Croatian Journal of Education
Vol.17; Sp.Ed.No.1/2015: pages: 37-48
Original research paper
Paper submitted: 9th March 2014
Paper accepted: 20th April 2014
doi: 10.15516/cje.v17i0.1484

Establishing Some Measures of Absolute and Relative Reliability of a Motor Tests

Ivan Šerbetar
Faculty of Teacher Education, University of Zagreb

Abstract

Relative and absolute reliability are discussed in this paper on the bases of some empirical motor data. Relative reliability was assessed via the calculation of intraclass correlation coefficients (ICC; Shrout & Fleiss, 1979; Nunnally & Berstein, 1994). Absolute reliability was assessed by calculating standard error of measurement (SEM), and additionally by calculating the smallest detectable change (SDC; de Veet et al., 2006a, b), a relatively new measure which originates from clinical disciplines but has an ever-growing use in other areas. Bland-Altman method (1986; 1995) for determining the limits of agreement and bias between two measurements, was also used. ICC coefficients were high with narrow limits of confidence but ICC masked some differences in trials revealed by SEM and Bland-Altman technique. As stated in Hopkins (2000), and Atkinson and Nevill (1998), more than one measure of reliability should be provided in reliability studies.

Key words: *Bland-Altman limits of agreement; intraclass correlation; reliability of the measurement; smallest detectable change (SDC); standard error of the measurement (SEM).*

Introduction

Data collection is a crucial part of a research process. It involves measurement, which can be defined as the assigning of numerical values to observations with the purpose to quantify the phenomena. Error minimization (reliability) during data collection is critically important (Atkinson & Nevill, 1998) to any research procedure. That emphasizes the importance of understanding measurement theory as well as two most valuable criteria for the judgment of the quality of measures – *validity* and *reliability*. The main concern of the present article is reliability which refers to the reproducibility and repeatability of the measure or variable (Hopkins, 2000).

The theory of measurement assumes that some amount of error is always included in the measurement regardless of the kind of measurement. Classical test theory deals with the *obtained* and *true scores.* Nunnally and Bernstein (1994) explained true score as the "average score that would be obtained over repeated testing" (p. 211) and *measurement error* as the reason of variation of true score over the testing. The sum of true and error score yields the obtained score. Error component of the score can be split into *systematic and random error*, the former can be further broken down into *constant error* and the *bias* while the latter, the random error, is the one which is largely present in behavioral research. Baumgartner and Jackson (1999) assume four sources of measurement error: "lack of agreement among scorers, lack of consistent performance by the individual tested, failure of an instrument to measure consistently and failure of the tester to follow standardized testing procedures" (p. 96).

Reliability is directly related to the error component of the score – the larger the error, the lower the reliability. Although there are many statistical procedures used in reliability estimation, all of them can be classified into one of the two types of reliability – *temporal stability reliability* or *internal consistency reliability* (Baumgartner, 1995). The former represents estimation of the stability of measures applied at different time points to the same subjects, while the latter represents the equivalence of items from the same test (internal consistency) or with the consistency of scoring different raters using the same instrument (interrater reliability). However, most recent theoretical approach distinguishes between *absolute* and *relative* reliability (Atkinson & Nevill, 1998; Hopkins, 2000; Weir, 2005). Relative reliability refers to the magnitude of the association of repeated measurements by quantifying correlation between repeated measures. It forms the ratio of total variability (between subjects/measurement) and individual variability (within subject/measurement) which produces the coefficient of reliability. Absolute reliability, on the other hand, refers to the variability of the scores from trial to trial (within subject/measurement) and it is not sample-dependent because the range of individual scores is not accounted.

This study is not considered a classical hypothesis testing but rather a comparative reliability estimation via relative and absolute approaches.

Intraclass Correlation

Common index of reliability which reflects the ratio between the variance of true score and the total variance on the test is the reliability coefficient which is a form of correlation coefficient. The most popular form of the correlation is *Pearson correlation,* sometimes called *interclass correlation* because the variables come from different "classes" – different categories of observations, i.e. motor skill and psychological trait. There are many studies which use Pearson correlation as the reliability coefficient, but apparently, that statistic has several weaknesses as a measure of reliability. Thomas and Nelson (2001) criticized Pearson *r* because it is a bivariate statistic whereas reliability involves univariate measures, and second, the computation is limited to only two scores while often more than two trials are analyzed. Furthermore, if multiple trials are analyzed, Pearson *r* does not examine different sources of variability. Bland and Altman (1995) complained against correlation coefficient because it cannot, on its own, assess systematic bias and it is greatly dependent on the sample, which is also pointed out by Atkinson and Nevill (1998). Similarly, Hopkins (2000) sees heterogeneity or spread of values between the participants as the main deficiency.

