KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: June 7, 2016
Revision received: October 12, 2016
Accepted: October 29, 2016
OnlineFirst: November 23, 2016

www.estp.com.tr
DOI 10.12738/estp.2017.1.0357 • February 2017 • 17(1) • 217–236

Research Article

An Alternative Method Used in Evaluating Agreement among Repeat Measurements by Two Raters in Education

Semra Erdoğan[1]
Mersin University

Gülhan Orekici Temel[2]
Mersin University

Hüseyin Selvi[3]
Mersin University

Irem Ersöz Kaya[4]
Mersin University

Abstract

Taking more than one measurement of the same variable also hosts the possibility of contamination from error sources, both singly and in combination as a result of interactions. Therefore, although the internal consistency of scores received from measurement tools is examined by itself, it is necessary to ensure inter-rater or intra-rater agreement in order to provide reliability. The biggest problem while conducting agreement analyses for obtained measurement results is deciding which statistical method to use. Inconsistency between measurements obtained by different methods over the same individual has been suggested as being similar to inconsistency between repeated measurements obtained by the same methods over the same individual. For this purpose, a new approach is proposed for estimating and defining an agreement coefficient between raters or methods. Based on this goal, an answer to the following question is sought: When the dependent/predicted variable has two categories (such as successful-unsuccessful, sick-healthy, positive-negative, exists-does not exist, etc.) and there are two raters who each undertake repeat measures, how does the method work in terms of disagreement functions and individual-agreement coefficient, as well as for different numbers of repeat measures and different sample sizes?

Keywords

Reliability • Methods comparison • Disagreement function • Inter-method agreement • Intra-method agreement • Individual agreement coefficient

1 **Correspondence to:** Semra Erdoğan (PhD), Department of Biostatistics and Medical Informatics, Faculty of Medicine, Mersin University, Mersin 33343 Turkey. Email: semraerdogann@gmail.com

2 Department of Biostatistics and Medical Informatics, Faculty of Medicine, Mersin University, Mersin 33343 Turkey. Email: gulhan_orekici@hotmail.com

3 Department of Medical Education, Faculty of Medicine, Mersin University, Mersin 33343 Turkey. Email: hsyn_selvi@yahoo.com.tr

4 Faculty of Technology, Department of Software Engineering, Mersin University, Mersin 33343 Turkey. Email: iremersoz@gmail.com

Citation: Erdoğan, S., Orekici Temel, G., Selvi, H., & Ersöz Kaya, I. (2017). An alternative method used in evaluating agreement among repeat measurements by two raters in education. *Educational Sciences: Theory & Practice*, *17*, 217–236. http://dx.doi.org/10.12738/estp.2017.1.0357

Criteria regarding whether a concept, theory, design, or even a whole discipline is actually scientific vary from one field to another. However, there are invariant criteria for all fields such as the abilities to *observe*, *measure*, *transmit*, *repeat*, *reproduce*, *verify*, and *falsify*. These criteria allow different scientists to monitor or determine whether theories or designs related to a specific case or concept are valid and reliable. These criteria even prepare the conditions for *measurability* and *reproducibility*, which allow the opportunity for further research, as well as protect scientists from being trapped in prejudices. One of the essential criteria for science is measurability. Hence, advances in science can be claimed to develop in parallel with advances in measurement science (Erdoğan, 2011; Karakaş, 1988).

In this respect, one can argue that compared to other science disciplines, advances in scientific fields where the investigated qualities can be directly measured are quicker, and therefore, quality measurements are comparatively easier to undertake.

The situation is completely different in science disciplines such as education and psychology where the investigated qualities cannot be directly measured. In these disciplines, one attempts to predict the conditions of the related quality based on responses provided to specific stimuli; in other words, measurement is indirect (Gulliksen, 1950). However, although indirect measurement makes it possible to measure qualities that cannot be directly measured, it may also radically increase the potential error sources involved in the process. While the direction and amount of these error sources are sometimes apparent and can be identified (i.e., fixed, systematic change), sometimes they cannot (random error). This fact makes it rather hard and complex to undertake quality measurement in sciences where indirect measurement is a necessity because error sources with unidentified directions and amounts damage data reliability and impair the accuracy of the procedural comparisons that use these measurements.

Scientists have developed various methods and techniques for examining reliability related to different error sources. Although these methods and techniques can be found under different classifications in different resources, they are simply classified by Crocker and Algina (1986) as methods based on multiple applications (such as equivalent forms and test-retest methods) or single application (split-half method or item-covariance-based methods). As the classification shows, some methods and techniques calculate error sources by using a single application to examine data reliability, whereas others rely on repeated measures, or scoring by multiple raters.

In scientific disciplines such as education and psychology, where the investigated qualities cannot be directly measured, written, oral, and kinetic exams that require scoring by more than one rater; procedural comparisons that compare new methods and techniques developed according to scientific and technological advances;

longitudinal studies; scale adaptation-development studies; and so on, are common. Reliability is the weakest link in studies where it is necessary to collect data from the same variable using different measurement tools, or to collect data from the same variable by using the same tool at different intervals (Güler & Gelbal, 2010). As a matter of fact, taking more than one measurement from the same variable also hosts the possibility of contamination from error sources as a result of interaction, both singly and in combination. Therefore, although the internal consistency of scores received from measurement tools is examined in itself, it is necessary to ensure inter-rater and intra-rater agreement in order to provide reliability (Güler & Gelbal, 2010; Lin, Hedayet, & Wu, 2012). In this context, agreement means similarities among measurements obtained by different inter-raters/methods. Disagreement can be defined as the difference among measurements (Barnhart, Song, & Haber, 2005).

The biggest problem while conducting agreement analyses on obtained measurement results is to decide which statistical method to use. Many agreement studies are seen to use classical statistical methods such as Pearson's correlation coefficient, regression analysis, or t-tests for dependent groups. Additionally, classical statistics methods like chi-squared test and Cohen kappa statistics are seen to be used widely in categorical measurements. Stralen, Dekker, Zoccali, & Jager (2012) have revealed that systematic error is disregarded when the Pearson correlation coefficient is used, when agreement between two continuous measurement methods is tested, or when the effect of prevalence and bias is not counteracted as a result of using Cohen kappa correlation while testing agreement between two categorical measurement methods and disregarding the different weight calculations for inconsistent cells (Stralen et al., 2012). Starting from this point of view, alternative methods have been developed apart from the known classical methods. As a matter of fact, the literature includes many different methods for examining agreement between two measurements (Lin et al., 2012). Intraclass correlation coefficient (ICC), Concordance correlation coefficient (CCC; developed by Lin), Bland&Altman plot developed by Bland and Altman, Type II regression methods (such as Deming regression method, Weighted Deming regression method and Passing-BaBlok methods), Scott's p statistics, Cohen's kappa statistics, G-index, Gwet's AC1 statistics, Fleiss kappa statistics, Krippendorf Alpha coefficient, Weighted kappa statistics, Kendall's W coefficient can be cited as common examples by taking their premises into consideration, such as measurement level, type of distribution, number of raters, and so on. Bland-Altman plots and concordance correlation coefficients can be argued as more common because they take random error and systematic change into consideration (Atkinson & Nevill, 1997; Erdoğan & Kanık, 2005; Işıkhan, Kılıçkap, & Alpar, 2013; Kanık & Erdoğan, 2004; Kanık, Erdoğan, & Orekici Temel, 2012; Kanık, Orekici Temel, & Ersöz, 2010). However, classical methods for the categorical variable mentioned above are not used in agreement analysis under conditions where each rater has multiple readings.

