KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: January 26, 2016
Revision received: April 26, 2016
Accepted: April 28, 2016
OnlineFirst: August 5, 2016

www.estp.com.tr
DOI 10.12738/estp.2016.4.0080 • August 2016 • 16(4) • 1381-1395

Research Article

Investigation of Coefficient of Individual Agreement in Terms of Sample Size, Random and Monotone Missing Ratio, and Number of Repeated Measures

Gülhan Orekici Temel[1]
Mersin University

Semra Erdoğan[2]
Mersin University

Hüseyin Selvi[3]
Mersin University

Irem Ersöz Kaya[4]
Mersin University

Abstract

Studies based on longitudinal data focus on the change and development of the situation being investigated and allow for examining cases regarding education, individual development, cultural change, and socioeconomic improvement in time. However, as these studies require taking repeated measures in different time periods, they may include various error sources that are exclusive of the investigated variable and that change over time. In Addition, the period of observation is critical in these studies, and observant (variable) attrition or the inability to obtain measurements for various reasons (death, moving away, resignation etc.) may be observed. This study presents the coefficient of individual agreement (CIA) developed for studies in which more than one rater are needed to provide repeated scoring and examines the applications of this agreement coefficient in relation to different sample size, missing value and number of repeated measures. In this context, trial scores were generated based on different sample sizes, missing value ratios, and the number of repeated measures provided by two raters, and the expected agreement between raters was simulated as very high. Related data were replicated 1000 times. Data analysis included the mean and standard deviation of agreement ratios calculated for each combination. According to the results, independently from sample size and number of retests, the CIA takes much lower values than the expected value when random missing value rates reaches 20%. Similar conclusions can be drawn for the existence of monotone missing values except where the sample size is 30 and the number of retests are two and three. It was observed that the CIA was higher than the expected value for the data consisting of two repeated evaluations of the raters with 5% or 10% missing values when the sample size was 30. In general, the agreement coefficient did not exceed the expected level in the situation of three and four replications of measurements. As a result, it can be concluded that an increase in the replication number of measurement decreased the obtained agreement values. The obtained data showed that missing value ratio caused differentiations in the CIA. As a result, increases in missing value ratio cause inconsistencies in CIA. Therefore, it is suggested that data obtained from studies be interpreted considering missing value ratio.

Keywords

Coefficient of individual agreement • Missing value • Scoring reliability • Inter-rater reliability • Simulation

1 Medical Faculty, Mersin University, Mersin Turkey. Email: gulhan_orekici@hotmail.com

2 Medical Faculty, Mersin University, Mersin Turkey. Email: semraerdogann@gmail.com

3 **Correspondence to:** Hüseyin Selvi (PhD), Medical Faculty, Mersin University, Mersin Turkey. Email: hsyn_selvi@yahoo.com.tr

4 Tarsus Technologies Faculty, Mersin University, Mersin Turkey. Email: iremersoz@gmail.com

Citation: Orekici Temel, G., Erdoğan, S., Selvi, H., & Ersöz Kaya, I. (2016). Investigation of coefficient of individual agreement in terms of different sample size, random and monotone missing ratio and number of repeated measures. *Educational Sciences: Theory & Practice, 16,* 1381-1395.

Science and technology are rapidly advancing, and it is a major priority of most countries to keep up with these developments. Advanced technologies and microelectronics have developed in the last quarter of the 20th century and have found applications in all areas—this is believed to constitute the third stage of the Industrial Revolution. This stage has allowed immense progress in technological development and has made it possible to move forward from an industrial society to a knowledge society. In a knowledge society, information technologies are at the center of production and the economy. Advanced information technologies allowing easy access to all parts of the world have facilitated knowledge attainment, trading knowledge and passing it from one to another. This stage, in which stupendous progress has been possible, has been termed the *Knowledge-Information Age*. Rapid spread of inventions such as computers and the Internet in this era has expedited scientific and technological advances and facilitated significant industrialization (Erdoğan, 2011).

Countries' development can now be measured with the level of generated and transferred knowledge, and societies that can train individuals with high self-esteem and who can research and question are at the top of this league table. Training individuals with these qualifications is only possible with well-planned, high quality and sustainable educational initiatives. Such an education requires continuous and long-term assessment of individuals and includes the identification and removal of deficiencies in education. This is feasible through studies based on longitudinal data.