A more appropriate technique uses ANOVA approach for assessing the reliability which is called *intraclass correlation* (*ICC*; Shrout & Fleiss, 1979; Nunnally & Bernstein, 1994), because it measures the correlation within a "class", with the repeated measurements of the same phenomenon used as variables. Although the ANOVA approach is known as a mean of quantifying the differences between the groups of subjects, ANOVA is also a method for establishing the magnitude of variation in different sources of variation. Simple ANOVA design contains three effects: total, between-groups and within groups, whereas repeated measures contains four: total, variations among individuals, variations among trials (or raters) and residual variations (interaction of trials and individuals), that allow computation of variance ratios which form intraclass correlation.

Overall estimation of the general form of ICC is (Weir, 2005; de Vet et al., 2006b):

$$ICC = \frac{\text{between subject variability}}{\text{between subject variability} + \text{error}} \tag{1}$$

It is clear from the Equation 1 that ICC is a relative measure of variability because the magnitude of ICC depends on the between-subjects variability.

To quantify ICC, repeated measures ANOVA is usually performed. Both, one or two-way can be applied while the choice depends on whether the variability is due to trials and error collapsed together (one-way models) or kept separated (two-way models; Weir, 2005). There are several different forms of ICC, but classical citation is 6 different forms of ICC by Shrout and Fleiss (1979). Their nomenclature contains two indexes in ICC models which designate one- or two-way model and fixed or random effects, respectively. Although the original model was later expanded by McGraw and Wong (1996) up to 10th version of ICC, the model named by Shrout and Fleiss (1979) as $ICC_{2,1}$ (two-way with random effects) will be used in the present study.

$$ICC = \frac{\text{between subject variability}}{\text{between subject variability} + \text{between trials variability} + \text{error variability}} \tag{2}$$

Standard Error of Measurement

The larger the variations of the obtained scores around the true score, the larger the measurement error. Indices of error are standard deviations of each subject, while the standard deviation of all errors in one measure is called the *standard error of measurement (SEM)*. Estimation of reliability, presented in the previous section falls in the class of relative reliability, while the standard error of the measurement is the measure of absolute reliability. SEM, denominated by Hopkins (2000) as the "typical error" is the measure of within-subject variation regarded as a "random variation in a measure when individual is tested many times" (Hopkins, 2000, p.2). Calculation of SEM is typically performed by multiplying the standard deviation by √1 and then subtracting the reliability coefficient. Expressed by de Vet et al. (2006a) the estimation is as follows:

$$SEM = SD \ x \sqrt{1 - ICC} \tag{3}$$

where the SD_{pooled} is an average of the standard deviations of two trials.

Weir (2005) argues that the calculation stated above could be substantially affected by the form of ICC, thus, to avert the uncertainty, SEM can also be estimated as a square root of the error variance in ANOVA (Stratford & Goldsmith, 1997; Weir, 2005 de Vet et al., 2006b):

$$SEM = \sqrt{S^2_{error}} \tag{4}$$

If SEM is calculated alternatively, in a manner expressed above, Hopkins (2000) suggests that, because one-way model combines random and systematic error, the error term from the two-way model should be employed:

$$SEM_{agreement} = \sqrt{S^2_{between\ trials} + S^2_{residual}} \tag{5}$$

Smallest Detectable Change

Smallest detectable change (SDC) is a benchmark for the interpretation of changes in scores. It is a measure of variation in scale due to the measurement error (van Kampen et al., 2013) and it is also known as *minimal detectable change* (MDC) or *smallest real difference* (Beckerman et al., 2001). SDC reflects the smallest amount of change in score which is outside an error and which is due to a real change in score and not due to the error in measurement. Calculation of SDC relies on SEM and therefore SDC is also expressed in original units of measurement with a confidence of 90% or 95%. SDC is used extensively in clinical and therapeutic research settings and in practice, but lately also in movement sciences (Weir, 2005; Smits-Engelsman et al., 2011; Schwenk et al., 2012; Holm et al., 2013). The common formula for SDC (e.g. de Vet et al., 2006a) is expressed as:

$$SDC_{95} = 1.96 \ x \sqrt{2} \ x \ SEM \tag{6}$$

The 1.96 in the SDC_{95} Equation represents the z-score at the 95% confidence level while the multiplier square root of 2 is contained because the measurement at 2 time points is considered. SDC, explained in a more practical context, means that, if the difference in scores emerges in magnitude greater than SDC, there is 95% probability that the difference was not due to the error or variation but resulted from the real difference in measurement. Because SDC is based on SEM, it can also be calculated from the variance error term (Bruynesteyn et al., 2005; Van Kampen et al., 2013) which is the approach taken in the present research.

Bland-Altman Limits of Agreement

The *Bland-Altman limits of agreement* method (LOA; Bland & Altman, 1986, 1995), also known as *Bland-Altman plot* or *difference plot,* is a method for graphical comparison of two techniques of measurements or two variables of interest. In this method, the differences between two variables are plotted against their averages. Horizontal reference lines on the plot represent the mean of the differences between the measurements, and limits of agreement, respectively, which are fitted on the plus and minus 1.96 SD, whereas along the x axis, the averages of the two measurements are displayed. The *Bland-Altman plot* is a suitable method to disclose the association between the differences and the averages, to check for any orderly bias and to uncover outliers. Basically, if the differences are small, and the mean of the differences is near zero, the test can be considered reliable. Atkinson and Nevill (1998) encourage the use of the LOA method, especially because of the exploration of heteroscedasticity that is inherent in this analysis.

Methods

In the context of a broader research project oriented to metric characteristics of motor tests, a large battery of tests was applied with younger school-aged children. For the present study, results of 6 motor tests, measured on 142 children aged 7, were chosen. Measurement sessions were held one week apart and all the measures were taken by a single tester.

Intraclass correlation coefficients were calculated with a two-way random model for absolute agreement (Baumgartner & Chung, 2001; de Vet et al., 2006b):

$$ICC_{agreement} = \frac{S^2_{subj}}{S^2_{subj} + S^2_{trials} + S^2_{error}} \qquad (7)$$

SEM was computed using a variance component from a two-way ANOVA according to the Equation 5.

The Bland-Altman method for assessing the agreement between the first and second trial was also used. Limits of agreement were defined as mean bias ±2 SD. Bland-Altman plots were created in *SigmaPlot*, while all the other computing was performed in *SPSS.*

Results

Descriptive statistics, as well as some indices of differences are shown in Table 1. The mean differences of all the measures, except *medicine ball throw*, were close to zero indicating no systematic differences between trials (Table 1). Exceptionally large range of values between trials was observed for *medicine ball throw*. Overall, SEM and SDC were quite large, especially for *medicine ball throw*, which is logical because SEM and, consequently SDC, rise as the value of measure rises. Both measures were also expressed in percentages thus further uncovering the width of error. Since SEM values can be translated to normal curve probabilities, Table 2 values can be applied to the practice. Using *broad jump* as the example, it can be expected with the probability of 68% that the values obtained on repeated measurement will be within ~ ±8 cm of the original values, i.e. there is 96% chance that the value of repeated measurement will be in ~ ±16 cm of the original value.

Given the highest and the lowest values of SDC calculated for the present variables, a change in the individual performance of less than one third of the mean cannot be detected beyond measurement error for the *sit and reach* and less than one fifth for *tapping* (calculated relatively to the mean of the three trials).

Table 1

Descriptive Values (N=142)

	Trial 1 Mean(SD)	Trial 2 Mean(SD)	Trial 3 Mean(SD)	Trial1-3 Mean Diff. (SD)	Range
Sit and reach	30.93 (7.93)	31.26 (7.85)	31.28 (7.88)	-.81 (3.59)	22
Grip strength	10.94 (2.68)	11.61 (2.67)	11.30 (2.65)	-.28 (1.31)	7.33
Broad jump	107.58 (18.75)	107.80 (19.91)	107.96 (19.74)	-1.41 (8.30)	54.67
Medicine ball throw	287.63 (88.83)	322.63 (96.76)	321.02 (101.46)	-41.99 (54.04)	300
Tapping	17.62 (3.21)	17.52 (3.32)	17.50 (3.36)	-.16 (1.41)	8
Shoulder flexibility	77.11 (10.98)	77.15 (11.69)	77.67 (11.26)	-1.36 (5.03)	33.33