A new approach has been suggested by Haber and Barnhart (2007) that can evaluate the harmony both between and within individual raters in cases where there are more than one measurement or observer. The individual agreement coefficient (CIA) is an agreement statistic in numerical (discrete, continuous) or categorical (nominal, ordinal, ratio, interval) structures that can be used in situations where single or more than one measurement have been taken. Unlike other agreement statistics, CIA is an approach that can evaluate the agreement of each rater individually or together. Haber and Barnhart (2007) proposed that disagreement between observations obtained from different methods is similar to the disagreement between observations made by the same method. In other words, replacing or interchanging one method with another does not substantially increase the disagreement between measurements obtained from the same individuals. Based on this information, they proposed a new coefficient of agreement that compares the disagreement between measurements made by different methods to the disagreement between replicated measurements made by the same method. This approach is the coefficient of individual agreement, described as a special disagreement function that can be used in cases where repeated measures are also continuous and categorical variables (Barnhart, Lokhnygina, Kosinki, & Haber, 2007; Haber & Barnhart, 2007; Haber, Gao, & Barnhart, 2007; Pan, Gao, Haber, & Barnhart, 2010; Pan, Haber, & Barnhart, 2011).

Material and Methods

Coefficient of Individual Agreement (CIA)

In order to define a coefficient of agreement, one first must decide how to quantify agreement between two methods or raters. In cases where there are only two raters, the rater's measurements are indicated by X or Y. Replicated measurements for the first rater (X) are indicated by X and X'; the disagreement function between two measurements is $G(X, X')$; two replicated measurements for the second rater (Y) are indicated by Y and Y', and the disagreement function between these two measurements is defined as $G(Y, Y')$. The quantity of disagreement between measurements obtained from the same individuals is presented by $G(X, Y)$. This disagreement function is assumed to be $G(X, Y) \geq 0$ and $G(X, X) = 0$ (Haber & Barnhart, 2007; Haber et al., 2007).

N denotes the number of subjects included in the study and is stated as $i = 1, 2, \ldots N$. X_{ik} denotes k-replicated measurement values of rater X obtained from subject i ($k = 1, 2, \ldots K_i$); Y_{il} denotes l replicated observations obtained from subject i ($l = 1, 2, \ldots L_i$). For a structure with two results, positive case values of X and Y will be equal to 1 and negative case values of X and Y will be equal to 0. For a positive case of rater X, $P(X_{ik} = 1) = \pi_i$ ($k = 1, \ldots K_i$); for a positive case of rater Y, $P(Y_{il} = 1) = \lambda_i$, ($l = 1, \ldots$

L_i). Disagreement functions specific to individuals (David & Skene, 1979; Gao, Pan, & Haber, 2012; Haber et al., 2007; Pan et al., 2010; Pan et al., 2011) are denoted as:

$$\begin{aligned} G_i(X,Y) &= P(X_{ik} \neq Y_{il}/i) \\ &= Pr(X_{ik}=1, Y_{il}=0/i) + Pr(X_{ik}=0, Y_{il}=1/i) \\ &= \pi_i(1-\lambda_i) + (1-\pi_i)\lambda_i \\ &= \pi_i + \lambda_i - 2\pi_i\lambda_i \end{aligned} \quad (1)$$

$$\begin{aligned} G_i(X,X') &= P(X_{ik} \neq X_{ik'}/i; k \neq k') \\ &= 2\pi_i(1-\pi_i) \end{aligned} \quad (2)$$

$$\begin{aligned} G_i(Y,Y') &= P(Y_{il} \neq Y_{il'}/i; l \neq l') \\ &= 2\lambda_i(1-\lambda_i) \end{aligned} \quad (3)$$

Total disagreement function, G, and the mean of disagreement functions for all individuals, G_i, are formulated as:

$$\overline{G} = 1/N \sum_{i=1}^{N} G_i \quad (4)$$

Haber and Barnhart (2007) assessed cases where there is one reference rater, in addition to those with none, while evaluating agreement between raters. Based on this, if no rater is considered a reference, CIA is formulated as in Equation 5. This equation's numerator provides the average disagreements between two repeated measurements taken from the same individuals by the same rater; its denominator provides disagreement between raters X and Y (David & Skene, 1979; Gao et al., 2012; Haber & Barnhart, 2007; Haber et al., 2007; Pan et al., 2010).

$$\psi^N = \frac{\left(\overline{G}(X,X') + \overline{G}(Y,Y')\right)}{\overline{G}(X,Y)} = \frac{\sum_i \left[\pi_i(1-\pi_i) + \lambda_i(1-\lambda_i)\right]}{\sum_i \left(\pi_i + \lambda_i - 2\pi_i\lambda_i\right)} \quad (5)$$

If measurements of an experienced or a reliable rater are to be compared with measurements of a new rater, the rater X should be provided as reference and CIA should be expressed as in Equation 6 when this new rater is compared with the measurements of Y. The equation's numerator provides the disagreement between repeated measurements of the reference rater; its denominator provides disagreement between raters X and Y (David & Skene, 1979; Haber et al., 2007; Pan et al., 2011).

$$\psi^R = \frac{\overline{G}(X,X')}{\overline{G}(X,Y)} = \frac{2\sum_i \pi_i(1-\pi_i)}{\sum_i(\pi_i + \lambda_i - 2\pi_i\lambda_i)} \quad (6)$$

Estimation

Probability for positive readings of measurements values (X_{ik}) of rater X repeated k. times and taken over i subjects is shown with $\hat{\pi}_i$; probability for positive readings of measurements values (Y_{il}) of rater Y repeated l times taken over i subjects is shown with $\hat{\lambda}_i$. There, the classification probabilities are estimated around $\hat{\pi}_i = T_i / K_i$ and $\hat{\lambda}_i = U_i / L_i$. Here, T_i shows the total numbers of positive readings, l, in repeated measurements that belong to rater X for i subjects; U_i shows the total number of positive readings, 1, in repeated measurements that belong to rater Y for i subjects. The unbiased estimators of the subject-specific disagreement functions are calculated as shown in Equations 7, 8, and 9 (Gao, 2010; Haber et al., 2007; Pan et al., 2011):

$$\hat{G}_i(X,Y) = \hat{\pi}_i + \hat{\lambda}_i - 2\hat{\pi}_i\hat{\lambda}_i \quad (7)$$

$$\hat{G}_i(X,X') = 2K_i\hat{\pi}_i(1-\hat{\pi}_i)/(K_i - 1) \quad (8)$$

$$\hat{G}_i(Y,Y') = 2L_i\hat{\lambda}_i(1-\hat{\lambda}_i)/(L_i - 1) \quad (9)$$

Estimations of the overall G are calculated as shown by Equations 10, 11, and 12:

$$\overline{\hat{G}}(X,Y) = \overline{G_i(X,Y)} \quad (10)$$

$$\overline{\hat{G}}(X,X') = \overline{G_i(X,X')} \quad (11)$$

$$\overline{\hat{G}}(Y,Y') = \overline{G_i(Y,Y')} \quad (12)$$

Table 1 summarizes the sample case where disagreement values may occur when two repeated measures are taken by two raters for a situation during dichotomous scoring. As X_{ik} and Y_{il} only take values 0 or 1, there are only two repeated measurements for each individual. Thus, probabilities $\hat{\pi}_i$ and $\hat{\lambda}_i$ for their positive readings will be 0.0, 0.5, or 1.0. If two measurements that were taken by rater X are equal ($X_{i1} = X_{i2}$), the value $\hat{\pi}_i$ will be 1 (2/2) or 0 (0/2); the disagreement value within rater X will be $\hat{G}_i(X, X') = 0$. The same thing also applies to rater Y. If measurements taken from raters are $X_{i1} \neq X_{i2}$ and $Y_{i1} \neq Y_{i2}$, then $\hat{\pi}_i$ and $\hat{\lambda}_i$ will be 0.5 (1/2), $\hat{G}_i(X, X') = 1$ and $\hat{G}_i(Y, Y') = 1$. If repeated measurements taken from both raters are equal to each other ($X_{i1} = X_{i2} = Y_{i1} = Y_{i2}$), $\hat{\pi}_i$ and $\hat{\lambda}_i$ will be 0 or 1, and $\hat{G}_i(X, Y') = 0$; this shows perfect agreement between raters X and Y (Table 1; Gao, 2010). The literature states that a value of 0.80 or higher is acceptable for agreement (Haber & Barnhart, 2007; Pan et al., 2011).

