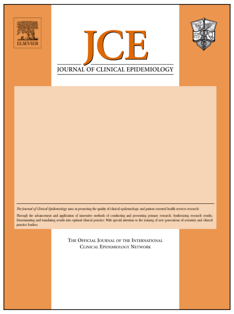


The measurement properties reliability and measurement error explained – a COSMIN perspective

Lidwine B. Mokkink, Iris Eekhout

PII: S0895-4356(25)00391-9

DOI: https://doi.org/10.1016/j.jclinepi.2025.112058

Reference: JCE 112058

To appear in: *Journal of Clinical Epidemiology*

Received Date: 1 July 2025

Revised Date: 16 October 2025

Accepted Date: 12 November 2025

The measurement properties reliability and measurement error explained – a COSMIN perspective

Lidwine B Mokkink[1,2], Iris Eekhout[1,3]

[1] Amsterdam UMC location Vrije Universiteit Amsterdam, Epidemiology and Data Science, Amsterdam, Netherlands;

[2] Amsterdam Public Health research institute, Methodology, Amsterdam, the Netherlands;

[3] Child Health, Netherlands Organisation for Applied Scientific Research (TNO), Leiden, the Netherlands;

Corresponding author:

LB (Wieneke) Mokkink, w.mokkink@amsterdamumc.nl. Department of Epidemiology and Data Science, Amsterdam UMC, Vrije Universiteit Amsterdam, Location AMC, J1B-225, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands

Keywords: reliability; measurement error; classical test theory; intraclass correlation coefficient; standard error of measurement; COSMIN

Abstract

Reliability and measurement error are related but distinct measurement properties. They are connected because both can be evaluated using the same data, typically collected from studies involving repeated measurements in individuals who are stable on the outcome of interest. However, they are calculated using different statistical methods and refer to different quality aspects of measurement instruments.

We explain that a measurement error refers to the precision of a measurement, that is, how similar or close the scores are across repeated measurements in a stable individual (variation within individuals). In contrast, reliability indicates an instrument's ability to distinguish between individuals, which depends both on the variation between individuals (i.e. heterogeneity in the outcome being measured in the population) and the precision of the score, i.e. the measurement error. Evaluating reliability helps to understand if a particular source of variation (e.g. occasion, type of machine, or rater) influences the score, and whether the measurement can be improved by better standardizing this source. Intraclass-correlation coefficients, standards error of measurement and variance components are explained and illustrated with an example.

Introduction

Reliability and measurement error are related but distinct measurement properties. They are connected because both can be evaluated using the same data, typically collected from studies involving repeated measurements in individuals who are stable on the outcome of interest. However, they are calculated using different statistical methods and refer to different quality aspects of measurement instruments.

In this COSMIN Perspective, a series of papers on measurement properties and measurement-related issues, we explain the relationship and distinctions between these two properties by explaining their concepts, the appropriate study design, and the relevant statistics and interpretation, using an illustrative example.

The concepts

Reliability and measurement error are measurement properties that should be understood within the framework of Classical Test Theory (CTT) (1). According to this theory, an observed score consists of a true but unknown score, and a measurement error. This means that every measurement instrument (as defined by its five components (2)) inherently contains some degree of error. In a study on reliability and measurement error, the focus is on the *value* (or consistency) of the observed score, rather than on the accuracy (i.e. adequacy of the representation of the outcome, or its validity) (1). The “true score” refers to the score without any measurement error, which can only be approximated by repeating the assessment of the score an infinite number of times (which is of course impossible).

Measurement error

A measurement error refers to how similar or close the scores are from repeated measurements in an individual who is stable on the outcome of interest (3). It focuses on the consistency or value of the measurements in a single individual, reflecting the precision of a score.