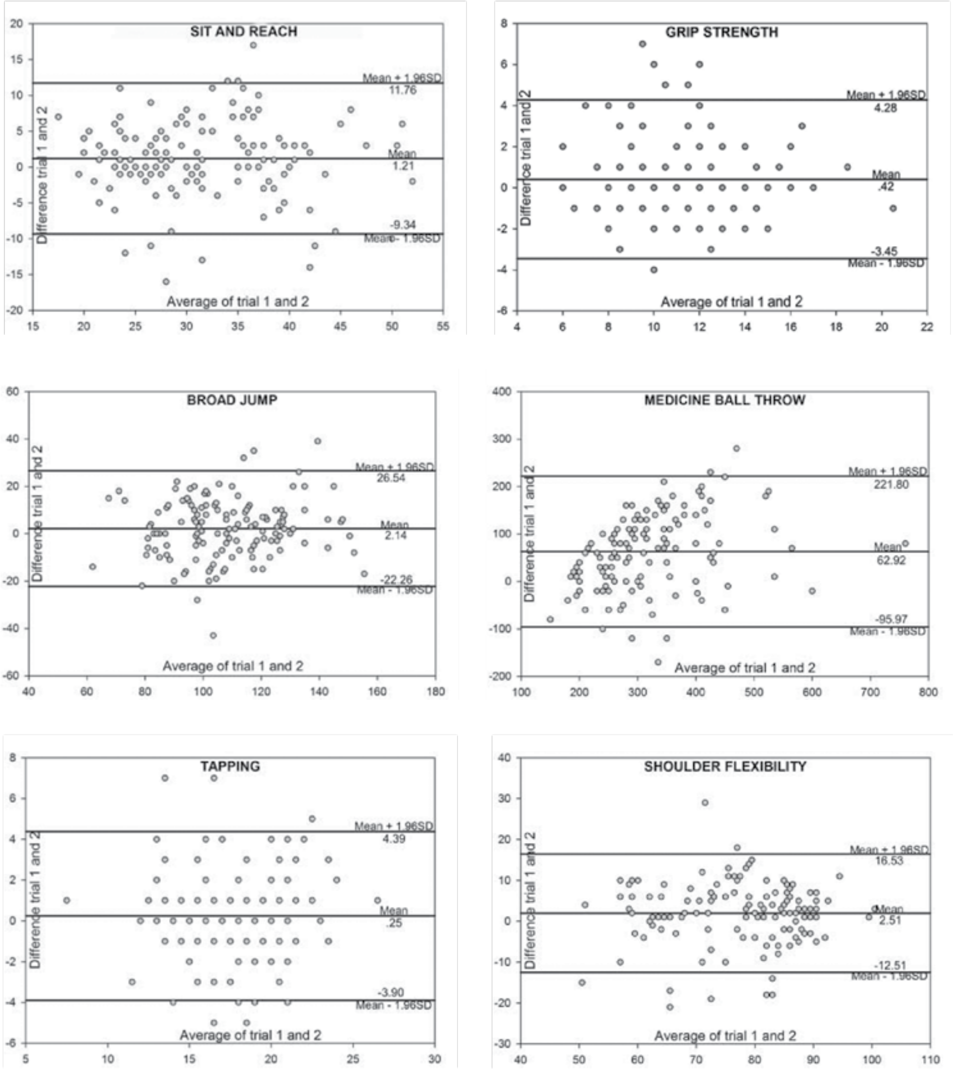
Intraclass correlation coefficients were all, except one, ≥ 90 (Table 2), which were, according to the common criteria, high values. Confidence intervals were narrow, except for the relatively large confidence interval obtained for *medicine ball throw.* There is a visible tendency that increases in ICC, followed by a decrease of confidence intervals as anticipated in Baumgartner and Chung (2001). Respecting ICC values, the magnitude of error was between 6 and 11% (1-ICC) of the total variance.