Table 1

Parametric Approach for Estimating Disagreement Functions for $K_i = L_i = 2$

$X_{i1}, X_{i2}, Y_{i1}, Y_{i2}$	$\hat{\pi}_i$	$\hat{\lambda}_i$	$\hat{G}_i(X, X')$	$\hat{G}_i(Y, Y')$	$\hat{G}_i(Y, Y')$
$X_{i1} = X_{i2} = Y_{i1} = Y_{i2}$	0 (0/2) or 1(2/2)	0 or 1	0	0	0
$X_{i1} = X_{i2} \neq Y_{i1} = Y_{i2}$	0 or 1	0 or 1	0	0	1
$X_{i1} \neq X_{i2}; Y_{i1} = Y_{i2}$	0.5 (1/2)	0 or 1	1	0	0.5
$X_{i1} = X_{i2}; Y_{i1} \neq Y_{i2}$	0 or 1	0.5	0	1	0.5
$X_{i1} \neq X_{i2}; Y_{i1} \neq Y_{i2}$	0.5	0.5	1	1	0.5

T_i: Agreeable/matching measurement of rater X,

K_i: Total numbers of measurements for rater X,

U_i: Agreeable/matching measurement of rater Y,

L_i: Total number of measurements for rater Y,

The literature presents various methods developed for use in different conditions that require the investigation of agreement between data and scoring reliability. CIA is one of these methods. All of these methods are known to have strengths and weaknesses.

Goodwin (2001) compared the values obtained from different methods (i.e., Pearson product-moment correlation coefficient, simple agreement coefficient, kappa statistics, and generalizability theory[G Theory]) to calculate scoring reliability and agreement between raters, and found that G Theory, which includes more than one error source in calculation at the same time, is more advantageous in these types of studies.

Similarly, Liao, Hunt, and Chen (2010) examined different methods (Pearson product-moment correlation coefficient, Kendall's *t* and *w* coefficients, Spearman's *p* coefficient and $r_{WG(I)}$ and $r_{WG(J)}$ coefficients) proposed by James (1982) for calculating inter-rater agreement and scoring reliability during performance assessment by using a simulated data set that included different agreement and correlation conditions. They found that the values obtained with these methods were relatively significantly different from each other.

In their study, Goodwin and Goodwin (1991) compared the different methods (G Theory, simple agreement coefficient, calculating correlation coefficient) used for calculating scoring reliability on a simulated data set. According to the results of the study, values obtained with the help of G Theory provided higher quality results when examining reliability.

Güler and Taşdelen Teker (2015) examined inter-rater reliability in essay-type items obtained by different methods (correlation, comparison of means, percentage of agreement, and G Theory) and concluded that the highest prediction of reliability

(0.90) was possible using G Theory. The value predicted through percentage of inter-rater agreement was found to be 58.9%. The simple correlational value was found to be positive (0.74) and high. Study results suggested using G Theory when examining scoring reliability as it has the ability to address many error sources at the same time, even though it seems to be the most complex method.

Kan (2005) examined the effect of using scoring rubrics and answer keys while grading written exams with variable scoring reliability, and found significant differences between the mean scores provided by teachers depending on whether they had or had not used scoring rubrics or answer keys at different times. They also concluded that the time lapse between grading affected the consistency of scoring.

A review of the literature shows that various methods have been developed to investigate scoring reliability and inter-rater agreement, along with many studies for identifying the strong and weak points of these methods. CIA, proposed by Haber and Barnhart (2007), is a somewhat new method that examines agreement between dates, and extensive studies do not exist in the literature about the functionality of this method. Therefore, this study aims to present the functions of CIA as proposed by Haber and Barnhart (2007) under different conditions.

Based on this goal, an answer to the following question is sought: When the dependent/predicted variable has two categories (such as successful-unsuccessful, sick-healthy, positive-negative, exists-does not exist, etc.) and there are two raters who each undertake repeated measures, how does the method work in terms of disagreement functions and CIA, as well as for different numbers of repeated measures and different sample sizes?

Study Data

Data used in the research were obtained using the simulation technique with the help of the package program, MatLab 7.0.

Procedure

During the data generation phase, two different raters, *X* and *Y*, took repeated measurements from each individual two, three, four, and five times. Conditions were examined where the measurements were scored using binary scoring (0-1) and the raters were dependent/independent. Here, rater independence means that agreement between a rater's repeated measures were taken into consideration, but inter-rater agreements were not investigated. Rater dependence in this context means when both agreement between a rater's repeated measures and inter-rater agreements are taken into consideration.

In this sense, data were first generated for raters who were independent of each other.

a) Measurements were taken from raters *X* and *Y* two, three, four, and five times (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}; Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5}).

b) There is a perfect agreement among the repeat measurements of each rater (X_{ik} = 0.90; Y_{il} = 0.90).

c) There is no agreement among the repeat measurements of either rater (X_{ik} = 0.10; Y_{il} = 0.10).

d) There is a perfect agreement among the repeat measurements of one rater and no agreement among the repeat measurements of the other rater (cases where the agreement between repeated measurements was 0.90 for rater *X*, where the agreement among repeated measurements was 0.10 for the rater *Y* (X_{ik} = 0.90; Y_{il} = 0.10), and its exact opposite (X_{ik} = 0.10; Y_{il} = 0.90).

e) There is moderate agreement among the repeat measurements of each rater (X_{ik} = 0.50; Y_{il} = 0.50).

For the case where raters were dependent on each other, each rater's repeated measures and the situation between the raters were also examined. Accordingly:

a) Measurements were taken from raters *X* and *Y* two, three, four, and five times (X_{i1}, $X_{i2,}$ X_{i3}, X_{i4}, X_{i5}; Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5}).

b) There is perfect agreement among all repeated measurements for each rater, and perfect agreement between raters as well (X_{ik} = 0.90; Y_{il} = 0.90; $X_{ik} = Y_{il}$).

c) There is perfect agreement among all repeated measurements for each rater but no agreement between raters (X_{ik} = 0.90 and Y_{il} = 0.90; $X_{ik} \neq Y_{il}$).

d) There is no agreement among the repeated measurements of either rater, but there is perfect agreement between raters (X_{ik} = 0.10, Y_{il} = 0.10, $X_{ik} = Y_{il}$).

e) There is no agreement among the repeated measurements of either rater nor between them (X_{ik} = 0.10, Y_{il} = 0.10, $X_{ik} \neq Y_{il}$).

f) There is moderate agreement among all repeated measurements of each rater, as well as perfect agreement between raters (X_{ik} = 0.50, Y_{il} = 0.50, $X_{ik} = Y_{il}$).

g) There is moderate agreement among all repeated measurements for each rater but no agreement between raters (X_{ik} = 0.50, Y_{il} = 0.50, $X_{ik} \neq Y_{il}$).