If we could observe an individual’s outcome through an infinite series of measurements, the resulting distribution of scores would approximate a normal curve centered on their mean score, which is regarded as the individual’s true score. The width of this (theoretical) distribution represents the measurement error (see Appendix 1). This measurement error should not be confused with the standard error of the mean in a sample, which represents the dispersion of sample means around the population mean. In CTT it is assumed that an individual with a low true score shows the same standard deviation in observed scores as an individual   with a high true score. Since it is impossible to measure a person an infinite number of times, we estimate the measurement error using two or three repeated measurements in multiple individuals who are stable on the measured outcome (see Figure 1).

___ Please, insert Figure 1 here_________

Figure 1. Data of two repeated measurements in persons stable on the outcome for estimating the measurement error; in panel A are first (T1) and second (T2) scores specified with average of the two measurements (i.e. middle dot) as a person's true value; in panel B are scores centered around a person's true score.

The measurement error of an instrument is expressed in the same unit as its corresponding score (i.e. its unit of measurement). This facilitates the interpretation. For example, when a person's weight is measured, the measurement error of a weighing scale is expressed in grams. Some instruments, like many patient-reported outcome measures (PROMs), don't have a meaningful unit of measurement, and hence neither does the measurement error.

Reliability

Reliability indicates an instrument's ability to distinguish between individuals – with various scores and despite measurement error (3). This ability depends on the variation between the persons (heterogeneity in the outcome being measured in the population), and the measurement error, i.e. variation within persons (differences in scores due to other sources of variation, such as variation due to a rater, the occasion, or the equipment used).

When individuals have distinctly different true scores, it is easier for the instrument to differentiate between them, even if there is some measurement error. Consider measuring the heights of children aged 0 to 18 years in a random sample of 30 individuals. Heights in this group might range from 50 cm (for infants) to 198 cm (for older teenaged boys). Even if the measurement error is relatively high, such as 2 cm, the large differences in true heights allow the instrument to maintain a consistent ranking of individuals from shortest to tallest. If the measurement is repeated, the order of individuals will remain largely the same, leading to high reliability. Now imagine measuring the heights of 30 children who are all close in age and size with an absolute difference in size of only 10 cm. In this case, the measurement error of 2 cm represents a larger proportion of the total variation. As a result, the ranking of individuals might differ between repeated measurements—one child might rank taller than another in the first measurement but shorter in the second. These inconsistencies in ordering reflect lower reliability. In the rare case where all individuals are exactly the same height, no instrument can distinguish between them, and the reliability would be zero.

When scores from repeated measurements are compared, reliability can be visualized as the consistency of the order of individuals between the two measurements. If the order changes between the first measurement (T1) and the second measurement (T2), reliability is lower.

The Design

The basic design of a study on reliability and measurement error involves *repeated measurements* in *individuals stable on the outcome of interest*. These repeated measurements are conducted under consistent measurement conditions, with one (or more) condition purposefully varied. This condition represents the source(s) of variation under investigation (e.g. rater, occasion or administration mode of a measurement instrument) (4). All other conditions are either held constant, or are assumed not to influence the outcome score.

In a reliability study, the goal is to assess the influence of various sources of variation on the score. While the individual's true score influences the observed score, other factors do as well. For example, if the goal is to understand the influence of different raters on the score and raters are varied across the repeated measurements, an inter-rater reliability study is conducted. If you are interested in the variation of occasion, the measurements are repeated over time in a test-retest design. To understand the influence of the same rater over time, an intra-rater reliability study is conducted with repeated measurements conducted by the same rater. Note, that test-retest and intra-rater studies can share the same design structure, repeated measures by the same rater over time, but the underlying assumption is different (3). In a test-retest study, we assume stability of raters, and investigate the influence of occasion. In an intra-rater design, we assume stability of occasions, and we investigate the influence of one rater.

All of these designs rest on the assumption of stability of the outcome in a person, that the true score of the individual remains unchanged between repeated measurements. The validity of this assumption depends on the nature of the measured outcome (some outcomes are more stable than others), and the time interval between repeated measurements. If measurements are repeated immediately, participants may become tired or get trained (e.g. in case of a performance-based test), or remember their scores (e.g. on questionnaires). If time between measurements increases, true scores may have changed, and the assumption of outcome stability does not hold.