Studies based on longitudinal data in general focus on the change and development in the situation that is being investigated and allow the examination of issues such as education, individual development, cultural change and socioeconomic development in time (Rajulton, 2001). In other words, data related to longitudinal studies are obtained from the same variable at different time frames and are intended to measure the development of the variable in time.

Assessment instruments used for this purpose may include paper and pencil and psychokinetic tests for basic competences as well as instruments that will assess affective competencies. Taking repeated measures or obtaining scores from multiple raters may be necessary to ensure reliability of the measurements obtained through these instruments.

However, reliability is observed as one of the weakest links in essays, verbal and kinetic examinations that involve multiple raters for gathering longitudinal data. Therefore, it is crucial to ensure inter-rater or intra-rater harmony in these studies for the reliability of scoring (Cohen, 1990; Güler & Gelbal, 2010; Lin, Hedayet, & Wu, 2012). In other words, when inter-rater reliability, measurement tools, and longitudinal studies are considered together, error-free measurement of the change and development observed in time will be seen as the function of assessment tools and inter-rater reliability.

In this context, Haber, Gao, and Barnhart (2007) proposed that disagreement between data obtained from the same individuals towards the same variable with different methods is similar to disagreement between repeated measures data obtained from same individuals towards the same variable with different methods, and thus developed the coefficient of individual agreement (CIA), a function of disagreement in which one of the raters acts as a reference to measure agreement or disagreement between methods.

The agreement calculations between raters differ according to the measurement level of measuring device (nominal, ordinal, interval etc.) and the number of raters (Carletta, 1996; Cohen, 1960). The most basic agreement coefficient is Cohen's kappa coefficient—an agreement coefficient in a measurement tool measured at the classification level of two raters. As an alternative to Cohen's kappa, Scott's π statistics, G index and Gwet's AC1 statistics, Fleiss Kappa, and Krippendorf alpha coefficient are also used widespread. Some of these coefficients might be used for agreement calculations between two raters (Cohen's Kappa, Scott's π statistics, G index), and others might be used for agreement calculations between two and more raters (Gwet's AC1 statistics, Fleiss Kappa, and Krippendorf alpha) (Kanık, Erdoğan, & Orekici, 2012; Kanık, Orekici Temel, & Ersöz Kaya, 2010). But all of these coefficients can be used for calculations on the agreement of two and more raters on a single measurement tool. So, these coefficients cannot be used for agreement calculations in a study with longitudinal type consecutive measures. The only coefficient that can be used for concordance calculation of consecutive measures of two or more raters is the individual agreement coefficient. Besides, these coefficients are affected from the missing value. The only coefficient of agreement that is not affected from the missing data in agreement studies is the Krippendorf alpha coefficient.

To define a coefficient of agreement, we first have to decide how we quantify the agreement between the two methods or observers. In cases where there are only two observers, measurements of these observers are indicated with X and Y. Replicated measurements for the first observer (X) is indicated with X and X' and disagreement function between two measurements is G(X,X'); two replicated measurements for the second observer (Y) is indicated with Y and Y' and disagreement function between these two measurements are defined as G(Y,Y'). The quantity of disagreement between measurements obtained from the same individuals is presented with G(X, Y). It is assumed that this disagreement function is $G(X,Y) \geq 0$ and $G(X,X') = 0$ (Haber & Barnhart, 2008; Haber et al., 2007).

To give a more detailed explanation, functions of disagreement for individuals are calculated as shown in Equations 1, 2 and 3 where "N" represents the number of individuals included in the study, "i" represents i^{th} individual, "X_{ik}" represents the measurement obtained by X rater for i^{th} individual for the k^{th} repeated measure and "Y_{il}" represents the measurement obtained by Y rater for i^{th} individual for the l^{th} repeated measure (Gao, Pan, & Haber, 2011; Haber et al., 2007).

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

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

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

– for cases scored in binary mode, X and Y values obtain the value of "1" in positive cases and "0" in negative cases

– for positive cases for X rater; $P(X_{ik} = 1) = \pi_i$ (k = 1 ,…, K_i)

– for positive cases for Y rater $P(Y_{il} = 1) = \lambda_i$ (l = 1 ,., L_i) (Gao et al., 2011; Haber et al., 2007)

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 number of positive readings (1) in repeated measurements that belong to the observer X for i. subjects and U_i shows the total number of positive readings (1) in repeated measurements that belong to the observer Y for i subjects (Haber et al., 2007; Pan, Haber, and Barnhart, 2011).