Considering the circumstances, the cases for different sample sizes (30, 100, and 200) were used for the case where disagreement functions and CIAs were also calculated. All operations were repeated 1,000 times and their averages were found.

Data Analysis

During data analysis, disagreement functions and individual agreement coefficients were first calculated for each data set. The next phase included calculating CIA means for each combination and standard deviation. The obtained values are reported in Tables 2 through 8.

Results

Cases Where Raters are Independent of Each Other

Disagreement functions and CIAs belonging to the case of perfect agreement among repeated measurements for both raters are presented in Table 2; disagreement functions and CIAs pertaining to cases of no agreement among repeated measurements for both raters are presented in Table 3. Disagreement functions and CIAs belonging to the case where agreement among repeated measurements is 0.90 for rater *X* and agreement among repeated measurements is 0.10 for rater *Y* are presented in Table 4. Disagreement functions and CIAs pertaining to cases where agreement among repeated measurements is 0.10 for rater *X* and agreement among repeated measurements is 0.90 for rater *Y* are presented in Table 5. Disagreement functions and CIAs for the case of moderate agreement (0.50) among repeated measurements for both raters are presented in Table 6.

Having examined cases where there is perfect agreement among repeated measurements for both raters *X* and *Y* where raters *X* and *Y* are independent of each other, disagreement functions and CIAs (Ψ^N and Ψ^R) are determined to be unaffected by sample size or the number of repeated measurements by raters. Having perfect agreement among repeated measurements by each rater does not require having perfect agreement between raters. Having perfect agreement among each rater's repeated measurements is not a sign of agreement between raters because the measurement results of raters do not depend on one another. When multiple measurements of each rater are in agreement with each other, disagreement functions belonging to the referenced rater are acknowledged as near 0, and the disagreement function belonging to two raters is closer to 0.5, depending on chance, which is expected.

Table 2

Cases where Agreement Among Repeated Measurements for Raters X *and* Y *is 0.90 (*X_{ik} *= 0.90,* Y_{i1} *= 0.90).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.050 ± 0.000	0.050 ± 0.000	0.501 ± 0.087	0.103 ± 0.020	0.103 ± 0.020
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.499 ± 0.089	0.092 ± 0.019	0.092 ± 0.019
	4	4	0.042 ± 0.004	0.042 ± 0.004	0.499 ± 0.092	0.087 ± 0.019	0.087 ± 0.020
	5	5	0.042 ± 0.004	0.042 ± 0.005	0.497 ± 0.086	0.087 ± 0.019	0.087 ± 0.020
100	2	2	0.050 ± 0.000	0.050 ± 0.000	0.500 ± 0.049	0.101 ± 0.010	0.101 ± 0.010
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.500 ± 0.048	0.090 ± 0.009	0.090 ± 0.009
	4	4	0.042 ± 0.002	0.042 ± 0.002	0.499 ± 0.047	0.085 ± 0.009	0.085 ± 0.009
	5	5	0.042 ± 0.002	0.042 ± 0.002	0.501 ± 0.047	0.085 ± 0.009	0.085 ± 0.010
200	2	2	0.050 ± 0.000	0.050 ± 0.000	0.499 ± 0.034	0.101 ± 0.007	0.101 ± 0.007
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.499 ± 0.033	0.089 ± 0.006	0.089 ± 0.006
	4	4	0.042 ± 0.001	0.042 ± 0.001	0.499 ± 0.034	0.085 ± 0.006	0.085 ± 0.007
	5	5	0.042 ± 0.002	0.042 ± 0.002	0.499 ± 0.034	0.085 ± 0.006	0.085 ± 0.007

Table 3

Cases where Agreement Among Repeated Measurements for Raters X *and* Y *is 0.10 (*X_{ik} *= 0.10,* Y_{i1} *= 0.10).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.450 ± 0.000	0.450 ± 0.000	0.500 ± 0.030	0.904 ± 0.054	0.904 ± 0.054
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.500 ± 0.032	0.804 ± 0.051	0.804 ± 0.051
	4	4	0.379 ± 0.011	0.380 ± 0.011	0.501 ± 0.031	0.761± 0.051	0.761 ± 0.052
	5	5	0.378 ± 0.014	0.378 ± 0.014	0.499 ± 0.033	0.760± 0.055	0.760 ± 0.058
100	2	2	0.450 ± 0.000	0.450 ± 0.000	0.501 ± 0.015	0.900 ± 0.028	0.900 ± 0.028
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.500 ± 0.017	0.801 ± 0.027	0.801 ± 0.027
	4	4	0.380 ± 0.006	0.380 ± 0.006	0.499 ± 0.017	0.762± 0.028	0.762± 0.028
	5	5	0.378 ± 0.007	0.378 ± 0.007	0.500 ± 0.018	0.757 ± 0.029	0.757 ± 0.031
200	2	2	0.450 ± 0.000	0.450 ± 0.000	0.501 ± 0.011	0.899 ± 0.020	0.899 ± 0.020
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.500 ± 0.012	0.800 ± 0.019	0.800 ± 0.019
	4	4	0.380 ± 0.004	0.380 ± 0.004	0.500 ± 0.012	0.761 ± 0.020	0.761 ± 0.021
	5	5	0.378 ± 0.005	0.378 ± 0.005	0.500 ± 0.012	0.756 ± 0.020	0.756 ± 0.021

Having examined cases with no perfect agreement among the repeated measurements of both raters *X* and *Y*, disagreement functions *G*(*X*, *X'*) and *G*(*Y*, *Y'*) and CIAs were observed to be unaffected by sample size; however, as the number of repeated measurements by each rater increases, these values drop, albeit only a little. The disagreement function *G*(*X*, *Y*) was observed to have an agreement of 0.50 in every case that depended on chance. In cases where disagreement functions *G*(*X*, *X'*) and *G*(*Y*, *Y'*) are around 0.50 and the value of the disagreement function *G*(*X*, *Y*) is 0.50, coefficients of individual agreement are expected to have a value closer to 1. CIAs with value between 0.76 and 0.90 can be said to have perfect agreement, even when there is no agreement between raters (Table 3).

Table 4

Cases where Agreement Among Repeated Measurements of Rater X *is 0.90 and Agreement Among Repeated Measurements of Rater* Y *is 0.10 (X_{ik} = 0.90, Y_{i1} = 0.10).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.050 ± 0.000	0.450 ± 0.000	0.500 ± 0.029	0.502 ± 0.029	0.100 ± 0.006
	3	3	0.044 ± 0.000	0.400 ± 0.000	0.503 ± 0.040	0.445 ± 0.036	0.089 ± 0.007
	4	4	0.042 ± 0.003	0.380 ± 0.011	0.500 ± 0.043	0.425 ± 0.038	0.085 ± 0.010
	5	5	0.042 ± 0.005	0.378 ± 0.014	0.501 ± 0.045	0.423 ± 0.041	0.085 ± 0.012
100	2	2	0.050 ± 0.000	0.450 ± 0.000	0.500 ± 0.015	0.501 ± 0.015	0.100 ± 0.003
	3	3	0.044 ± 0.000	0.400 ± 0.000	0.500 ± 0.022	0.446 ± 0.020	0.089 ± 0.004
	4	4	0.042 ± 0.002	0.380 ± 0.006	0.501 ± 0.024	0.422 ± 0.021	0.085 ± 0.006
	5	5	0.042 ± 0.002	0.378 ± 0.007	0.500 ± 0.024	0.421 ± 0.022	0.084 ± 0.006
200	2	2	0.050 ± 0.000	0.450 ± 0.000	0.500 ± 0.011	0.501 ± 0.011	0.100 ± 0.002
	3	3	0.044 ± 0.000	0.400 ± 0.000	0.500 ± 0.015	0.445 ± 0.013	0.089 ± 0.003
	4	4	0.042 ± 0.001	0.380 ± 0.004	0.501 ± 0.017	0.422 ± 0.015	0.084 ± 0.004
	5	5	0.042 ± 0.002	0.378 ± 0.005	0.501 ± 0.017	0.420 ± 0.016	0.084 ± 0.005