Table 2

Values of Reliability Measures - ICC, SEM and SDC

VARIABLE	$ICC_{2,1}$ (95%CI)	$SEM_{agreement}$	%SEM	$SDC_{95\%}$	$\%SDC_{95\%}$
Sit and reach	.93(.90 .95)	3.43	10.97	9.50	3.03
Grip strength	.91(.88 .94)	1.27	11.23	3.52	31.15
Broad jump	.94(.92. 95)	8.08	7.48	22.37	20.72
Medicine ball throw	.90(.79 .94)	53.99	16.81	149.55	46.58
Tapping	.94(.92 .96)	1.34	7.66	3.71	21.52
Shoulder flexibility	.89(.86 .92)	5.81	7.48	16.11	20.74

Bland-Altman scatterplots were created to estimate the disagreement between two measurements as a function of the mean of the two measurements. Due to software limitations, only the first and the second trials were evaluated. The charts are shown in Figure 1. Mean differences were all, except *medicine ball throw*, near zero and positively biased, which means that the values of retest were somewhat larger than in the first trial. In all the panels it can be seen that there is not much change in the differences as the mean increased while the variation in the data was adequately constant.



Note. Line in the middle – bias or mean absolute agreement; upper and lower lines – upper and lower limits of agreement

Figure 1. Bland-Altman Plots

Upper confidence interval for the *medicine ball throw* was more than twice as broader than the lower confidence interval (i.e. limits of agreement - LOA), clearly indicating that there is an effect of learning or motivation (or both) in the second trial. All other LOAs were relatively narrow although the scores were scattered widely among them. Number of the observations, which exceeded LOA, across the variables was: 10 (7%), 6 (4%), 6 (4%), 5 (3.5%), 9 (6%), 9 (6%). Given that according to the normal distribution theory, 5% percent of observations should fall in the range above or below 2 SD, present results are approximately in concordance with the theory. Slight tendency to heteroscedasticity is visible on *medicine ball throw* panel, where the increase in scores is accompanied with the increase in the amount of error.

Discussion

The study attempted to empirically relate different approaches to reliability, based on Hopkins' (2000) strategy that in every reliability study at least three trials should be performed, and that more than one measure should be employed. The sample size was one of convenience, although substantially larger than the recommended "around 30" by Baumgartner and Jackson (1999) or "at least 50" as suggested by Baumgartner and Chung (2001, p. 187).

The intraclass correlation coefficient was computed with a two-way random effects model with absolute agreement. Although high ICC coefficients were obtained, one should bear in mind that ICC is a ratio index of within and between subject variability, therefore agreement between groups of subjects is assessed in repeated measures, which does not provide information about the amount of individual change or error in scores. Furthermore, ICC is dependent of the variability in the sample and thus assignment to the other populations may be ambiguous. Weir (2005) showed that if subjects differ little from each other, ICC values will also be small, regardless of the small trial-to-trial variability. Also, if subjects differ significantly, ICC can be large even if variability of trials is large. As cited in Atkinson and Nevill (1998) "ICC is affected by sample heterogeneity to such degree that a high correlation may still mean unacceptable measurement error for some analytical goals... and it should not be employed as the sole statistic" (p. 228). Hence, the use of the second measure of reliability is necessary, as evidenced by the variable *medicine ball throw*, for which high ICC coefficient was obtained, but Bland-Altman method showed substantial bias and large disproportion of the limits of agreement. However, it should be stressed that only the first and second trial were assessed with Bland-Altman method.

Despite some criticism of the limits of agreement method pronounced by Hopkins (2000) who was preliminary concerned with the bias of the limits of agreement caused by the sample size, it was found that in the estimation of reliability, the plots were useful in visualizing outliers and especially in exposing the relationship between the trials. The differences between the first two trials were positively biased and a tendency to heteroscedasticity was observed.

The observed error between trials in the repeated measurements can be quantified as SEM or SDC. Both measures are expressed in the original units of measurement which allow direct comparison of the scores in repeated trials. SEM represents the stability or variability of response and defines the range of the scores which can be expected in the repeated test. Traditional problem for researchers is how to determine whether the change in score or an error in measurement is significant. Calculating SDC statistic is a possible way to resolve the problem, because SDC symbolizes a minimal change in score which cannot be attributed to the measurement error. Hence, if the scores are at the level of SDC or higher, that is either due to real change in ability or due to inaccuracy of the instrument (considering measured ability constant), not due to an error in measurement.