To calculate CIA by using the function of disagreement; Equation 9 is used when there is no referent rater and Equation 10 is used when rater X is used as a referent (Haber et al., 2007; Pan et al., 2011).

$$\psi^N = \frac{\left[\overline{G}(X,X') + \overline{G}(Y,Y')\right]}{2} \Big/ \overline{G}(X,Y) = \frac{\sum_i \left[\pi_i(1-\pi_i) + \lambda_i(1-\lambda_i)\right]}{\sum_i (\pi_i + \lambda_i - 2\pi_i\lambda_i)} \quad (4)$$

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

Coefficients of individual agreement (ψ^N, ψ^R) in general take a value between 0 and 1. If repeated measurements of each two observers are in agreement, it is expected that disagreement functions of both observers would be similar. It is believed that for an acceptable agreement, coefficients of individual agreement should be at least 0.80 (Haber & Barnhart, 2008; Pan et al., 2011).

The coefficient of agreement presented here can be used to determine scoring reliability in longitudinal studies where developments of individuals in time are examined. Similarly, the coefficient of agreement points to the need for scoring by more than one rater and that repeated measures can be taken; while it can be used in the field of educational sciences in studies related to adaptation and development of scales, psychokinetic tests such as written and oral exams, and instruments of assessment related to affective characteristics, it can also be utilized in medicine for the prognosis of specific diseases, in studies to identify personnel, and in many fields such as economy and politics. As an alternative, there is the total intra ratio (TIR) method; although the TIR is equivalent to CIA, the method uses an alternative statistical inference approach using GEE methodology that works for both continuous and categorical data (Lin et al., 2012).

However, as these studies require obtaining repeated measures in various time frames, they include error sources that are exclusive of the investigated variable that may develop over time, and they require extensive budget, time, and effort. In addition, period of observation is critical in these studies and observant (variable) attrition or inability to obtain measurements for various reasons (death, moving away, resignation etc.) may be observed (Molenberghs & Kenward, 2007). In these situations, such studies cannot be completed or simply such studies can be accepted as biased.

In longitudinal studies where repeated measures are taken at different time periods, the missing data mechanism that occurs due to not taking any measures might have two forms: random and monotone structures. In random missing data structure, lost observations might be monitored in any location, and the alignment types of lost data in repeated measures have no significance. However, the alignment type of lost data in monotone missing value structure is significant, and it is the most frequent missing value structure witnessed in longitudinal studies where repeated measures are taken. In such studies, the initial measures are complete, but the measures after the loss begins are considered as lost (Allison, 2009; Enders, 2010). Most research in the literature generally arrives at a consensus that missing value ratio and insufficient sample size cause bias, however these studies have not found any common ground for the upper limit of missing value ratio. Some of these studies state that missing value ratio should not exceed 5%, some 10% and some 20% (Bennett, 2001; Peng, Harwell, Liou, & Ehman, 2007; Schafer, 1999).

Acock (2005) proposed that the existence of missing values in data cause bias based on the research findings, and that making provision against the missing value problem in the stage of data collection will be the best solution. Çokluk and Kayrı (2011) demonstrated that an increase in missing value amount or assignment of approximate value to these values by several methods negatively affects the reliability and validity of measurement tools.

Allison (2009) stated that the solution by deleting index (in case the missing value amount is high) and missing value cannot satisfy the randomness property and produces non-objective estimations.

Demir and Parlak (2012) studied how much attention is paid to the missing value problem in education research in Turkey and which methods are being used to examine the existence of missing values. They searched 708 articles published between 2009 and 2011 and found out that there are serious problems in the reporting of the missing value process. They reported that there is an almost complete lack of interest towards the missing value problem in education research in Turkey. Similarly Delice (2010) studied the 90 masters theses published between 1997 and 2007 in Turkey under the scope of education research in terms of universe and sample. They reported the following: most of the theses involved just a few statements for the universe and the sample sections; a few gave detailed selected sample tables without explaining the selection criterion; none provided the requirements of the method by random sampling that is commonly used.

Katsis and Limakopoulou (2005) stated that a sample size that is insufficient or selected with an inappropriate method causes biased results. In agreement to Katsis and Limakopoulou, studies determining absorbable missing value amount with sufficient sample size varies according to number of referees and the agreement statistics. Sample size tables were formed for determining sufficient sample sizes for agreement statistics (Machin, 2011). However, such a study aimed at CIA statistics has not yet been published. Therefore, the current study intended to present the coefficient of agreement proposed by Haber et al. (2007) and identify how this coefficient works in terms of missing value. In this context, answers to the following question are sought: "How does individual coefficient of agreement operate in relation to number of repeated measures, type and amount of missing value, and different sample sizes?"

