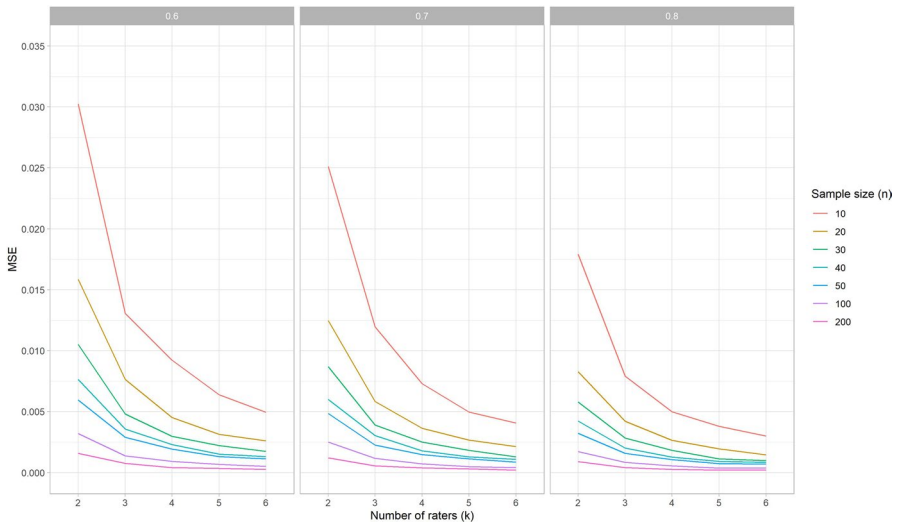


Fig. 3 MSE values of $SEM_{agreement}$ estimations with different sample sizes, plotted against k per condition r (one rater systematic differs; $v=1$)

especially for a sample size up to $n=50$. Moreover, we see the distance between the curves decreases when n increases, especially in the curves up to $n=40$. The MSE values for condition $n=30$ and $k=4$ for any of the three conditions r is very similarly compared to the condition $n=50$ and $k=3$ when $r=0.6$, or $n=40$ and $k=3$ when r is higher.

So, we can conclude that the influence of the correlation r on the MSE value for SEM estimations is similar to the influence of r on the MSE values for ICC estimations.

In both $SEM_{one-way}$ and $SEM_{agreement}$ models all measurement error is taken into account (see “Appendix 1”), so the resulting SEM estimates are equal between these models (Mokkink et al. 2022). The MSE values for $SEM_{consistency}$ are nearly the same if no rater deviates or when one rater deviates. When no rater deviates, the MSE values for the $SEM_{one-way}$ and $SEM_{agreement}$ are only slightly lower compared to the $SEM_{consistency}$ (data available in the online application). However, aberrant from the MSE results for the ICC estimations (see Fig. 1), the MSE values for the $SEM_{one-way}$ and $SEM_{agreement}$ increase when one of the raters systematically deviates (see Fig. 4).

3.3 Coverage of the confidence intervals of ICCs

With no systematic difference between raters the coverage of the 95% confidence intervals around the ICC estimation was as expected, i.e., around the 0.95 for all three types of ICCs. As there were no differences found for the simulation study conditions (i.e., r , v , n and k) the results for coverage are only separated per type of ICC (Fig. 5, left panel).

The coverage of the $ICC_{consistency}$ is very similar when one or two raters deviate, compared to the situation when no rater deviates. However, when one of the raters deviates the lowest coverage of the 95% confidence intervals around the $ICC_{one-way}$ estimation decreases (i.e., under-coverage) and the highest coverage increases (i.e., over-coverage) (Fig. 5, middle panel). While this change in coverage disappears again when two raters deviate (Fig. 5, right panel). Note that in this latter scenario always more than three raters are involved. Furthermore, the $ICC_{agreement}$ showed an over-coverage when one or two raters

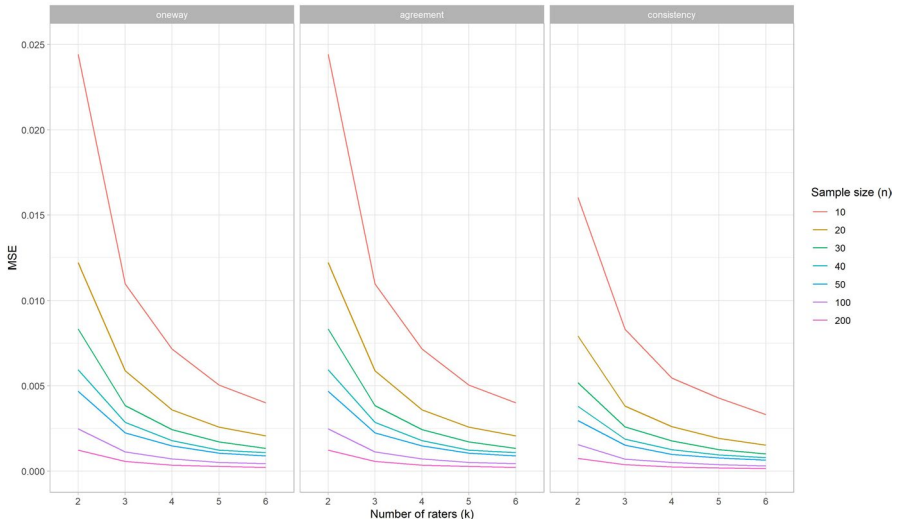


Fig. 4 MSE values of SEM estimations for different sample sizes, plotted against k per type of SEM model (one rater systematically deviates; $v = 1$; averaged over all conditions)