Table 5

Cases where Agreement Among Repeated Measurements of Rater X *is 0.10, and Agreement Among Repeated Measurements of Rater* Y *is 0.90 (X_{ik} = 0.10 Y_{i1} = 0.90).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.450 ± 0.000	0.050 ± 0.000	0.500 ± 0.029	0.502 ± 0.023	0.903 ± 0.053
	3	3	0.400 ± 0.000	0.044 ± 0.000	0.498 ± 0.042	0.450 ± 0.038	0.809 ± 0.068
	4	4	0.380 ± 0.011	0.042 ± 0.004	0.499 ± 0.044	0.426 ± 0.040	0.767 ± 0.072
	5	5	0.377 ± 0.014	0.042 ± 0.005	0.501 ± 0.045	0.422 ± 0.041	0.760 ± 0.074
100	2	2	0.450 ± 0.000	0.050 ± 0.000	0.500 ± 0.016	0.501 ± 0.016	0.901 ± 0.028
	3	3	0.400 ± 0.000	0.044 ± 0.000	0.499 ± 0.022	0.447± 0.020	0.804 ± 0.036
	4	4	0.380 ± 0.006	0.042 ± 0.002	0.500 ± 0.023	0.423 ± 0.021	0.761 ± 0.038
	5	5	0.378 ± 0.007	0.042 ± 0.003	0.500 ± 0.024	0.421 ± 0.022	0.760 ± 0.039
200	2	2	0.450 ± 0.000	0.050 ± 0.000	0.500 ± 0.011	0.500 ± 0.011	0.900 ± 0.021
	3	3	0.400 ± 0.000	0.044 ± 0.000	0.500 ± 0.015	0.445 ± 0.014	0.801 ± 0.025
	4	4	0.380 ± 0.004	0.042 ± 0.001	0.500 ± 0.017	0.422 ± 0.015	0.760 ± 0.027
	5	5	0.378 ± 0.005	0.042 ± 0.002	0.500 ± 0.017	0.420 ± 0.015	0.757 ± 0.027

In case of perfect agreement among repeated measurements by one of the raters (0.90) and no agreement among repeated measurements by the other rater (0.10; see Tables 4 and 5), disagreement-function values within each rater and CIA (Ψ^N) show similar results. The disagreement function for both raters, *G*(*X*, *Y*), provides a value around 0.50 statistically, while CIA (Ψ^N) has a value between 0.40 and 0.50; this value demonstrates moderate agreement between both raters. CIA (Ψ^R) varies depending on the level of agreement among repeated measurements of rater *X*, as rater *X* is taken as a reference. If agreement among repeated measurements by rater *X* is 0.90, Ψ^R has at most a value of 0.10 (Table 4), and if agreement among rater *X*'s repeated measurements is 0.10, Ψ^R has a value greater than 0.70 (Table 5).

When Table 6 is examined, the disagreement values among the repeated measurements of both raters are between 0.21 and 0.25; the disagreement value among the measurements of each rater can be seen affected by the number of raters' measurements, and that the disagreement probability shows a decrease from 0.50 to 0.27 as the number of raters' measurements increases. When CIAs are examined, they can be seen to be unaffected by sample size, yet show an increase depending on the increase in the number of raters' measurements. These coefficients take a value between 0.50 and 0.79.

Table 6

Cases where Agreement Among Repeated Measurements of Raters X *and* Y *is 0.50 (*$X_{ik} = 0.50, Y_{i1} = 0.50$*).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.250 ± 0.000	0.250 ± 0.000	0.503 ± 0.065	0.506 ± 0.068	0.506 ± 0.068
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.337 ± 0.045	0.672 ± 0.090	0.672 ± 0.090
	4	4	0.211 ± 0.008	0.211 ± 0.008	0.285 ± 0.031	0.748 ± 0.080	0.748 ± 0.080
	5	5	0.210 ± 0.010	0.210 ± 0.010	0.268 ± 0.024	0.789 ± 0.072	0.788 ± 0.077
100	2	2	0.250 ± 0.000	0.250 ± 0.000	0.501 ± 0.038	0.502 ± 0.037	0.502 ± 0.037
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.336 ± 0.027	0.666 ± 0.053	0.666 ± 0.053
	4	4	0.211 ± 0.005	0.211 ± 0.004	0.288 ± 0.018	0.737 ± 0.046	0.737 ± 0.048
	5	5	0.210 ± 0.006	0.210 ± 0.005	0.267 ± 0.013	0.789 ± 0.039	0.789 ± 0.042
200	2	2	0.250 ± 0.000	0.250 ± 0.000	0.500 ± 0.025	0.502 ± 0.025	0.502 ± 0.025
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.336 ± 0.019	0.664 ± 0.038	0.664 ± 0.038
	4	4	0.211 ± 0.003	0.211 ± 0.003	0.286 ± 0.013	0.738 ± 0.034	0.739 ± 0.035
	5	5	0.210 ± 0.004	0.210 ± 0.004	0.267 ± 0.009	0.789 ± 0.029	0.789 ± 0.030

Consequently, on the condition that measurements taken by the two raters are independent, if there is perfect agreement among repeated measurements obtained from the same rater, the disagreement functions and CIAs approach 0. When there is no agreement among the repeated measurements obtained from the same rater, while disagreement functions are closer to 0.50, CIAs are closer to 1. If agreement among repeated measurements of both raters differs, Ψ^n is around 0.50 and Ψ^R shows differentiation. While calculating Ψ^R, rater X is regarded as the reference. Thus, if agreement among repeated measurements belonging to rater X is high, Ψ^R is closer to 0; if there is no or lower agreement among repeated measurements, Ψ^R is closer to 1 (Tables 4 and 5). Moderate agreement among the repeated measurements of the raters can be concluded as moderate agreement between the raters when the disagreement values among the repeated measurements of the raters are around 0.25.

Cases Where Raters are Inter-Dependent

CIAs pertaining to cases where the agreement among repeated measurements of both raters is 0.90 and agreement between raters is 0.99 are presented in Table 7. CIAs pertaining to cases where the agreement among repeated measurements of each

rater is 0.90 with no agreement between raters (0.01) are presented in Table 8. CIAs pertaining to cases where the agreement among repeated measurements of each rater is 0.10 and agreement between raters is 0.99 are presented in Table 9. CIAs pertaining to cases where the agreement among repeated measurements of each rater is 0.10 and the agreement between raters is 0.01 are presented in Table 10. CIAs pertaining to cases where the agreement among repeated measurements of each rater is 0.50 and agreement between raters is 0.99 are presented in Table 11. CIAs pertaining to cases where the agreement among repeated measurements of each rater is 0.50 and where the agreement between raters is 0.01 are presented in Table 12.

In case where repeated measurements taken from both raters are equal and in agreement, all disagreement functions are very close to 0. Therefore, CIAs (Ψ^N and Ψ^R) will be equal to or very close to 1. While Ψ^N and Ψ^R are equal to 1 in small samples (n = 30), they have a value between 0.91 and 0.98 in medium and large samples. Additionally, Ψ^N and Ψ^R approach 1 as the number of repeated measurements taken by raters increases (Table 7). Although there is a similar combination in Table 2, the disagreement function $G(X, Y)$ obtained a value around 0.50 statistically because raters X and Y were independent of each other. Here, because raters X and Y are interdependent, even when they are in intra-agreement, a value close to 0 is obtained.

Table 7

Cases where Agreement Among Repeated Measurements of Both Raters X *and* Y *is 0.90 and Agreement Between The Two Raters is 0.99 (*$X_{ik} = 0.90$, $Y_{il} = 0.90$; $X_{ik} = Y_{il}$*).*

N	*K*	L	*G*(*X*,*X'*)	*G*(*Y*,*Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.050 ± 0.000	0.050 ± 0.000	0.050 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.044 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	4	4	0.042 ± 0.004	0.042 ± 0.004	0.042 ± 0.004	1.000 ± 0.000	1.000 ± 0.000
	5	5	0.042 ± 0.004	0.042 ± 0.004	0.042 ± 0.004	1.000 ± 0.000	1.000 ± 0.000
100	2	2	0.050 ± 0.000	0.050 ± 0.000	0.060 ± 0.005	0.917 ± 0.083	0.917 ± 0.083
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.047 ± 0.004	0.954 ± 0.067	0.954 ± 0.067
	4	4	0.042 ± 0.002	0.042 ± 0.002	0.044 ± 0.003	0.967 ± 0.049	0.967 ± 0.050
	5	5	0.042 ± 0.003	0.042 ± 0.003	0.043 ± 0.003	0.974 ± 0.039	0.974 ± 0.041
200	2	2	0.050 ± 0.000	0.050 ± 0.000	0.055 ± 0.004	0.910 ± 0.058	0.910 ± 0.058
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.047 ± 0.003	0.953 ± 0.049	0.953 ± 0.049
	4	4	0.042 ± 0.001	0.042 ± 0.001	0.044 ± 0.002	0.967 ± 0.036	0.967 ± 0.036
	5	5	0.042 ± 0.002	0.042 ± 0.002	0.043 ± 0.002	0.975 ± 0.028	0.975 ± 0.030

In the case of perfect agreement among repeated measurements taken from both raters but the results of both raters differ from each other, while the disagreement functions $G(X, X')$ and $G(Y, Y')$ are very close to 0, the disagreement function $G(X, Y)$ is around 0.70. In this case, a value closer to 0 is obtained for CIAs Ψ^N and Ψ^R. The coefficients Ψ^N and Ψ^R are also observed to be unaffected by sample size or the number of repeated measurements from the raters (Table 8).

Table 8

Cases where Agreement Among Repeated Measurements of Both Raters X *and* Y *is 0.90 but Agreement Between The Two Raters is 0.10 (*$X_{ik} = 0.90$, $Y_{il} = 0.90$; $X_{ik} \neq Y_{il}$*).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.050 ± 0.000	0.050 ± 0.000	0.700 ± 0.005	0.072 ± 0.005	0.072 ± 0.005
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.682 ± 0.044	0.065 ± 0.004	0.065 ± 0.004
	4	4	0.042 ± 0.004	0.042 ± 0.004	0.675 ± 0.044	0.063 ± 0.006	0.063 ± 0.007
	5	5	0.042 ± 0.005	0.042 ± 0.005	0.678 ± 0.044	0.062 ± 0.006	0.062 ± 0.008
100	2	2	0.050 ± 0.000	0.050 ± 0.000	0.700 ± 0.003	0.072 ± 0.003	0.072 ± 0.003
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.683 ± 0.024	0.065 ± 0.002	0.065 ± 0.002
	4	4	0.042 ± 0.002	0.042 ± 0.002	0.677 ± 0.024	0.062 ± 0.003	0.062 ± 0.004
	5	5	0.042 ± 0.002	0.042 ± 0.003	0.677 ± 0.025	0.062 ± 0.003	0.062 ± 0.004
200	2	2	0.050 ± 0.000	0.050 ± 0.000	0.700 ± 0.018	0.072 ± 0.002	0.072 ± 0.002
	3	3	0.044 ± 0.000	0.044 ± 0.000	0.683 ± 0.017	0.065 ± 0.002	0.065 ± 0.002
	4	4	0.042 ± 0.001	0.042 ± 0.001	0.679 ± 0.017	0.062 ± 0.002	0.062 ± 0.003
	5	5	0.042 ± 0.002	0.042 ± 0.002	0.676 ± 0.017	0.062 ± 0.002	0.062 ± 0.003

Comparing Tables 7 and 8, when the agreement between raters is high and self-agreements are perfect, CIAs obtain a value closer to 1; but if the agreement between raters is low then CIAs obtain a value closer to 0.

If there is no intra-agreement among multiple measurements taken by both raters but the results of both raters are in agreement with each other, CIAs, Ψ^N and Ψ^R, are observed to have a value between 0.99 and 1. CIAs Ψ^N and Ψ^R are acknowledged to be unaffected by sample size or the number of repeated measurements taken by raters (Table 9).

Table 9

Cases where Agreement Among Repeated Measurements of Both Raters X *and* Y *is 0.10, and Agreement Between The Two Raters is 0.99 (*$X_{ik} = 0.10$, $Y_{il} = 0.10$, $X_{ik} = Y_{il}$*).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.450 ± 0.000	0.450 ± 0.000	0.450 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.400 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	4	4	0.379 ± 0.011	0.379 ± 0.011	0.379 ± 0.011	1.000 ± 0.000	1.000 ± 0.000
	5	5	0.379 ± 0.014	0.379 ± 0.014	0.379 ± 0.014	1.000 ± 0.000	1.000 ± 0.000
100	2	2	0.450 ± 0.000	0.450 ± 0.000	0.450 ± 0.001	0.989 ± 0.011	0.989 ± 0.011
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.402 ± 0.004	0.995 ± 0.009	0.995 ± 0.009
	4	4	0.380 ± 0.006	0.380 ± 0.006	0.381 ± 0.007	0.996 ± 0.007	0.996 ± 0.007
	5	5	0.378 ± 0.007	0.378 ± 0.007	0.379 ± 0.007	0.997 ± 0.005	0.997 ± 0.005
200	2	2	0.450 ± 0.000	0.450 ± 0.000	0.455 ± 0.004	0.989 ± 0.008	0.989 ± 0.008
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.402 ± 0.003	0.995 ± 0.006	0.995 ± 0.006
	4	4	0.380 ± 0.004	0.380 ± 0.004	0.381 ± 0.004	0.996 ± 0.005	0.996 ± 0.005
	5	5	0.378 ± 0.005	0.378 ± 0.005	0.379 ± 0.005	0.997 ± 0.003	0.997 ± 0.004

Having examined Table 10, CIAs Ψ^N and Ψ^R are observed to change from 0.82 to 0.91, irrespective of sample size, in the case where agreement is observed neither among multiple measurements taken from either rater nor between these raters' results. Based on this information, the following can be said by looking at the

simulation results: CIAs Ψ^N and Ψ^R have a value that is really close to 1 if there is no intra-agreement among repeated measurements from either rater without looking at whether the raters are in agreement with each other (Tables 9 and 10).

Table 10

Cases where Agreement Among Repeated Measurements of Both Raters X *and* Y *is 0.10 But Agreement Between The Two Raters is 0.10 (*$X_{ik} = 0.10$, $Y_{il} = 0.10$, $X_{ik} \neq Y_{il}$*).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.450 ± 0.000	0.450 ± 0.000	0.498 ± 0.030	0.907 ± 0.055	0.907 ± 0.055
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.474 ± 0.022	0.846 ± 0.037	0.846 ± 0.037
	4	4	0.380 ± 0.011	0.380 ± 0.011	0.461 ± 0.019	0.825 ± 0.037	0.825 ± 0.041
	5	5	0.378 ± 0.014	0.379 ± 0.014	0.456 ± 0.017	0.830 ± 0.037	0.829 ± 0.043
100	2	2	0.450 ± 0.000	0.450 ± 0.000	0.500 ± 0.016	0.902 ± 0.028	0.902 ± 0.028
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.474 ± 0.012	0.845 ± 0.022	0.845 ± 0.022
	4	4	0.380 ± 0.006	0.380 ± 0.006	0.462 ± 0.011	0.823 ± 0.022	0.823 ± 0.023
	5	5	0.378 ± 0.007	0.378 ± 0.007	0.456 ± 0.009	0.830 ± 0.020	0.830 ± 0.023
200	2	2	0.450 ± 0.000	0.450 ± 0.000	0.500 ± 0.011	0.900 ± 0.021	0.900 ± 0.021
	3	3	0.400 ± 0.000	0.400 ± 0.000	0.475 ± 0.010	0.843 ± 0.017	0.843 ± 0.017
	4	4	0.380 ± 0.004	0.380 ± 0.004	0.462 ± 0.007	0.823 ± 0.014	0.822 ± 0.016
	5	5	0.378 ± 0.005	0.378 ± 0.005	0.456 ± 0.006	0.829 ± 0.014	0.829 ± 0.016

In the case where agreement among the repeated measurements obtained from each rater is moderate but both raters are in agreement with each other, all disagreement functions are observed to have a value between 0.21 and 0.25 with CIAs Ψ^N and Ψ^R having values between 0.98 and 1; additionally, Ψ^N and Ψ^R coefficients are not affected by sample size or the number of repeated measurements taken by the raters in this situation (Table 11).

Table 11

Cases where Agreement Among Repeated Measurements of Raters X *and* Y *is 0.50 and Agreement Between The Two Raters is 0.99 (*$X_{ik} = 0.50$, $Y_{il} = 0.50$, $X_{ik} = Y_{il}$*)*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.250 ± 0.000	0.250 ± 0.000	0.250 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.222 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	4	4	0.211 ± 0.008	0.211 ± 0.008	0.211 ± 0.008	1.000 ± 0.000	1.000 ± 0.000
	5	5	0.210 ± 0.010	0.210 ± 0.010	0.210 ± 0.010	1.000 ± 0.000	1.000 ± 0.000
100	2	2	0.250 ± 0.000	0.250 ± 0.000	0.255 ± 0.005	0.981 ± 0.019	0.981 ± 0.019
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.225 ± 0.004	0.990 ± 0.015	0.990 ± 0.015
	4	4	0.211 ± 0.004	0.211 ± 0.004	0.212 ± 0.010	0.994 ± 0.010	0.994 ± 0.010
	5	5	0.210 ± 0.006	0.210 ± 0.006	0.212 ± 0.006	0.995 ± 0.009	0.994 ± 0.009
200	2	2	0.250 ± 0.000	0.250 ± 0.000	0.255 ± 0.003	0.981 ± 0.013	0.981 ± 0.013
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.225 ± 0.003	0.989 ± 0.011	0.989 ± 0.011
	4	4	0.211 ± 0.003	0.211 ± 0.003	0.213 ± 0.004	0.993 ± 0.008	0.993 ± 0.008
	5	5	0.210 ± 0.004	0.210 ± 0.004	0.212 ± 0.004	0.995 ± 0.006	0.995 ± 0.007

In the case where agreement among the repeated measurements obtained from each rater is moderate but both raters are in disagreement with each other; CIAs are observe to have values between 0.28 and 0.30, and they are unaffected by sample size or the number of repeated measurements taken by the raters (Table 12).

Table 12

Cases where Agreement Among Repeated Measurements of Both Raters X *and* Y *is 0.50 and Agreement Between The Two Raters is 0.10 (*$X_{ik} = 0.50$, $Y_{il} = 0.50$, $X_{ik} \neq Y_{il}$*).*

N	*K*	*L*	*G*(*X*, *X'*)	*G*(*Y*, *Y'*)	*G*(*X*, *Y*)	Ψ^N	Ψ^R
30	2	2	0.250 ± 0.000	0.250 ± 0.000	0.750 ± 0.000	0.333 ± 0.000	0.333 ± 0.000
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.759 ± 0.007	0.293 ± 0.003	0.293 ± 0.003
	4	4	0.211 ± 0.008	0.211 ± 0.008	0.757 ± 0.010	0.279 ± 0.009	0.279 ± 0.011
	5	5	0.210 ± 0.010	0.210 ± 0.010	0.754 ± 0.010	0.279 ± 0.010	0.279 ± 0.014
100	2	2	0.250 ± 0.000	0.250 ± 0.000	0.750 ± 0.000	0.333 ± 0.000	0.333 ± 0.000
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.759 ± 0.004	0.293 ± 0.001	0.293 ± 0.001
	4	4	0.211 ± 0.005	0.211 ± 0.004	0.756 ± 0.005	0.279 ± 0.005	0.279 ± 0.007
	5	5	0.210 ± 0.006	0.210 ± 0.005	0.753 ± 0.006	0.279 ± 0.006	0.279 ± 0.008
200	2	2	0.250 ± 0.000	0.250 ± 0.000	0.750 ± 0.000	0.333 ± 0.000	0.333 ± 0.000
	3	3	0.222 ± 0.000	0.222 ± 0.000	0.760 ± 0.003	0.293 ± 0.001	0.293 ± 0.001
	4	4	0.211 ± 0.003	0.211 ± 0.003	0.756 ± 0.004	0.279 ± 0.004	0.279 ± 0.005
	5	5	0.210 ± 0.004	0.210 ± 0.004	0.753 ± 0.004	0.279 ± 0.004	0.279 ± 0.005

Discussion and Conclusion

When there is perfect agreement between the repeated measurements of the raters, individual agreement coefficients (CIAs) change according to whether the raters are dependent on each other or not. If the raters are independent from each other, CIAs are observed to be near 0. If they are dependent on each other, CIAs are observed to be near 1 when there is intra-agreement among all the measurements of each rater, and CIAs are near 0 when there is no agreement whatsoever (Tables 2, 7, and 8). In addition, in cases where agreement among the repeated measurements of each rater is low, moderate, or high, CIAs have values close to 1 when there is agreement between raters (Tables 7, 9, and 11). Furthermore, if there is moderate agreement among the repeated measurements of each rater, CIAs show moderate agreement when the measurements of the raters are independent of each other. However, if the measurements of the raters are dependent on each other, CIAs are naturally calculated as 1 when agreement between raters is high, and calculated at near 0 when this inter-agreement is low (Tables 6, 11, and 12). This finding shows that agreement between raters is not enough on its own to increase or decrease the CIA. In other words, intra-agreement among each raters' repeated measures does not necessarily mean there is agreement between the raters.

In cases where there were agreements in one of the rater's repeated measures but not in the other's, CIAs Ψ^N and Ψ^R grew apart; if there was perfect agreement in rater *X*'s repeated measures, the Ψ^R coefficient obtained a value close to 0. Otherwise, it had a value close to 1. Ψ^N has a value around 0.50 when the raters are switched under these conditions (Table 4 and 5).

One can argue that CIA is affected by raters' level of agreement in terms of repeated measures and whether all measurements by these two raters are in agreement. Therefore, it would not be accurate to regard raters as independent while interpreting CIAs. On the other hand, findings have shown that Ψ^N and Ψ^R coefficients are not radically affected by sample size or the number of repeated measures. This finding can be regarded as important for CIA-use in agreement studies.

However, if there is a high level of disagreement within one or both raters' repeated measures, caution is necessary when interpreting the results. As a matter of fact, examining Tables 3, 9, and 10, and especially Table 8, show that agreement between raters increases CIA, even when repeated measures from each rater were in disagreement. In other words, inter-rater agreement was found to be high even though scoring reliability was low. This finding implies that ensuring inter-rater agreement in studies that examine the agreement between CIA and inter-rater agreement does not mean that scoring reliability will also be ensured. Also, in the case where the resulting variable is in a binary state, chance agreement and disagreement values are seriously affected as a result of having only two repeat measurements. Therefore, individual agreement coefficients reach higher values than they should (Table 10). This finding can be interpreted as a disadvantage of CIA. Therefore, this point needs to be taken into consideration when interpreting the findings of studies that include CIA.

Due to the limited number of studies in the literature related to CIA operations in different situations, a direct comparison of this study's findings with the results of other studies is not possible. Hence, future studies on the operations of individual agreement coefficient in different situations can be suggested, as well as on how its strong and weak aspects can contribute to the literature.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Atkinson, G., & Nevill, A. (1997). Comment on the use of concordance correlation to assess the agreement between two variables. *Biometrics, 53*(2), 775–777.

Barnhart, H. X., Lokhnygina, Y., Kosinski, A. S., & Haber, M. (2007). Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. *Journal of Biopharmaceutical Statistics, 17*, 721–738. http://dx.doi. org/10.1080/10543400701329497

Barnhart, H. X., Song, J., & Haber, M. J. (2005). Assessing intra, inter and total agreement with replicated readings. *Statistics in Medicine*, *24*, 1371–1384. http://dx.doi. org/10.1002/sim.2006

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Rinehart and Winston Inc.

David, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of rater error-rates using the EM algorithm. *Applied Statistics, 28,* 20–28.

Erdoğan, E. (2011). *Bilim ve metafizik üzerine tarihsel bir soruşturma* [An historical inquiry into science and metaphysics]. İstanbul, Turkey: Arkeoloji ve Sanat Yayınları.

Erdoğan, S., & Kanık, E. A. (2005, September). *Rasgele ve sistematik hataların sınıf içi ve uyum korelasyon katsayıları ve Bland & Altman yöntemi üzerine etkileri: Bir simülasyon çalışması* [Classrooms and the cohesive correlation coefficients of random and systematic errors: A simulation of its effects on the Bland & Altman method]. Paper presented at VIII. Ulusal Biyoistatistik Kongresi, Bursa, Turkey.

Gao, J. (2010). *Assessing observer agreement for categorical observations*. Research document. Retrieved from https://etd.library.emory.edu/view/record/pid/emory:7t409

Gao, J., Pan, Y., & Haber, M. (2012). Assessment of rater agreement for matched repeated binary measurements. *Computational Statistics and Data Analysis, 56*(5), 1052–1060. http://dx.doi.org/10.1016/j.csda.2011.11.005

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5,* 13–14. http://dx.doi.org/10.1207/S15327841MPEE0501_2

Goodwin, L. D., & Goodwin, W. L. (1991). Using Generalizability theory in early childhood special education. *Journal of Early Intervention, 15,* 193–204. http://dx.doi.org/ 10.1177/105381519101500208

Gulliksen, H. (1950). *Theory of mental tests.* New York, NY: Wiley.

Güler, N., & Gelbal, S. (2010). Studying reliability of open ended mathematics items according to the classical test theory and generalizability theory. *Educational Sciences: Theory & Practice, 10,* 1011–1019.

Güler, N., & Taşdelen Teker, G. (2015). The evaluation of rater reliability of open-ended items obtained from different approaches. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(1)*,* 12–24. http://dx.doi.org/10.21031/epod.63041

Haber, M., & Barnhart, H. X. (2007). A general approach to evaluating agreement between two raters or methods of measurement from quantitative data with replicated measurement. *Statistical Methods in Medical Research*, *14*, 1–19. http://dx.doi.org/10.1177/0962280206075527

Haber, M., Gao, J., & Barnhart H. X. (2007). Assessing observer agreement in studies involving replicated binary observations. *Journal of Biopharmaceutical Statistics, 17*(4)*,* 757–766. http://dx.doi.org/10.1080/10543400701329547

Işıkhan, S., Kılıçkap, M., & Alpar, R. (2013). Graphical and regression methods which are used to examine agreement of two measurement techniques: An applied and scenarios-based study. *Türkiye Klinikleri Journal of Biostatistics, 5*(1), 8–18.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67*(2)*,* 219–229.

Kan, A. (2005). The effection of using grading scale and response key to grader's reliability. *Eurasian Journal of Educational Research*, *19,* 207–219.

Kanık, E. A., & Erdoğan, S. (2004). Değerlendiriciler arası uyumun saptanması [Determining compatibility among evaluators]. *Mersin Üniversitesi Tıp Fakültesi Dergisi, 5*(4), 430–437.

Kanık, E. A., Erdoğan, S., & Orekici Temel, G. (2012). Agreement statistics impacts of prevalence between the two clinicians in binary diagnostic tests. *Journal of Inonu University Medical Faculty, 19*(3), 153–158. http://dx.doi.org/10.7247/jiumf.19.3.5

Kanık, E. A., Orekici Temel, G., & Ersöz Kaya, I. (2010). Effect of sample size, the number of raters and the category levels of diagnostic test on Krippendorff alpha and the Fleiss kappa statistics for calculating inter rater agreement: A simulation study. *Türkiye Klinikleri Journal of Biostatistics, 2*(2), 74–81.

Karakaş, S. (1988). *Bilimsel psikoloji: Temel ilkeler*. Ankara, Turkey: TBMM Vakfı Ofset Tesisleri.

Liao, S. C., Hunt, E., & Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *ANNALS Academy of Medicine Singapore, 39*(8), 613–618.

Lin, L., Hedayet, A. S., & Wu, W. (2012). *Statistical tools for measuring agreement.* New York, NY: Springer.

Pan, Y., Gao, J., Haber, M., & Barnhart, H. X. (2010). Estimation of coefficients of individual agreement (CIA's) for quantitative and binary data using SAS and R. *Computer Methods and Programs in Biomedicine, 98*(2), 214–219. http://dx.doi.org/10.1016/j.cmpb.2009.12.002

Pan, Y., Haber, M., & Barnhart, H. X. (2011). A new permutation-based method for assessing agreement between two raters making replicated binary readings. *Statistics in Medicine, 30*(8), 839–853. http://dx.doi.org/10.1002/sim.4136

Stralen, K. J., Dekker, F. W., Zoccali, C., & Jager, K. J. (2012). Measuring agreement. More complicated than it seems. *Nephron Clinical Practice, 120*(3), 162–167. http://dx.doi.org/10.1159/000337798