Method

As the study examines the operations of the individual coefficient of agreement in different missing value situations, it is a correlation research (Karasar, 2011).

Research Data

Data used in the study were obtained as a result of simulation. Matlab 7.0 software was used in data simulation and analyses.

Procedure

This study has focused on the status of impact on the agreement of decisions given by two raters for the tests in repeated (consecutive, longitudinal) structures by the

missing value generations that might occur in different structures. Therefore, different trial plans were prepared with two, three, and four repetitions in which the agreement expected between the raters was very high (0.90), and the sample sizes were 30, 100, and 500 (Matala, 2008), for a test where two raters decide within a binary structure.

Within the framework of the trial plans prepared, data were generated in all possible combinations where the agreement between raters was expected to be high and different lost value structures were established as shown in Tables 1–4. As it is suggested that missing value rates in longitudinal studies should not exceed 5%–10% and 20% (Bennett, 2001; Peng et al., 2007; Schafer, 1999), missing data rates are considered to be 5%–10% and 20%. Besides, missing value mechanisms in longitudinal studies can be seen in two different forms. In random structures, losses can be monitored at any location, and the alignment of lost data has no significance (random). In such studies, monotone lost data structure, in which the alignment of lost data is significant, can frequently be witnessed. In such studies, the initial measures are complete; but the measures after the loss begins are considered as lost (monotone). Therefore, the random and monotone missing data rates in this data set are generated to be 5%–10% and 20%, respectively, of the data set. Data sets in each combination were replicated 1000 times (Harwell, Stone, Hsu, & Kirisci, 1996). As a result of this study, a total of 71,000 data were produced: 27,000 in Table 5; 27,000 in Table 6; 18,000 in Table 7; and 9000 in Table 8. The distribution of 1000 data in each combination will be considered as appropriate for normal distribution according to centralized limit theory; average (μ) and standard deviation ($\boldsymbol{\sigma}$) values for individual agreement coefficient will be provided.

In the framework of trial plans, the data were generated in all possible combinations to provide very high inter-rater agreement and different missing value structures were generated as presented in Tables 1–4. Random and monotone missing data ratios in the data set were generated so as to include 5%, 10% and 20% missing data and the simulation for the data set of each combination was iterated 1000 times.

The trail plans for random and monotone missing data structure are given in the following four tables. The parameter *n* used in the tables indicates the sample size. It took the value of 30, 100, and 500 in our study.

Table 1
Random Missing Data Trial Plan

	Two Replicated			**Three Replicated**			**Four Replicated**		
Sample Size	T1	T2	T1	T2	T3	T1	T2	T3	**T4**
1	X	X	X	X	Missing	X	X	Missing	**X**
2	Missing	X	Missing	X	X	Missing	X	X	Missing
.	.	.	.	.		.	.		
.	.	.	.	.		.	.		
.	.	.	.	.		.	.		
n	**X**	Missing	**X**	Missing	**X**	**X**	Missing	**X**	**X**

Table 1 presents a visual presentation of the trial plan developed for the data structure that includes random missing values. The data structure is designed to ensure that it includes 5%, 10%, and 20% missing values randomly distributed.

Tables 2–4 demonstrate the trial plans for the data structure including monotone missing values.

Table 2
Trial Plan I with Monotone Missing Values

	Two Replicated			**Three Replicated**			**Four Replicated**		
Sample Size	T1	T2	T1	T2	T3	T1	T2	T3	**T4**
1	X	Missing	X	X	Missing	X	X	X	**X**
2	X	X	X	X	X	X	X	X	Missing
.	.	Missing	.	.	.	.	.	.	**.**
.	.	.	.	.	.	.	.	.	**.**
.	.	.	.	.	.	.	.	.	**.**
n	**X**	**X**	**X**	**X**	Missing	**X**	**X**	**X**	Missing

Table 2 presents a visual presentation of the trial plan developed for the data structure that includes monotone missing values. In the designed data structure, only the last measurement involves missing values at the ratio of 5%, 10%, and 20%.