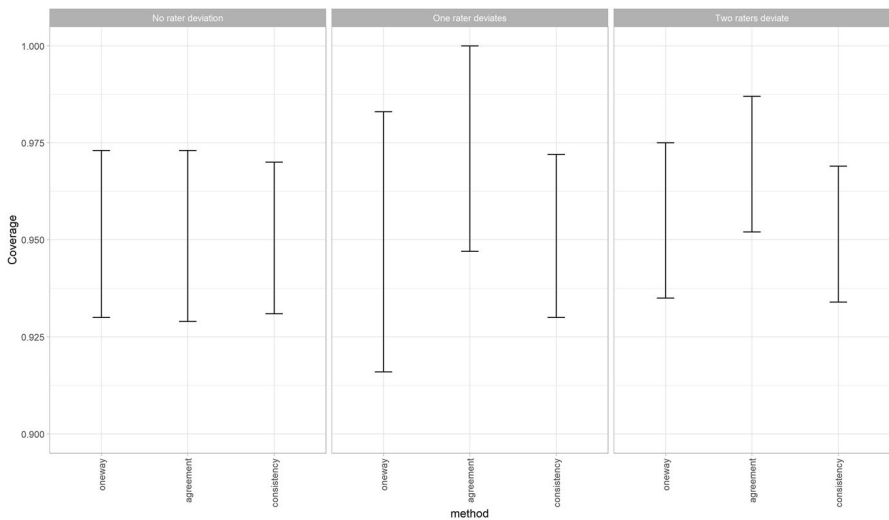


Fig. 5 Lowest and highest coverage of the 95% confidence intervals around the ICC estimations over all conditions of n , k , r and v (left and middle panel $k = 2-6$, right panel $k = 4-6$)

systematically deviated from the other raters, as the lowest value and the highest value for the coverage of the 95% confidence intervals around the $ICC_{agreement}$ both increase (Fig. 5 middle and right panel). A coverage of 1 means that the ICC of the population always fell within the 95% confidence intervals of the ICC estimation. This was due to the fact that the width of the confidence intervals around these estimations were very large, i.e., confidence interval width around 1.

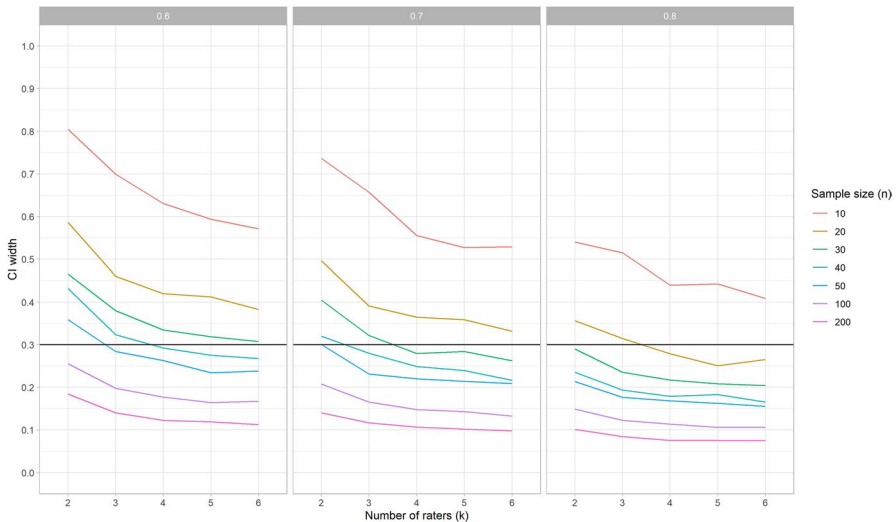


Fig. 6 95% confidence interval width for the ICC for $r=0.6, 0.7$ or 0.8 ($v=1$, no raters deviate; averaged over de three ICC models)

3.4 Influence of various conditions on the width of the 95% confidence intervals of ICCs

When no rater deviates, the 95% CI width around the ICC is the same for the different variances (v) and the different ICC methods (one-way, agreement or consistency). However, the correlation r does impact the width of the 95% CI: an increase of r leads to a decrease of the width (i.e., smaller confidence intervals) (Fig. 6). This means that when we expect the ICC to be 0.7 (i.e., we assume the measurements will be correlated with 0.7) the required sample size will be larger to obtain an ICC with the same precision than when we expect the ICC to be 0.8.

When one rater deviates, the width of the 95% CI does not change for the $ICC_{consistency}$, but it does increase for $ICC_{agreement}$, and even more for $ICC_{one-way}$ (see Fig. 7).

The 95% CI width around the ICC estimation for specific conditions can be used to determine what the optimal trade-off is for the sample size of patients and the number of repeated measures in these situations. In Fig. 6 (where we show results averaged over the three effects models) we can see that in the situation that no rater deviates, and $v=1$, and we wish to estimate an ICC for three raters ($k=3$), we need between 40 and 50 patients to obtain a CI width around the point estimation of 0.3 when $r=0.6$ (i.e., ± 0.15) (Fig. 6, left panel). If r is 0.7, then 30 patients is enough to reach the same precision (Fig. 6, middle panel), while if $r=0.8$ 20 patients is sufficient (Fig. 6, right panel). When one of the raters deviates, the chosen ICC method impacts the 95% CI width, in addition to the r (Fig. 7). To come to a 95% CI width of 0.3 around the point estimate when $r=0.8$, $v=1$, for a $ICC_{agreement}$ the sample size should be increased to 40, while the $ICC_{one-way}$ would require a sample size of 50.

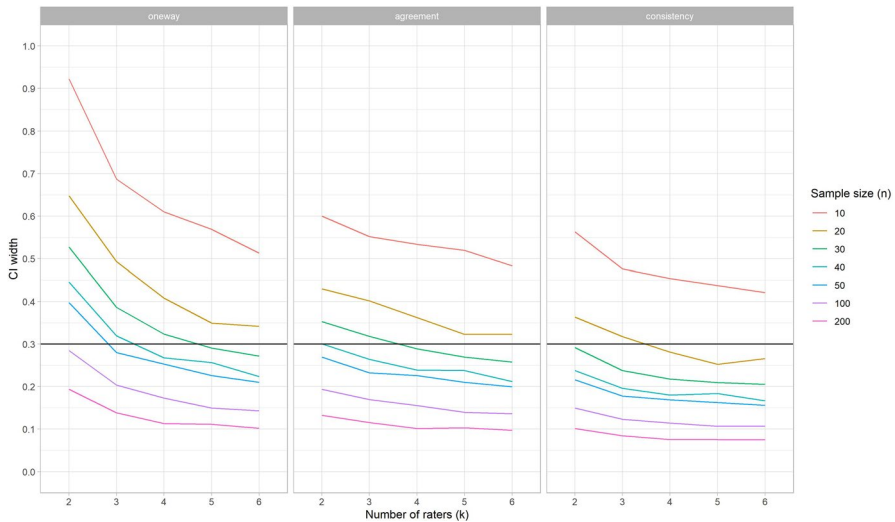


Fig. 7 95% confidence interval width for the three ICC models, when one rater deviates ($r=0.8$, $v=1$)

3.5 Influence of various conditions on the width of the 95% confidence intervals of SEMs

The CI width for SEM estimation decreases when r increases (Fig. 8), similar as for ICC. However, in general, the width for SEM was smaller than for ICCs (Fig. 6).