For any measurement instrument, different sources of variation can influence the score. In order to decide which sources to study for a particular measurement instrument, one can consider the components of the measurement to facilitate designing the study (4), specifically for more complex designs, in which multiple sources are varied across the repeated measurements. For example, Nederpelt et al. evaluated 6 MRI-based parameters of brain atrophy, in 7 different brain regions (i.e. 42 parameters). Per parameter per brain region they conducted reliability studies, varying the brand of the machine (3 types of machines) and the occasions (test and retest), leading to 6 scans per patient. Subsequently, to provide the images both fillings and non-fillings were used. And last, each scan was analyzed using 6 different software packages. In the end per parameter, 72 scores were obtained per patient (see Figure 2).

This data can be analyzed at the same time, and variance components are obtained per source of variation. The sources with relatively high variance components are most influencing the scores, and could be improved, e.g. by better standardizing (e.g. improving instructions), or by restricting the source (e.g. measure a person on the very same brand of machine).

In a reliability study measurements need to be repeated at least twice, but simulation studies suggest that three times is most efficient (5), in terms of the least total numbers of measurements. However, this may not be feasible for some outcomes or measurement instruments. For recommendations on the appropriate sample size for patients and repeated measurements, we refer to the Sample size Decision Assistant, which helps to decide on the number of study participants (patients) and repeated measurements (for example number of raters) (6).

Illustrative example

The following example regarding facial tension of six babies illustrated how to examine the reliability and measurement error. Table 1 shows scores on the item 'facial tension' of the COMFORT scale (5) for 6 babies from 4 health care professionals (i.e. raters) who independently rated each baby at the same time. This setup ensures the assumption of stability is met. The mean scores for the babies ranged from 4 to 7.5, while the mean scores given by the raters ranged from 2.5 to 7.7.

_________Please, insert Table 1 here______________

Table 1. Artificial raw data on facial tension of six babies by four raters on the COMFORT scale ranging from 1-10 (high score means more discomfort).

The statistics

Using the same dataset, it is possible to calculate parameters for both reliability and measurement error. For continuous scores the most commonly used parameters are intraclass correlation coefficients (ICCs) for reliability, and the standard error of measurement (SEM) for single scores. There are various types of ICCs, and SEMs, each suited to different study designs and assumptions. In Mokkink et al. (3) these types are described in detail. Here we focus on the most commonly applied versions: $\text{ICC}_{\text{agreement}}$, and the $\text{SEM}_{\text{agreement}}$. Both parameters are based on variance components (σ^2). A variance component reflects the influence of a source of variation on the score and can be estimated with an repeated-measures ANOVA or multi-level analysis.

In the example study design, individual babies are the individuals being measured, and the raters are the measurement condition (i.e. source of variation) that is purposefully varied across the repeated measurements. This allows us to estimate several variance components: variation between babies, (i.e. σ^2_{babies}), reflecting the heterogeneity in the sample, and the rater variance (i.e. σ^2_{rater}), representing the systematic difference between the raters. Additionally, there is a residual error (also called a random error; $\sigma^2_{\text{residual}}$), capturing the remaining unexplained variance. The residual error includes random fluctuation, but also any other sources of variation that were not explicitly controlled for in the study design (i.e. no predefined measurement conditions).

In our example, individual variance components were: σ^2_{babies} is 2.56, σ^2_{rater} is 5.24, and $\sigma^2_{\text{residual}}$ is 1.02.

The $\text{ICC}_{\text{agreement}}$ is calculated by the formula:

ICC_{agreement} = $\frac{\sigma^2_{babies}}{\sigma^2_{babies}+\sigma^2_{rater}+\sigma^2_{residual}}$

And the SEM$_{\text{agreement}}$ is calculated by the formula:

SEM$_{\text{agreement}}$ = $\sqrt{}(\sigma^2_{rater} + \sigma^2_{residual})$

For our example this results in an ICC$_{\text{agreement}}$ of 0.29 (0.019 – 0.761), and a SEM$_{\text{agreement}}$ of 2.5.