Table 3
Trial Plan II with Monotone Missing Values

	Three Replicated			**Four Replicated**			
Sample Size	T1	T2	T3	T1	T2	T3	**T4**
1	X	Missing	Missing	X	X	X	**X**
2	X	X	X	X	X	Missing	Missing
:	.	.	.	.	.	.	**.**
:	.	.	.	.	.	.	**.**
:	.	.	.	.	.	.	**.**
n	**X**	Missing	Missing	**X**	**X**	Missing	Missing

In Table 3, the trial plan II is given, which shows the data structure that includes missing values just at the last two measurements with the ratio of 5%, 10%, and 20%. Therefore, the plan does not involve the situation of the two replications.

Table 4
Trial Plan III with Monotone Missing Values

	Four Replicated			
Sample Size	T1	T2	T3	**T4**
1	X	X	X	**X**
2	X	Missing	Missing	Missing
.	.	.	.	**.**
.	X	X	X	**X**
.	.	.	.	**.**
n	**X**	Missing	Missing	Missing

The last trial plan with monotone missing values is built to examine the existence of missing values that only fall into the last three of the four measurements. The structure is given in Table 4.

Data Analysis

Data analysis first included the calculation of the coefficient of agreement values for the data combinations created according to the given scenarios. Then the mean and the standard deviation of the coefficient were calculated over the values obtained from 1000 iterations of each combination in the study.

Results

In this study, the CIA was evaluated for studies in which more than one rater are needed to provide repeated scoring, and the effectiveness of the coefficient was examined with respect to different sample sizes, rates and distributions of missing values, and numbers of measurement replications. First, data with three different rates (5%, 10%, and 20%) of missing values randomly distributed were generated, including high expected agreement (0.90) between the two raters. The applications were repeated with three different sizes of data (30, 100, and 500), for three different numbers of replication (two, three, and four), in a combinatorial way. Afterwards, the applications of similar nature were performed again with the data containing the monotone missing values. The results section presents the findings obtained from the applications for the research data mentioned above.

The coefficients of agreement were found for each combination. In addition the means and the standard deviations of agreement ratios were calculated for all possible combinations. Table 5 shows the findings obtained from simulation data for trial plans including random missing values in all consecutive test measures.

Table 5
Findings for Trial Plans with Random Missing Data

	N=30			N=100			N=500		
Ratio	**Two Rep.**	**Three Rep.**	**Four Rep.**	**Two Rep.**	**Three Rep.**	**Four Rep.**	**Two Rep.**	**Three Rep.**	**Four Rep.**
	μ ± σ	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**
5%	0.93 ± 0.15	0.88 ± 0.08	0.88 ± 0.07	0.90 ± 0.09	0.87 ± 0.04	0.87 ± 0.04	0.90 ± 0.03	0.87 ± 0.01	0.87 ± 0.01
10%	0.91 ± 0.16	0.87 ± 0.10	0.87 ± 0.08	0.86 ± 0.12	0.85 ± 0.05	0.86 ± 0.04	0.86 ± 0.05	0.85 ± 0.03	0.86 ± 0.02
20%	0.78 ± 0.23	0.82 ± 0.17	0.83 ± 0.13	0.77 ± 0.16	0.80 ± 0.09	0.82 ± 0.07	0.77 ± 0.06	0.80 ± 0.04	0.82 ± 0.03

The results in cases where 20% of the data set includes random missing data show that the individual coefficient of agreement calculated for each combination is lower than the

expected agreement. While individual coefficient of agreement increases according to consecutive number of tests, it decreases based on sample size when consecutive number of tests is constant. When consecutive number of tests is two, individual coefficient of agreement reaches the highest level (0.77–0.78) regardless of sample size.

For two consecutive tests, in cases of 5% random missing values, the individual coefficient of agreement is calculated at the expected level (0.90) independently, whereas it is higher than the expected level (0.93) when the sample size is 30.

When random missing value ratios are 10% and 20% of the data set, the agreement values of all three cases of replication do not show any change between the sample size of 100 and 500; neither of them achieve a higher value than the expected agreement. As is mentioned above, in the second part of the study the effectiveness of the CIA was analyzed for the data containing monotone missing values. Table 6 presents the findings obtained from simulation data for the trial plan I, which includes missing values only in the last measurement of the consecutive test measures.