When one rater deviates, the width of the 95% CI does not change for $SEM_{consistency}$, but it does increase for $SEM_{agreement}$ and $SEM_{one-way}$ (see Fig. 9). In general, the width of the 95% CI is lower for SEM than it is for ICC. This means that in general, the SEM can be estimated with more precision than the ICC under the same conditions.

3.6 Online application that shows the implications for decisions about the sample sizes in reliability studies

As shown in the results of our simulation study, sample size recommendations are dependent on the specific conditions of the study design at hand. Therefore, based on these simulation studies, we have created a Sample size decision assistant that is freely available as an online application to inform the choice about the sample size and number of repeated measurements in a reliability study.

The Sample size decision assistant shows the implications of decisions about the study design on the power of the study, by using any of the three procedures described in the methods section (i.e., the width of the confidence interval (CI width) procedure, the CI lower limit procedure, and the MSE ratio procedure). Each procedure requires some assumptions about the study design as input, as described in Table 3. When you choose to use either the CI lower limit procedure or the MSE ratio procedure, you are asked to indicate what the target design is. The target design is the intended sample size of patients or the number of repeated measurements (e.g., raters), decided upon at the start of the study. For the MSE ratio procedure you are also asked to indicate the

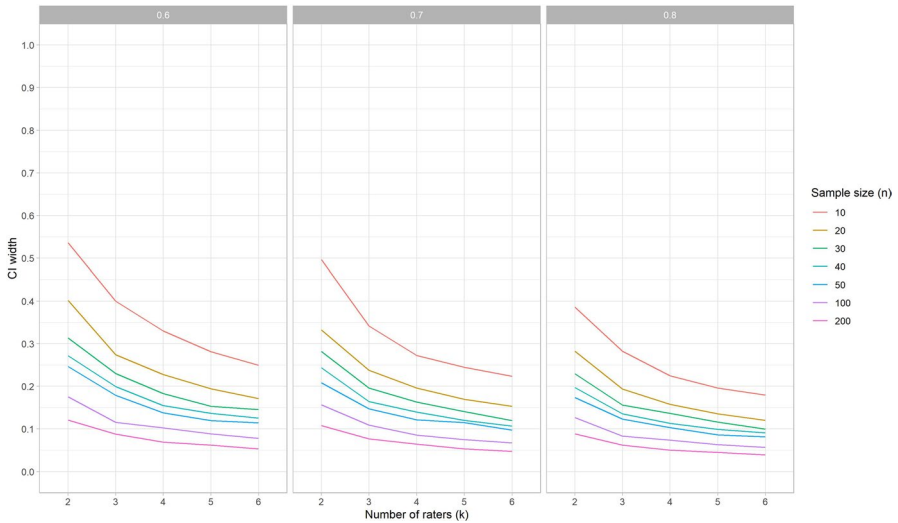


Fig. 8 95% confidence interval width for the ICC for $r=0.6, 0.7$ or 0.8 (averaged over SEM models, $v=1$, no raters deviate)

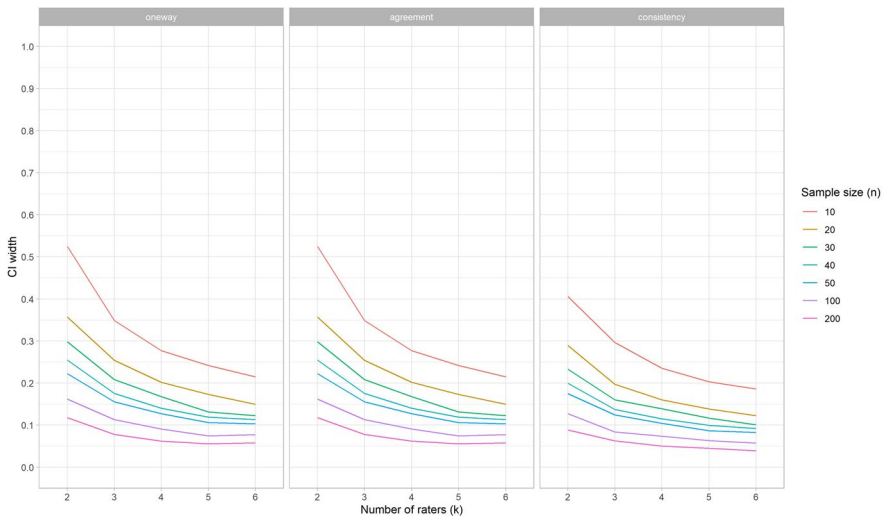


Fig. 9 95% confidence interval width for the three SEM models, when one rater deviates ($v=1, r=0.8$)

adapted design, which refers to the number of patients or repeated measurements of the new design, e.g., the numbers that are included in the study so far. For both procedures you are asked to indicate the target width of the 95% CI of the parameter of interest. The width depends of the unit of measurements. As the range of the ICC is always between 0 and 1, the range of the target width is fixed, and it is set default at 0.3 in the online application. However, the SEM depends on the unit of measurement, and changes across conditions v . Therefore, in the online application, the range for the target width of the 95% CI for the SEM changes across conditions v , and various default settings are used.