SDC is typically calculated to one of two degrees of confidence, e.g. SDC90 (90% confidence) and SDC95 (95% confidence), and both measures have been found in studies (Holm et al., 2013). In the present study SDC was calculated using the usual 95% confidence limits. Because of that and because of large $\mathrm{SEM}_{\mathrm{agreement}}$ values, caused by the included systematic variability, the SDC values were large. Therefore, Hopkins' (2000) argumentation should be considered. In his opinion 95% confidence limits are too stringent to use as a threshold in the decision whether or not the real change has occurred, and he recommended using 1.5 or 2.0 times the SEM instead of 2.77 times the SEM. In the particular case in this study, relations of the SDC measures and results should be taken with caution because they are interpreted relatively to the mean and they are representative for a strictly limited population. However, practical use of SDC is highly recommended to the practitioners because, with respect to the practical context of individual measure and measurement procedure, SDC can reveal whether the real progress in the measured ability has occurred.

Conclusion

In the study, test-retest reliability was assessed using several procedures belonging to relative and absolute reliability methods. First, ICC with two-way random effects model (ICC_{21}), with absolute agreement, was conducted to include any systematic variability in test repetitions. SEM and SDC were calculated from ANOVA variance components (between trials and residual variance). SDC refers to minimal within-subject change which cannot be attributed to measurement error but rather indicates real change in the measured ability. Bland-Altman limits of agreement were created to visually describe differences between the first two trials against averages in scores. Because of software limitations, only the first and the second trials were assessed with Bland-Altman scatterplots.

Although ICC coefficients were high with narrow limits of confidence, ICC masked some individual differences in trials revealed by SEM and Bland-Altman technique.

ICC is regularly reported in studies, but it barely assesses measurement error relatively to the between-subject variability for the measured subjects. Hence, SEM and SDC are more practical measures for research but also to the practitioners. As stated in Hopkins (2000), and Atkinson and Nevill (1998), more than one measure of reliability should be provided in reliability studies, and for ICC the interval of confidence should always be stated.

References

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, 26, 217–38. http://dx.doi.org/10.2165/00007256-199826040-00002

Baumgartner, T. A. (1995). Objectivity and reliability of the 90-degree push-up for college students. *Research Quarterly for Exercise and Sport*, 66 (Supp.), A-47.

Baumgartner, T. A., & Jackson, A. S. (1999). *Measurement for evaluation in physical education and exercise science* (6th ed.). Madison, WI: WCB/McGraw-Hill.

Baumgartner, T. A., & Chung, H. (2001). Confidence Limits for Intraclass Reliability Coefficients. *Measurement in physical education and exercise science*, 5(3), 179–188. http://dx.doi.org/10.1207/S15327841MPEE0503_4

Beckerman H., Roebroeck M.E., Lankhorst, G.J., Becher, J.G., Bezemer, P.D., & Verbeek A.L.M. (2001). Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research*, 10, 571-578. http://dx.doi.org/10.1023/A:1013138911638

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327, 307-310. http://dx.doi.org/10.1016/S0140-6736(86)90837-8

Bland, J.M., & Altman, D.G. (1995). Comparing two methods of clinical measurement: a personal history. *International Journal of Epidemiology,* 24 Suppl. 1, 7-14. http://dx.doi.org/10.1093/ije/24.Supplement_1.S7

Bruynesteyn, K., Boers, M., Kostense, P., Van der Linden, S., & Van der Heijde, D. (2005). Deciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change. *Annals of the Rheumatic Diseases,* 64, 179–82. http://dx.doi.org/10.1136/ard.2003.018457

De Kegel, A., Baetens, T., Peersman, W., Dhooge, I., Rijckaert, J., Cambier, D., & Van Waelvelde, H. (2012). Ghent Developmental Balance Test: a new tool to evaluate balance performance in toddlers and preschool children. *Physical Therapy,* 92, 841–852. http://dx.doi.org/10.2522/ptj.20110265

de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol,D. L., & Bouter, L. M. (2006a). Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes*, 4, 54. http://dx.doi.org/10.1186/1477-7525-4-54

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006b). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology,* 59, 1033–1039. http://dx.doi.org/10.1016/j.jclinepi.2005.10.015