Table 6
Findings for Trial Plan I with Monotone Missing Data

	N=30			N=100			N=500		
Ratio	**Two Rep.**	**Three Rep.**	**Four Rep.**	**Two Rep.**	**Three Rep.**	**Four Rep.**	**Two Rep.**	**Three Rep.**	**Four Rep.**
	μ ± σ	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**
5%	0.96 ± 0.11	0.90 ± 0.04	0.88 ± 0.06	0.92 ± 0.08	0.88 ± 0.03	0.88 ± 0.04	0.93 ± 0.04	0.88 ± 0.01	0.88 ± 0.02
10%	0.96 ± 0.11	0.90 ± 0.05	0.88 ± 0.06	0.91 ± 0.09	0.88 ± 0.04	0.87 ± 0.04	0.90 ± 0.04	0.88 ± 0.01	0.88 ± 0.02
20%	0.90 ± 0.17	0.90 ± 0.08	0.88 ± 0.06	0.86 ± 0.12	0.88 ± 0.05	0.88 ± 0.04	0.86 ± 0.05	0.88 ± 0.02	0.87 ± 0.02

In the light of the results given in Table 6, the expected 0.90 individual coefficient of agreement increased to 0.96 when the sample size in two consecutive measurements was 30, and the missing value ratio was 5% or 10% in the data. Individual coefficient of agreement was 0.86–0.93 for the other combinations.

Table 7 shows the results for the applications of the data with missing values in the last two measurements of the consecutive test measures.

Table 7
Findings for Trial Plan II with Monotone Missing Data

	N=30		N=100		N=500	
Ratio	**Three Rep.**	**Four Rep.**	**Three Rep.**	**Four Rep.**	**Three Rep.**	**Four Rep.**
	μ ± σ	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**	**μ ± σ**
5%	0.89 ± 0.06	0.88 ± 0.06	0.88 ± 0.04	0.87 ± 0.04	0.88 ± 0.01	0.88 ± 0.02
10%	0.90 ± 0.08	0.88 ± 0.07	0.87 ± 0.05	0.87 ± 0.04	0.87 ± 0.02	0.88 ± 0.02
20%	0.87 ± 0.12	0.88 ± 0.08	0.84 ± 0.07	0.87 ± 0.04	0.85 ± 0.03	0.87 ± 0.02

Table 7 shows that individual coefficients of agreement were found to be 0.84–0.90 and were not affected by sample size, the number of consecutive tests and the missing value ratio. Trial plan II provides lower agreement values for each combination.

CIA was calculated as 0.84–0.85 in the combination where the number of consecutive tests was three, the sample size was 100 and 500, and the missing value rate was 20%.

In trial plan III, the data comprising decisions of the raters for four replicated measurements whose last three measures include missing values were evaluated (see Table 8).

Table 8
Findings for Trial Plan III with Monotone Missing Data

	N=30	N=100	N=500
Ratio	**Four Rep.**	**Four Rep.**	**Four Rep.**
	$\mu \pm \sigma$	**$\mu \pm \sigma$**	**$\mu \pm \sigma$**
5%	0.88 ± 0.06	0.87 ± 0.04	0.87 ± 0.02
10%	0.87 ± 0.07	0.87 ± 0.04	0.87 ± 0.02
20%	0.86 ± 0.10	0.85 ± 0.05	0.85 ± 0.02

According to the results reported in Table 8, the CIA values are 0.85–0.88 in trial plan III applications. It is clearly seen that the CIA is not affected by the sample size or the missing value ratios. However, in general, the agreement values are lower than the expected agreement value for all combinations in trial plan III.

Discussion

It is widely known that tools of assessment need to satisfy two basic criteria, reliability and validity, to ensure quality of measurements and to be the basis of correct decisions. Reliability is defined as the ability of an assessment tool to provide error-free measurements and is a prerequisite for validity (Erkuş, 2006).

In the field of education and psychology, one of the variables that can affect the reliability of scores obtained from a tool of measurement is inter-rater reliability (Aiken, 2000; Crocker & Algina, 1986).

Therefore there are many methods in the literature that can be used in examination of agreement between referees. However the biggest problem while conducting agreement analyses for obtained measurement results is to decide the statistical method to be used. It is seen in many agreement studies that classical statistical methods such as Pearson's correlation coefficient, regression analysis, and *t*-tests for dependent groups are used. However, it was observed that the results obtained from these known classical methods are inaccurate, and this directs research towards developing alternative methods. The statistical method to be used varies depending on with what variable the measurement result is identified or in other words whether outcome variable is constant, discrete,

categorical or sequenced, whether the measured property (variable) fulfills the condition for normality, on the number of observers, if used the number of diagnostic tests and the number of categories in diagnostic test (Lin et al., 2012).