Table 3 Choices and assumptions per approach that are available in the online application (<https://irise.ekhout.shinyapps.io/ICCpower/>)*Width of the CI procedure*

Choice of type of ICC or SEM (one-way random effects model, two-way random effects model, or two-way mixed effects model)

Choice about the target width of the 95% CI for the parameter

Assumption about correlation between the repeated measurements (e.g., the raters) (i.e., 0.6, 0.7 or 0.8)

Assumption about occurrence of systematic difference between repeated measurements (e.g., raters)

Assumption about variance in score (i.e., 1, 10, or 100)

CI lower limit method

Choice of the acceptable lower limit of 95% confidence interval

Target design

MSE ratio procedure

Choice of type of ICC or SEM

Choice about the target width of the 95% CI for the parameter

Assumption about correlation between the repeated measurements (e.g., the raters)

Assumption about occurrence of systematic difference between repeated measurements (e.g., raters)

Assumption about variance in score

Target design

Adapted design

Target design refers to the intended number of patients or number of repeated measurements chosen at the start of the study as described in the study protocol; adapted design refers to the number of patients or repeated measurements in the new design, e.g., included so far in the study

In the design phase of a study, before the data collection has started, two approaches can be used. For example, to obtain the sample size recommendation to obtain the ICC_{agreement} with the CI width procedure, we need to make some assumptions on the correlation between the repeated measurements, the presence of a systematic difference and expected variance in score. If we assume the measurements will be correlated with 0.8 (in other words, you expect to find an ICC of 0.8), with no systematic difference between the measurements (e.g., the raters), and the expected variance between the score is 10. Based on this information, we will get an overview as shown in Fig. 10.

By scrolling over the different blocks in the online application, we can easily see what the consequence is for the width of the CI around the estimated ICC when adding an extra rater or including more patients. For example, when we use 3 raters and 20 patients, the estimated width of the CI around the ICC estimation is 0.293; or when $k=2$ and $n=30$ the width of the CI is 0.278; and when $k=2$ and $n=25$ the width is 0.33. In the online application this information automatically pops up.

If we compare the results for various conditions in the application, we see that the impact of whether or not a systematic difference exist on the sample size recommendations is much larger than the impact of different values for the variance between the scores, specifically when in the one-way random effects model, or the two-way random effects model for agreement.

The second procedure that can be used in the design phase is the CI lower limit procedure. This procedure is developed by Zou for ICC_{one-way}. Note that procedure may lead to an overestimation of the required sample size for ICCs based on a two-way effects model (see results, and (Donner and Eliasziw 1987)). An example to use this procedure: if we expect the ICC to be 0.8, and we accept a lower CI limit of the ICC

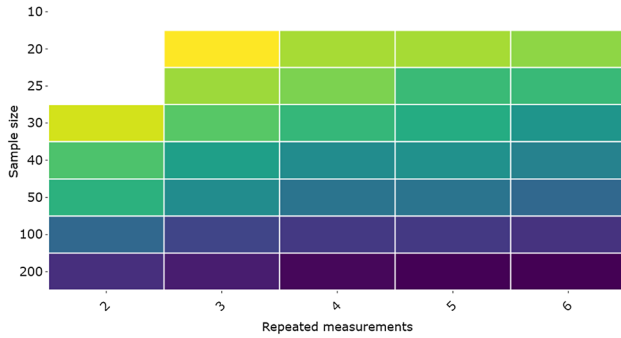


Fig. 10 Print screen of the results of the expected width of the 95% confidence interval of the ICC_{agreement} for sample size and rater combination based on the CI width procedure under the expected conditions $r=0.8$, $v=10$, and no systematic difference between raters

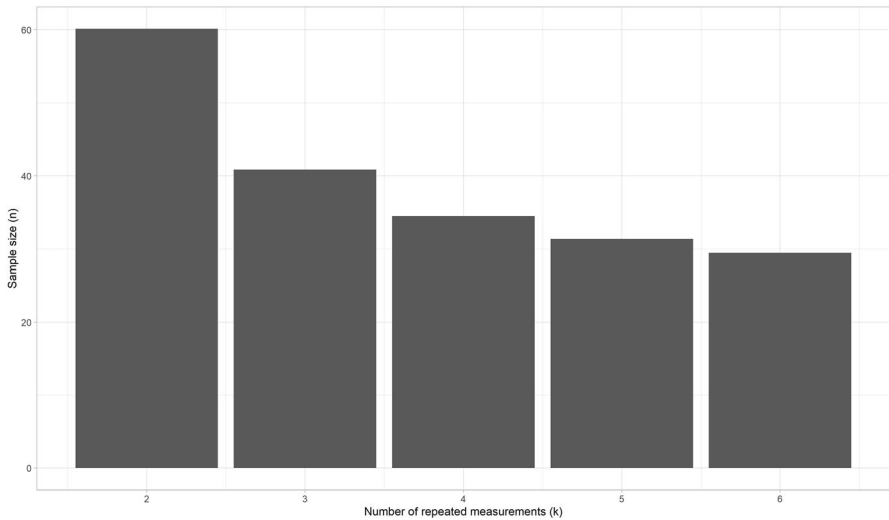


Fig. 11 Recommendations for n and k using the CI lower limit procedure ($r=0.8$, acceptable lower bound of CI is 0.65, ICC_{one-way})

of 0.65, depending on the number of repeated measurements that will be collected, the adequate sample size is given (see Fig. 11). For example, for $k=3$, a sample size of 40 is appropriate (under the given conditions). As this procedure is based on a formula, it can be used beyond the conditions chosen in the simulated data.

The third procedure, the MSE ratio procedure, is most suitable when we have started the data collection and realize that the target design cannot be reached. In that case we want to know how an adapted design compares to our target design that was described in the study protocol. Suppose that patients were observed in clinical practice and scored by three raters ($k=3$) at (about) the same time. We envisioned 50 patients (i.e., target design). The number of raters cannot be changed anymore, as patients will possibly have changed on the construct measured, or it is logistically impossible to invite the same patients to come

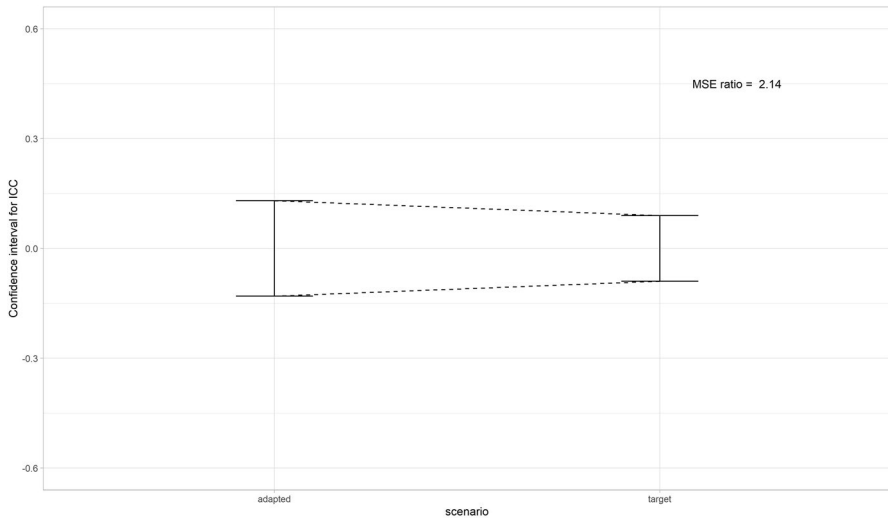


Fig. 12 Print screen of the expected decrease in width of the 95% confidence interval of the ICC between the adapted ($k=3$, $n=25$) and target design ($k=3$, $n=50$) based on the MSE ratio procedure

back for another measurement. Based on the results of previous studies, or by running preliminary analyses on the collected data within this study, we can make assumptions about: the expected correlation between the raters (i.e., the repeated measurements; e.g., 0.8), whether we expect one of these raters to systematically deviate from the others (e.g., no), and the expected variance in score (e.g., 10). Suppose we have collected data of three raters that each measured 25 patients; this is our adapted design. Now, we can see how much the 95% CI will increase, when we don't continue collecting data until we have included 50 patients (i.e., your target design) (Fig. 12). The 95% CI will increase approximately from 0.2 that we would have had if we measured 50 patients three times (i.e., target design) to 0.3 now in the adapted design.

Another way to use this method, is to see how much one of the two variables n or k should increase to preserve the same level of precision as in the target design. For example, in the target design 3 raters would assess 25 patients. As one of the raters dropped out, there are only 2 raters in the adapted design. The MSE ratio in this scenario was 1.43. To achieve the same level of precision in the adapted design with 2 raters as in the target design ($n=25$, $k=3$), the sample size should be increased by 1.43, resulting in a sample size of $n=36$.

4 Discussion

From the simulation studies we learn that most gain in precision (i.e., largest change in MSE values) can be obtained by increasing an initially small sample sizes or small number of repeated measures. For example, an increase from 2 to 3 raters gains more precision than from 4 to 5 raters, or when the sample size is increased from 10 to 20 compared to an increase from 40 to 50. Moreover, results show that the expected ICC (i.e., correlation between the repeated measurements), and the presence of a systematic difference have

most influence on the precision of the ICC and SEM estimations. Specifically, when the correlation increases the precision increases (i.e., smaller MSE values, and smaller width of CI). When one rater deviates the MSE values for $ICC_{one-way}$, $SEM_{one-way}/SEM_{agreement}$ increase, the coverage of the $ICC_{agreement}$ and the $ICC_{one-way}$ changes and the width of the CI increases for the $ICC_{one-way}$ and $ICC_{agreement}$, but not for the $ICC_{consistency}$. For example, to achieve an estimation with a width of the confidence interval of approximately 0.3 using an $ICC_{agreement}$ model when one of the three raters systematically deviates, the sample size needs to be around 40 (when $r=0.7$) or 30 (when $r=0.8$). When no systematic difference occurs between the three repeated measurements, the required sample size when $r=0.7$ or $r=0.8$, can be lowered to approximately $n=35$ or $n=20$, respectively, to obtain an estimation of the $ICC_{agreement}$ with a CI width of 0.3.

Throughout this paper, we used ‘raters’ as the source of variation that varied across the repeated measurements, but the results are not limited to the use of raters as the source of variation. Accordingly, all results and recommendations also hold for other sources of variation, however feasibility of the recommendations may differ. For example, in a test–retest reliability study ‘occasion’ is the sources of variation of interest. However, it may not be feasible to have three repeated measurements of patients as patients may not be stable between three measurements. When only two repeated measurements can be obtained, sample size requirements increase. Note, that we only took one-way or two-way effect models into account, and we cannot generalize these results to three-way effects models. We did not simulate conditions of n between 50 and 100. Therefore, we can only roughly recommend that when there is a systematic difference between the repeated measurements, required sample size will increase up to 100, specifically when the $ICC_{one-way}$ model is used, and likely around 75 when the $ICC_{agreement}$ is used. Recommendations for specific conditions can be found in the online application (<https://iriseekhout.shinyapps.io/ICCpower/>).

The selected sample of patients should be representative of the population in which the instrument will be used, as the variation of the patients will influence the ICC value. The result of the study can only be generalized to this population. The same holds for the selection of professionals that are involved in the measurements and any other source of variation that is being varied across the repeated measurements. Selecting only well trained raters in a reliability study will possibly decrease the variation between the raters, and subsequently influence the ICC and SEM estimation. Therefore, it is important to well-consider which patients and which professionals and other sources of variation are selected for the study. For an appropriate interpretation of the ICC and SEM values, complete reporting of research questions and rationale of choices made in the design (i.e., choice in type and number of patients, raters, equipment, circumstances etc.) is indispensable (Mokkink et al. 2022).

As measurements can be expensive and burdensome to patients and professionals, we do not recommend to collect more data than required to estimate ICC or SEM values as this would lead to research waste. Therefore, it is important to involve these feasibility aspects in the decisions of the optimal sample size and repeated measurements. When a systematic difference between raters occurs, we showed that the use of a one-way model requires a higher sample size compared to two-way random effect models for agreement, which subsequently requires a higher sample size than the two-way mixed model for consistency (see Fig. 7). The difference in data collection between the models, is that two-way effect models require extra predefined measurement conditions (Mokkink et al. 2022), e.g., only rater A and B are involved and measure all patients, while in one-way effects models no measurement conditions are defined, and any rater could measure the patient at any

occasion. As the goal of a reliability study is often to understand the influence of a specific source of variation (e.g., the rater) on the score (i.e., its systematic difference), a two-way random effects model is often the preferred statistical method (Mokkink et al. 2022). We have showed that this is also the most efficient model precision-wise.