Donoghue, D., & Stokes, E.K. (2009). How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *Journal of Rehabilitation Medicine,* 41, 343-346. http://dx.doi.org/10.2340/16501977-0337

Holm, I., Tveter, A.T., Smith Aulie, V., & Stuge, B. (2013). High intra- and inter-rater chance variation of the movement assessment battery for children 2, ageband 2. *Research in Developmental Disabilities*, 34, 795-800. http://dx.doi.org/10.1016/j.ridd.2012.11.002

Hopkins, W. G. (2000). Measures of Reliability in Sports Medicine and Science. *Sports Medicine*, 30(1), 1-15. http://dx.doi.org/10.2165/00007256-200030010-00001

McGraw, K.O., & Wong S.P. (1996). Forming some inferences about some intraclass correlation coefficients. *Psychological Methods,* 1, 30-46. http://dx.doi.org/10.1037/1082-989X.1.1.30

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin,* 86, 420-428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Schreven, S., Toussaint, H.M., Smeets, J.B., & Beek, P.J. (2013). The effect of different inter-pad distances on the determination of active drag using the Measuring Active Drag system. *Journal of Biomechanics*, 46, 1933–1937. http://dx.doi.org/10.1016/j.jbiomech.2013.05.020

Schwenk, M., Gogulla, S., Englert, S., Czempik,A. & K. Hauer (2012). Test–retest reliability and minimal detectable change of repeated sit-to-stand analysis using one body fixed sensor in geriatric patients. *Physiological Measurement,* 33, 1931-1946. http://dx.doi.org/10.1088/0967-3334/33/11/1931

Smits-Engelsman, B.C., Niemeijer, A.S., & van Waelvelde, H. (2011). Is the Movement Assessment Battery for Children-2nd edition a reliable instrument to measure motor performance in 3 year old children? *Research in Developmental Disabilities*, 32(4), 1370-1377. http://dx.doi.org/10.1016/j.ridd.2011.01.031

Stratford, P.W., & Goldsmith, C.H. (1997). Use of standard error as a reliability index of interest: An applied example using elbow flexor strength data. *Physical Therapy,* 77, 745–750.

Thomas, J.R., & Nelson, J.K. (2001). *Research methods in physical activity* (4th ed.). Champaign, IL: Human Kinetics.

Van Kampen, D. A., Willems, W. J., van Beers Loes, W. A. H., Castelein, R. M., Scholtes, V., & Terwee, C. B. (2013). Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *Journal of Orthopaedic Surgery and Research,* 8, 40. http://dx.doi.org/10.1186/1749-799X-8-40

Weir, J.P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research,*19, 231.

Ivan Šerbetar
Faculty of Teacher Education, University of Zagreb
Ante Starčevića 55, 40000 Čakovec, Croatia
ivan.serbetar@ufzg.hr

Određivanje nekih mjera apsolutne i relativne pouzdanosti motoričkih testova

Sažetak

Rad u kojem se raspravlja o relativnoj i apsolutnoj pouzdanosti utemeljen je na empirijskim motoričkim podacima. Relativna pouzdanost određena je izračunavanjem koeficijenata intraklas korelacije (ICC; Shrout i Fleiss, 1979; Nunnally i Berstein, 1994). Određivanje apsolutne pouzdanosti izvedeno je izračunavanjem standardne pogreške mjerenja (SEM) i dodatno izračunavanjem najmanjih uočljivih promjena (smallest detectable change; SDC; de Veet i sur., 2006a,b), relativno nove mjere koja potječe iz kliničkih disciplina, a sve veću primjenu nalazi i u drugim područjima. U radu je upotrijebljen i Bland-Altmanova metoda (1986, 1995) određivanja razina slaganja (limits of agreement; LOA), odnosno biasa između dva mjerenja. Dobiveni su visoki ICC koeficijenti i suženi limiti pouzdanosti, no ICC koeficijenti su prikrili neke razlike u ponovljenim izvedbama testa koje su uočene primjenom SEM i Bland-Altman metode. Kao što su naveli Hopkins (2000), Atkinson i Nevill (1998) u istraživanjima pouzdanosti treba primijeniti više mjera.

Ključne riječi: *Bland-Altmanove razine slaganja (LOA); intraklas korelacija; najmanja uočljiva promjena (SDC); pouzdanost mjerenja; standardna pogreška mjerenja (SEM).*