In longitudinal studies where inter-rater reliability is investigated, there may be various error sources that are time based and exclusive of the investigated variable since longitudinal studies involve taking repeated measures in different time frames (Twisk & Vente, 2002). One of these error sources is attrition of observers (variables) from observations or inability to measure them for various reasons (Bernhard, Celia, & Caotes, 1998; Molenberghs & Kenward, 2007; Woodward, Smith, & Tunsatall Pedoe, 1991).

As is the case in many statistical methods, existence of missing data gives rise to bias in studies, and the assessment of missing data requires paying attention to several variables such as type of missing data, sample size, and level of measurement used in the study (Molenberghs & Kenward, 2007). In Schafer's (1999) study, existence of missing data is presented as a source of bias, and it was stated that the ratio of missing data in a data set should be less than 5% (Schafer, 1999). Bennett (2001) reported that the possibility of bias in findings should be kept in sight when missing value ratio is over 10%. Furthermore, in their study, Peng et al. (2007) indicated that missing value ratios should be lower than 20% to avoid deterioration.

The study examined functioning of the CIA proposed by Haber et al. (2007) under the conditions of the existence of missing data, for this purpose all the possible scenarios for the properties of data with missing values were evaluated. The scenario varies depending on the three different sample sizes (30, 100, and 500), and the missing values at three different ratios (5%, 10%, and 20%). All scenarios were executed for two different types of missing value distribution. Firstly, missing values were placed into the data randomly; then simulations were performed for the data that consisted of monotone distributed missing values. For the situation of monotone distribution of missing values, there are three possible combinations in which the data comprises the decisions of two raters for two to four replicated measurements. In the study, each application was iterated for 1000 times to maintain reliability.

According to the results, independent of sample size and number of replication, the CIA takes the values much lower than the expected value when random missing value rates reach 20%. Similar conclusions can be drawn for the existence of monotone missing values except where the sample size was 30 and the number of replications was two or three. These findings are consistent with the findings obtained by Schafer (1999), Bennett (2001) and Peng et al. (2007). In addition, it was observed that the agreement decreased in all combinations when missing data ratio in the data set increased and independent of sample size and missing value ratio, the CIA was approximately the same where the number of consecutive tests was three or four.

The findings of the study demonstrate that the coefficient of individual agreement was higher than the expected value for the data consists of two repeated evaluations of the raters with 5% or 10% missing values when the sample size was 30. These findings are parallel to those of Schafer (1999). In general, the agreement coefficient did not outreach the expected for the situation of three or four replications of measurements.

As sample size increased, generally, the CIA values approached the expected agreement value but the difference between the agreement values for the sample size of 100 and 500 is less than the difference between the sample size of 30 and 100. This fact partially coincides with the statements of Katsis and Limakopoulou (2005) and Delice (2010). As a matter of fact, an increase in sample size caused an approach to the expected value up to a point but this approach was limited in case of higher sample values.

Based on the findings obtained at the end of the simulations, it can be said that increase in replication number of measurements decreased the obtained agreement values. The case becomes clear as missing value ratio increases. Although the study was performed by simulation technique, it is considered that the case can be explained via general limitations of classical test theory.

As a result, it can be concluded that an increase in missing value ratio generates inconsistencies in the coefficient of individual agreement. When sample size increased, the found values for the CIA were closer to the expected value, in general. Therefore, it is suggested that findings in studies should be interpreted by taking missing value ratios into consideration and sample size should be ensured to be at sufficient level.

Conversely, CIA is the only coefficient that can be used to measure agreement between consecutive evaluations of two or more raters. Although the coefficient was found to be affected by a missing value ratio in the study, it can be said that this is not a serious influence, and the differences between the calculated CIA values and the high expected value remain limited.

References

Acock, A. A. (2005). Working with missing values. *Journal of Marriage and Family*, *65*, 1012–1028.

Allison, P. D. (2009). *Missing data. Quantitative methods in psychology*. London, UK: Sage.

Aiken L., (2000). *Psychological testing and assessment*. Boston: Allyn & Bacon.

Bernhard, J., Cella, D. F., Coates, A. S., Fallowfield, L., Ganz, P. A., Mainpur, C. M., … Hürny, C. (1998). Missing quality of life data in cancer clinical trials: Serious problems and challenges. *Statistics in Medicine, 17,* 517–532.

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25,* 464–469.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22*(2), 249–254.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1990). A coefficient of agreement for nominal scales. *Educational and psychological Measurement, 20*(1), 37–46.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. California, CA: Holt, Rinehart and Winston Inc.

Çokluk, Ö., & Kayri, M. (2011). Kayıp değerlere yaklaşık değer atama yöntemlerinin ölçme araçlarının geçerlik ve güvenirliği üzerindeki etkisi [The effects of methods of ımputation for missing values on the validity and reliability of scales]. *Educational Sciences: Theory & Practice, 11*, 289–309.

Delice, A. (2010). Nicel araştırmalarda örneklem sorunu [The sampling ıssues in quantitative research]. *Educational Sciences: Theory & Practice, 10,* 1969–2018.

Demir, E., & Parlak, B. (2012). Türkiye'de eğitim araştırmalarında kayıp veri sorunu [Missing value issue at educational sciences in Turkey]. *Journal of Measurement and Evaluation in Education and Psychology*, *3*, 230–241.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

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

Erkuş, A. (2006). *Sınıf öğretmenleri için ölçme ve değerlendirme: Kavramlar ve uygulamalar* [Measurement and evaluation for primary school teacher: Concepts and applications]. Ankara, Turkey: Ekinoks Yayınları.

Gao, J., Pan, Y., & Haber, M. (2011). Assessment of observer agreement for matched repeated binary measurements. *Computational Statistics and Data Analysis*, *56*, 1052–1060.

Güler, N., & Gelbal, S. (2010). Açık uçlu matematik sorularının güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi [*Studying reliability of open ended mathematics ıtems according to the classical test theory and generalizability theory*]. *Educational Sciences: Theory & Practice, 10,* 989–1019.

Haber, M., & Barnhart, H. X. (2008). A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurement. *Statistical Methods in Medical Research*, *17*, 151–169. https://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*, 757–766.

Harwell, M., Stone, C., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101–125.

Karasar, N. (2011). *Bilimsel araştırma yöntemi* [Scientific research method]. İstanbul, Turkey: Nobel Akademik Yayıncılık.

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. *Journal of Biostatistics*, *2*(2), 74–81.

Kanık, E. A., Erdoğan, S., & Orekici Temel, G. (2012). Agreement statistics impacts of prevalence between the two clinicians in binary diagnostic tests. *Journal of İnönü University Medical Faculty*, *19*(3), 153–158.

Katsis, A., & Limakopoulou, A. (2005). *The determination of the optimal sample size for reliability scales in social sciences*. Greek Statistical Institute Proceedings 18th Panhellenic Conference Statistics. Greece. Abstract retrieved from: http://stat-athens.aueb.gr/~esi/proceedings/18/pdf/435-440.pdf

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

Machin, D., Campbell, M., Tan, S. B., & Tan, S. H. (2011). *Sample size tables for clinical studies*. London, UK: John Wiley&Sons.

Matala, A. (2008). *Sample size requirement for Monte Carlo simulations using Latin hypercube sampling* (Report No. Mat-2.4108). Helsinki University of Technology. Department of Engineering Physics and Mathematics Systems Analysis Laboratory. Retrieved from http://salserver.org.aalto.fi/vanhat_sivut/Opinnot/Mat-2.4108/pdf-files/emat08.pdf

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. London, UK: John Wiley&Sons.

Pani, Y., Haber, M., & Barnhart, H. X. (2011). A new permutation-based method for assessing agreement between two observers making replicated binary readings. *Statistics in Medicine*, *30*, 839–853.

Peng, C. Y., Harwell, M. R., Liou, S. M., & Ehman, L. H. (2007). Advances in missing data methods and implications for educational research. In S. S. Sawilowsky (Ed.), *Real data analysis* (pp. 31–78). Charlotte, NC: Information Age.

Rajulton, F. (2001). The fundamentals of longitudinal research: An overview [Special Issue on Longitudinal Methodology]. *Canadian Studies in Population*, *28*(2), 169–185.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*, 3–15.

Twisk, J., & Vente, W. (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology*, *55*(4), 329–337.

Woodward, M., Smith, W. C., & Tunstall Pedoe, H. (1991). Bias from missing values: Sex differences in implication of failed venipuncture for the Scottish Health Study. *International Journal of Epidemiology, 20*, 379–383.