Our recommendations are in line with other recommendations. Previous studies showed that the sample size is dependent on the correlation between the repeated measurements (Shoukri et al. 2004), and that adding more than three repeated measurements gains only little precision (Giraudeau and Mary 2001; Shoukri et al. 2004). However, we provide recommendations under more conditions, i.e., for three types of effect models, and with and without systematic differences. Moreover, we present our recommendations in a user-friendly way by the development of the Sample size decision assistant, that is available in the online application.

As an example, we used an appropriate width of the confidence interval around the point estimate of 0.3. We could have chosen another width. Zou (2012) used 0.2 as an appropriate interval, which we considered quite small. However, in the online application the consequences on precision with a width of 0.2 can be examined as well.

4.1 Strength and limitations

In this study we considered a large variety of conditions for the variables n , k , v , r . In contrast to previous studies on required sample sizes, we used three different and commonly used statistical models to estimate the parameters, and incorporated systematic differences between the repeated measurements. Moreover, we investigated the bias and precision of the ICC as well as of the SEM.

There are some limitations. Using 1000 samples seems arbitrarily. We calculated Monte Carlo standard errors (Morris et al. 2019), resulting in standard errors < 0.0001 in all simulation conditions for the MSE estimate. So we conclude that 1000 samples are enough to obtain reasonably precise estimates. In this simulation study, it was not feasible to calculate confidence intervals around the SEM for each sample, as we would have to use bootstrap techniques in each of the $2 \times 360 \times 1000$ and 216×1000 samples. Therefore, we could not evaluate the coverage for SEM. The precision of SEM is reflected in the MSE, and with this MSE we can compute the confidence interval width for each condition. This confidence width is used to reflect the gain precision of SEM in the MSE ratio procedure and to use as a criterion itself in the CI width procedure. The results for bias and MSE of SEM showed similar trends for precision and accuracy for SEM estimation as for ICC estimation.

Generally, we can see that the SEM can be estimated with more precision than the ICC. When in doubt, we propose to use recommendations on sample size and number of repeated measurements for ICC.

Furthermore, as conclusion based on simulation studies are restricted to the conditions investigated, our study is limited in that aspect. We only simulated three conditions of the correlations between repeated measurements ($r = 0.6, 0.7$ and 0.8) and we concluded that the presence of a systematic difference has most influence on the width of the confidence interval, specifically with the larger correlations (0.7 and 0.8). As we did not simulate the condition 0.9 , we don't know to what extent that holds for this condition. Moreover, we did not simulate any condition for n between 50 and 100. Therefore, we cannot give precise recommendations for when $k=2$, as it is likely that the appropriate sample size in this situation will be between 50 and 100. Last, we simulated a systematic difference in one ($k=2-6$) or two ($k=2-4$) raters. However, the way the two raters deviated in the latter

condition was the same (i.e., by increasing the average score of the rater with 1 standard deviation in score). Other ways that raters may deviate were not investigated. Nevertheless, we feel that the use of one standard deviation deviance for one or two raters demonstrates a sufficient difference to test the relative performance of the two-way effect models, but not too large to be unrealistic.

We used multilevel methods to estimate the variance components that are subsequently used to calculate ICCs and SEMs. These methods are robust against missing data and able to deal with unbalanced designs. However, investigating the impact of missing data on the precision of the ICC and SEM estimations was beyond the scope of this study.

Our findings are utilized in an online application. The different ways in which this tool can be used provides insight into the influence of various conditions on the sample sizes and into the trade-off between various choices. Using this tool enables researchers to use the study findings to estimate required sample sizes for number of patients or number of raters (or other repetitions) for an efficient design of reliability studies. We aim to continue to improve the design and layout of the app to improve usability and user-friendliness of the application and to broaden the scope of the recommendations to match with the demands of users.

Appendix 1: Model specifications of ICCs and SEMs and the Agree package for R

The Agree package is developed to calculate the reliability and measurement error between the scores of multiple raters or repeated measurements in stable patients (Eekhout 2022). 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. The research question together with the corresponding design of the study determine the best statistical model to analyze the data (Mokkink et al. 2022). In this “Appendix,” we will summarize these statistical models, and subsequently, we provide the a brief explanation of the functions from the Agree package that were used and present the R code that we used in the simulation studies. For more information on specific statistical models (including the multi-level models) we refer to Mokkink et al. (Mokkink et al. 2022).

Statistical models for ICCs and SEMs

One-way random effects model

In the design of the one-way random effects model the observers are unknown, so the effect of observers is not present in the model. This model is specified in Eq. 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 observation j of patient i , β_0 denotes the overall population mean of the observations, 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 test scores with a mean of 0 and a variance of σ_e^2 . In this model the observed test scores are only explained by the differences between patients.

The $ICC_{one-way}$ is the variance between the subjects (σ_j^2) divided by the sum of the subject variance (σ_j^2) and the residual variance (σ_e^2). The $ICC_{one-way}$ is computed as follows: $ICC_{one-way} = \sigma_j^2 / (\sigma_j^2 + \sigma_e^2)$. The $ICC_{one-way}$ assumes that each subject is rated by a different set of raters, that are randomly selected from a larger population of judges (Shrout and Fleiss 1979). The $SEM_{one-way}$ is the square root of the error variance (i.e., $SEM_{one-way} = \sqrt{\sigma_e^2}$). For the $ICC_{one-way}$, and $SEM_{one-way}$ only the level 1, the patient level, is random. The rater variance is not used.

Two-way random effects model for agreement

In the design of the two-way random effects model of agreement and the two-way mixed effects model of consistency the observers 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 observers, as presented in Eq. 2.

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

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

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

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

where x_{ij} denotes the score for observation j of patient i , β_0 denotes the overall population mean of the observations, a_{0i} denotes the random patient effect with a mean of 0 and a variance of σ_{a0}^2 , c_{0j} denotes the random observer effect with a mean of 0 and a variance of σ_{c0}^2 and e_{ij} denotes the residual variance in the observed test scores with a mean of 0 and a variance of σ_e^2 . This model accounts for systematic differences between observers represented in the random effect of the observers. The $ICC_{agreement}$ is the variance between the subjects (σ_j^2) divided by the sum of the subject variance (σ_j^2), rater variance (σ_k^2) and the residual variance (σ_e^2). The $ICC_{agreement}$ is computed as follows: $ICC_{agreement} = \sigma_j^2 / (\sigma_j^2 + \sigma_k^2 + \sigma_e^2)$ (Shrout and Fleiss 1979). All subjects are rated by the same set of raters, and the rater variance is taken into account in the calculation of the ICC. The $SEM_{agreement}$ is the square root of the sum of the rater variance and the error variance (i.e., $SEM_{agreement} = \sqrt{\sigma_r^2 + \sigma_e^2}$). For the $ICC_{agreement}$ and the $SEM_{agreement}$ both the level 1 and level 2 are random.