Reporting and interpretation

We recommend reporting the ICC and its 95% confidence interval, the SEM, and the different variance components, that is the variance components of the individuals (between-person), the variance components of the source(s) of variation varied across the repeated measurements, and the residual variance.

A commonly used threshold for sufficient reliability is ICC > 0.70 (6). The ICC in our example score is 0.29, indicating insufficient reliability. This low ICC suggests that the measurement instrument is not capable of reliably distinguishing between babies within this population. In such cases, the instrument is not suitable for use in decision-making or evaluation. The 95% CI ranges between 0.02 and 0.76, which indicates an uncertain point estimate of the ICC (due to a small sample size).

The SEM provides insight into the precision of an observed score. In our example, the SEM is 2.5. This means that if the observed score of a baby would be 5, their true score is likely (that is, we are 95% confident) to lie within the range of 5 +/- 1.96 * SEM, i.e. between 0.1 and 9.9. This range covers nearly the entire scale, indicating substantial measurement error and an imprecise observed score. In such cases, the observed scores offer little confidence about a baby's true score on the measured outcome.

In the light of these results, the question arises: how can this measurement be improved? If a particular source has a large variance component, it indicates a systematic influences on the measurement. By examining the variance components of the different sources of variance in the example study, we see that the σ^2_{rater} is larger than the residual error ($\sigma^2_{residual}$) and the individual variance (σ^2_{babies}). This suggests that the 'rater' is a substantial source of measurement error. To improve the reliability of the instrument, it is important to standardize this source, for example by improving the instructions to the raters or providing better training. Qualitative research (e.g. focusing on the content validity of the measurement instrument(7)) can results in better understanding how to improve the measurement instrument. Such improvements effectively result in a revised version of the instrument, and a new reliability study is needed to evaluate whether these changes have enhanced its performance.

Conclusions

Although the results in our example may seem extreme, substantial measurement error is often a common challenge in practice. When reducing measurement error through design improvements is not feasible, a practical alternative is to conduct repeated measurements and use their average as the final score, provided this approach is clinically and ethical feasible. Regardless of the approach, it is important to always consider measurement error when interpreting results.

Further reading:

Mokkink LB, Eekhout I, Boers M, van der Vleuten CPM, de Vet HCW. Studies on Reliability and Measurement Error of Measurements in Medicine - From Design to Statistics Explained for Medical Researchers. Patient Relat Outcome Meas. 2023;14:193-212.

This article explains how to formulating research questions for a study on reliability and measurement error, Generalizability theory, how to choose between the various ICC and SEM formulas and it provide syntaxes for calculating ICCs and SEMs in SPSS and R.

Buijs GS, ter Wee MA, Klein C, Mokkink LB, Dobbe JGG, Maas M, Schafroth MU, Streekstra GJ, Blankevoort L, Kievit AJ. Operator variation in applying a knee loading device for evaluation of tibial component loosening in total knee arthroplasty. Clinical Biomechanics. 2025;126:106531.

This article shows how a reliability study can be designed using the components of measurement instruments.