Two-way mixed effects model for consistency

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

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

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

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

where x_{ij} denotes the score for observation j of patient i , β_0 denotes the overall population mean of the observations, a_{0i} denotes the random patient effect with a mean of 0 and a variance of σ_{a0}^2 , c_1 denotes the fixed observer effect (so this effect does not vary between observers as opposed to the observer effect (c_{0j}) specified in Eq. 2) and e_{ij} denotes the residual variance in the observed test scores with a mean of 0 and a variance of σ_e^2 . The $ICC_{\text{consistency}}$ is the variance between the subjects (σ_j^2) divided by the sum of the subject variance (σ_j^2) and the residual variance (σ_e^2). The rater variance is not used to calculate the ICC and can therefore also be considered as a fixed effect. The $ICC_{\text{consistency}}$ is computed as follows: $ICC_{\text{consistency}} = \sigma_j^2 / (\sigma_j^2 + \sigma_e^2)$ (Shrout and Fleiss 1979). The $SEM_{\text{consistency}}$ is the square root of the error variance (i.e., $SEM_{\text{consistency}} = \sqrt{\sigma_e^2}$). For the $ICC_{\text{consistency}}$ and $SEM_{\text{consistency}}$ the level 1 (subject) is a random effect and the level 2 (rater) is fixed.

The Agree package for R

The package can be installed directly from GitHub by:

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

In the Agree package the `icc()` function computes the parameter for both the one- and two-way effects models. The `icc()` function can be used to estimate the reliability parameters (variance components and ICC's), the 95% confidence intervals for the ICCs, and the SEM parameters. The confidence intervals for ICC one-way and consistency are computed with the exact F method. $F = (k * \sigma_j^2 + \sigma_e^2) / \sigma_e^2$, with $df1 = n - 1$ and $df2 = n(k - 1)$ (Shrout and Fleiss 1979). For the $ICC_{\text{agreement}}$ an approximate CI was derived, which accounts for the three independent variance components (Satterthwaite 1946; Fleiss and Shrout 1978).

Simulation application

For the simulation study we used the `icc()` function from the Agree package and specified that the three types of reliability and measurement error should be computed from the same model by using `icc(data, onemodel = TRUE)`. Consequently, first the two-level model was estimated with a random intercept for both patients and raters. The $ICC_{\text{agreement}}$ was obtained with the estimated variance components from this model (i.e., for patient, rater and residual). For the $ICC_{\text{consistency}}$, only the estimated variance at the subject level and the error variance were used to compute the ICC. Accordingly, these variance components are

adjusted for rater variance, but rater variance is not used to compute the ICC. To obtain the ICC_{one-way}, the estimated rater variance is part of the error variance. The subject variance is computed by subtracting the rater variance from the sum of subject variance over the raters, which is then averaged (i.e., $\sigma_j^2 = ((k * \sigma_j^2) - \sigma_k^2) / k$).

R code for simulation study

```
#simulation conditions.
k=c(2,3,4,5,6) #number of raters.
#k <-k[2].
cor=c(0.6, 0.7, 0.8) #raw correlation between the repeated measurements.
#cor <-cor[2].
n=c(10,20,30,40,50,100,200) #sample size (patients).
#n <-n[2].
vari=c(1, 10, 100) #variance in scores.
#vari <-vari[1].
#means when 1 rater deviates.
means <-c(sqrt(vari), rep(0,(k-1))) #means in scores, when first rater deviates with 1
standard deviation.
means <-means * vari.
#covariance.
icc_cor <-matrix(cor,k,k) #correlation matrix.
diag(icc_cor) <-1.
icc_cov <-icc_cor*vari.
#simulation data sample.
data1 <-as.data.frame(MASS::mvrnorm(means, icc_cov, n=n)).
#compute ICC types and SEM.
icc(data1, onemodel = TRUE).
```

Appendix 2: MSE values of ICC_{agreement} estimations for k per sample size n for each condition r, when one rater systematically deviates

See Fig. 13.

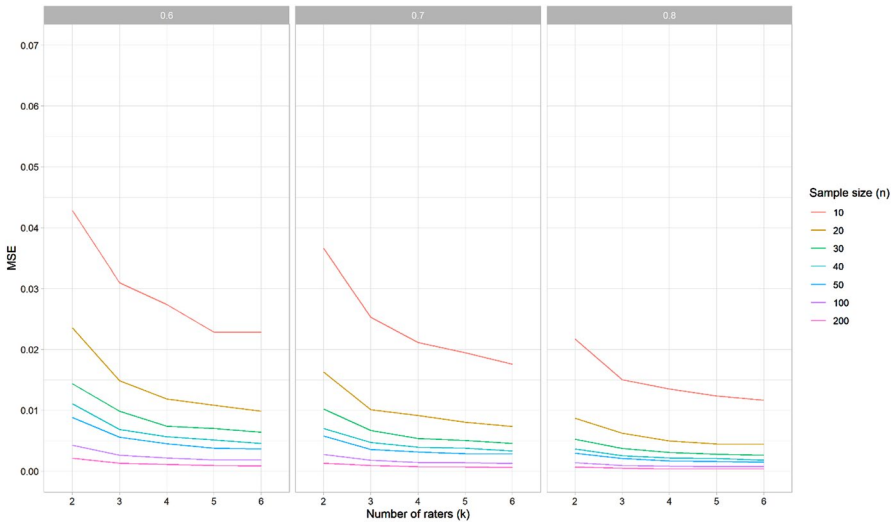


Fig. 13 MSE values of $ICC_{\text{agreement}}$ estimations for k per sample size n for each condition r (one rater systematically deviates; averaged over all conditions of v)

Author contributions LM, HdV and IE contributed to the study conception and design. Material preparation, data simulation and analysis were performed by LM, SD, and IE. The first draft of the manuscript was written by LM and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work is part of the research programme Veni (received by LM) with Grand No. 91617098 funded by ZonMw (The Netherlands Organisation for Health Research and Development). The funding body has no role in the study design, the collection, analysis, and interpretation of data or in the writing of this manuscript.

Data availability The syntax used to generate the datasets that were analyzed during the current study is available in “Appendix 1”.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval and consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bland, J.M.: How can I Decide the Sample Size for a Study of Agreement Between Two Methods of Measurement? <https://www-users.york.ac.uk/~mb55/meas/sizemeth.htm> (2004)
- Burton, A., Altman, D.G., Royston, P., Holder, R.L.: The design of simulation studies in medical statistics. *Stat Med* **25**(24), 4279–4292 (2006). <https://doi.org/10.1002/sim.2673>
- de Vet, H.C., Terwee, C.B., Knol, D.L., Bouter, L.M.: When to use agreement versus reliability measures. *J. Clin. Epidemiol.* **59**, 1033–1039 (2006)
- de Vet, H.C., Terwee, C.B., Mokkink, L., Knol, D.L.: *Measurement in Medicine*. Cambridge University Press, Cambridge (2011)
- Donner, A., Eliasziw, M.: Sample size requirements for reliability studies. *Stat. Med.* **6**(4), 441–448 (1987). <https://doi.org/10.1002/sim.4780060404>
- Dikmans, R.E.G., Nene, L.E.H., Bouman, M.B., de Vet, H.C.W., Mureau, M.A.M., Buncamper, M.E., Winters, H.A.H., Ritt, M.J.P.F., Mullender, M.G.: The aesthetic items scale: a tool for the evaluation of aesthetic outcome after breast reconstruction. *Plast. Reconstr. Surg. Glob. Open* **5**(3), e1254 (2017). <https://doi.org/10.1097/GOX.0000000000001254>
- Eekhout I.: Agree: Agreement and Reliability Between Multiple Raters. R package version 0.1.8. 2022. <https://github.com/iriseekhout/Agree/> (2022). Accessed 8 March 2022
- Eekhout, I., Enders, C.K., Twisk, J.W.R., de Boer, M.R., de Vet, H.C.W., Heymans, M.W.: Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Struct. Equ. Model. Multidiscip. J. Clin. Epidemiol.* **22**(4), 1–15 (2015)
- Eekhout, I., Mokkink, L.B.: Estimating ICCs and SEMs with Multilevel Models. <https://www.iriseekhout.com/r/agree/> (2022)
- Fleiss, J.L., Shrout, P.E.: Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika* **43**(2), 259–262 (1978)
- ICC power, shiny app.io.: Eekhout, I., Mokkink, L.B. [Mobile application software]. Retrieved from <https://iriseekhout.shinyapps.io/ICCPower/> (2022)
- Giraudeau, B., Mary, J.Y.: Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Stat. Med* **20**(21), 3205–3214 (2001). [https://doi.org/10.1002/sim.935\[pil\]](https://doi.org/10.1002/sim.935[pil])
- Lu, M.J., Zhong, W.H., Liu, Y.X., Miao, H.Z., Li, Y.C., Ji, M.H.: Sample size for assessing agreement between two methods of measurement by Bland-Altman method. *Int. J. Biostat.* (2016). <https://doi.org/10.1515/ijb-2015-0039>
- Mokkink, L.B., Eekhout, I., Boers, M., van der Vleuten, C.P., De Vet, H.C.: *Studies on Reliability and Measurement Error in Medicine—from Design to Statistics Explained for Medical Researchers* (2022)
- Morris, T.P., White, I.R., Crowther, M.J.: Using simulation studies to evaluate statistical methods. *Stat. Med.* **38**, 2074–2102 (2019). <https://doi.org/10.1002/sim.8086>
- Mosmuller, D., Tan, R., Mulder, F., Bachour, Y., de Vet, H., Don Griot, P.: The use and reliability of SymNose for quantitative measurement of the nose and lip in unilateral cleft lip and palate patients. *J. Craniomaxillofac. Surg.* **44**(10), 1515–1521 (2016). <https://doi.org/10.1016/j.jcms.2016.07.022>
- Mulder, F.J., Mosmuller, D.G.M., de Vet, H.C.W., Moues, C.M., Breugem, C.C., van der Molen, A.B.M., Don Griot, J.P.W.: The cleft aesthetic rating scale for 18-year-old unilateral cleft lip and palate patients: a tool for nasolabial aesthetics assessment. *Cleft Palate Craniofac. J.* **55**(7), 1006–1012 (2018). <https://doi.org/10.1597/16-123>
- Nunnally, J.C., Bernstein, I.H.: *Psychometric Theory*, 3rd edn. McGraw-Hill, New York (1994)
- Saito, Y., Sozu, T., Hamada, C., Yoshimura, I.: Effective number of subjects and number of raters for inter-rater reliability studies. *Stat. Med.* **25**, 1547–1560 (2006)
- Satterthwaite, F.E.: An approximate distribution of estimates of variance components. *Biom. Bull.* **2**, 110–114 (1946)
- Streiner, D.L., Norman, G.: *Health Measurement Scales. A Practical Guide to their Development and Use*, 4th edn. Oxford University Press, New York (2008)
- Shoukri, M., Asyali, M.H., Donner, A.: Sample size requirements for the design of reliability study: review and new results. *Stat. Methods Med. Res.* **13**, 251–271 (2004)
- Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428 (1979)
- Venables, W.N., Ripley, B.D.: *Modern Applied Statistics with S-plus*. Springer, New York (2002)
- Walter, S.D., Eliasziw, M., Donner, A.: Sample size and optimal designs for reliability studies. *Stat. Med.* **17**, 101–110 (1998)

- Walton, M.K., Powers, J.H., III, Hobart, J., Patrick, D., Marquis, P., Vamvakas, S., Isaac, M., Molsen, E., Cano, S., Burke, L.B.: Clinical outcome assessments: a conceptual foundation—report of the ISPOR clinical outcomes assessment emerging good practices task force. *Value Health* **18**, 741–752 (2015)
- Zou, G.Y.: Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Stat. Med.* **31**, 3972–3981 (2012)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.