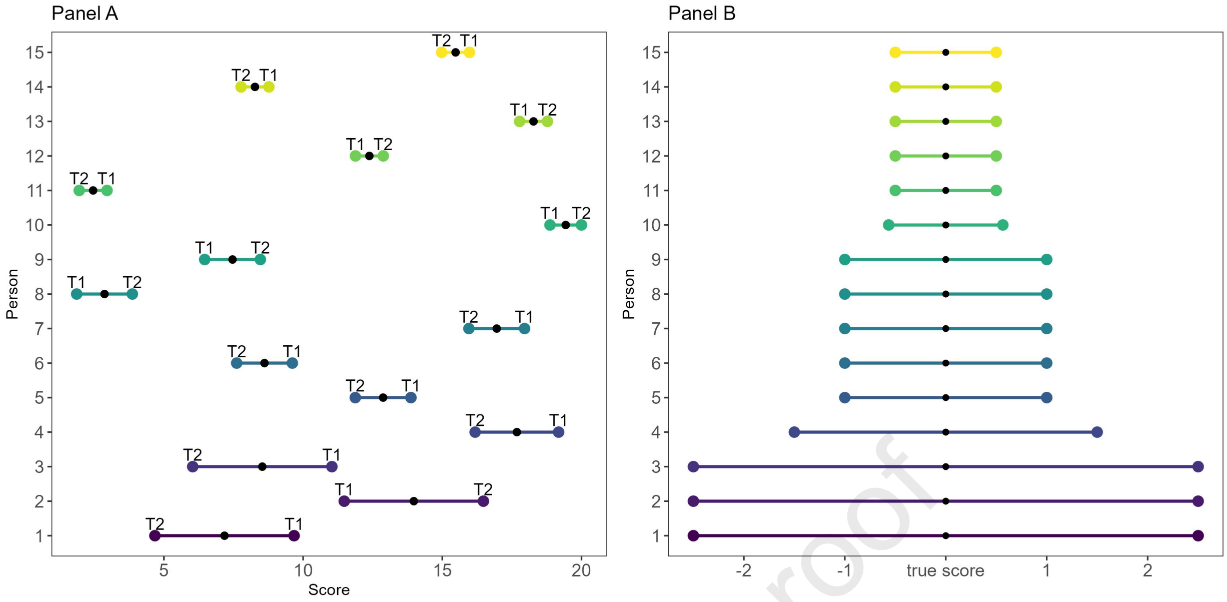
References

1. Streiner DL, Norman G. Health Measurement Scales. A practical guide to their development and use. 4th edition ed. New York: Oxford University Press; 2008.
2. Mokkink LB, Herbelet S, Terwee CB, Boers M. At the CORE of measurement - What do you want to measure? And how do you want to measure it? A COSMIN Perspective. J Clin Epidemiol. 2025:111836.
3. Mokkink LB, Eekhout I, Boers M, van der Vleuten CPM, de Vet HCW. Studies on Reliability and Measurement Error of Measurements in Medicine - From Design to Statistics Explained for Medical Researchers. Patient Relat Outcome Meas. 2023;14:193–212.
4. Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, 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).
5. Ambuel B, Hamlett KW, Marx CM, Blumer JL. Assessing distress in pediatric intensive care environments: the COMFORT scale. J Pediatr Psychol. 1992;17(1):95–109.
6. Mokkink LB, Elsman EBM, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. Qual Life Res. 2024;33(11):2929-39.
7. Mokkink LB, Herbelet S, Tuinman PR, Terwee CB. Content validity: judging the relevance, comprehensiveness, and comprehensibility of an outcome measurement instrument - a COSMIN perspective. J Clin Epidemiol. 2025;185:111879.

4 raters

6 babies	1	2	3	4	Mean
1	9	2	5	8	6.0
2	6	1	3	2	4.0
3	8	4	6	8	6.5
4	7	1	2	6	4.0
5	10	5	6	9	7.5
6	6	2	4	7	4.75
	7.7	2.5	4.3	6.7	5.46

	4 raters				
6 babies	1	2	3	4	Mean
1	9	2	5	8	6.0
2	6	1	3	2	4.0
3	8	4	6	8	6.5
4	7	1	2	6	4.0
5	10	5	6	9	7.5
6	6	2	4	7	4.75
Mean	7.7	2.5	4.3	6.7	5.46


Panel A
Person
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
T2 T1
T2 T1
T1 T2
T1 T2
T2 T1
T1 T2
T1 T2
T1 T2
T2 T1
T2 T1
T2 T1
T2 T1
T2 T1
T1 T2
T2 T1
5
10
15
20
Score
Panel B
Person
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
-2
-1
true score
1
2

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: