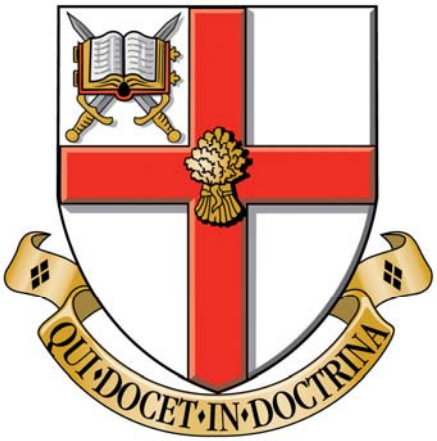

University of Chester

Author(s): Kevin L Lamb

Title: Test-retest reliability in quantitative physical education research: A commentary

Date: 1998

Originally published in: European Physical Education Review

Example citation: Lamb, K. L. (1998). Test-retest reliability in quantitative physical education research: A commentary. *European Physical Education Review*, 4, 145-152

Version of item: Author's post-print

Available at: http://hdl.handle.net/10034/29375

TEST-RETEST RELIABILITY ANALYSIS IN QUANTITATIVE PHYSICAL EDUCATION RESEARCH: A COMMENTARY

K.L. Lamb (1998)

Published in *European Physical Education Review*, 4, 145-152

TEST-RETEST RELIABILITY ANALYSIS IN QUANTITATIVE PHYSICAL EDUCATION RESEARCH: A COMMENTARY

Abstract

This paper highlights an important statistical development for exercise and physical education research. Traditionally, the Pearson and intraclass correlation coefficients have been liberally used by researchers to quantify the test-retest reliability of many performance, behavioural, and physiologically-related measurements. The suitability of these forms of analyses has recently been challenged by British exercise scientists, who argue that they do not really address what they are meant to, that is, the level of agreement between repeated measurements or scores. As a consequence, our existing knowledge of the reliability of such measurements is questionable and deserves to be re-established with a more appropriate statistical technique. Accordingly, the 95% Limits of Agreement method is presented and offered as an essential supplement for future measurement and evaluation research.

Introduction

The application of scientific principles to the study of sport and exercise demands of its investigators an understanding of the research process, and in particular the inter-related issues of research design and statistics. Moreover, it is assumed that the highly regarded peer-review process offered by respected academic journals acts to filter-out research that fails to show good regard for these issues. However, it is quite clear to the informed reader of sport/exercise literature that the above status is not universal, resulting in some research whose conclusions must at best be regarded as dubious, and at worst spurious. One sports science journal editor has boldly highlighted the “use and abuse of statistics in sport and exercise sciences” (Bartlett, 1997) and in doing so cautioned both authors and referees of the unacceptability of poorly designed studies and/or incorrect statistical techniques.

Despite its lengthy history of quantitative research activity, especially in the area of measurement and evaluation, and the existence of numerous focused text books (e.g. Thomas and Nelson, 1996; Morrow, Jackson, Disch, and Mood, 1995; Safrit and Wood, 1989; Baumgartner and Jackson, 1987; Bosco and Gustafson, 1983; Cohen and Holliday, 1979; Morehouse and Stull, 1975) and academic journals (e.g. *Research Quarterly for Exercise and Sport, Quest, European Journal of Physical Education,* and *European Physical Education Review* and its predecessor *Physical Education Review*), the discipline of physical education (PE) seems to be especially susceptible to the above criticism. The lack of research emphasis within its professional training programme is a likely explanation, but as the area of paediatric exercise-related research continues to expand, it is time for the situation to improve.

Worthy of particular scrutiny are all studies that have reported on measurements made of particular human attributes, such as anthropometric dimensions, physiological functions, performance indicators, attitudes, and behaviours. For the current journal this includes approximately 31% of all articles published during the last 10 years (1988-97), and probably larger proportions in the PE-related journals mentioned above. Most of these papers presented their data with some form of accompanying statistical analysis, on which subsequent interpretations and conclusions were based. Fundamental in each case is the premise that the measures or tools used were deemed to be valid, and thereby also reliable (repeatable). However, it has recently emerged that this might not have been the case. A movement in sports science research has questioned the appropriateness of the statistical techniques traditionally used to reflect measurement reliability. By implication, this threatens the credibility of the findings in past studies that, quite naturally, selected their measures based on such statistics. Furthermore, such a realisation has crucial implications for current and future empirical investigations. Therefore the purpose of this paper is to highlight the key details of this argument in order that the readership of this journal may at least be in a position to view existing knowledge from a more critical perspective, and at best design future investigations accordingly.

Measurement Reliability

The messages conveyed by recent articles addressing the issue of test-retest reliability (Atkinson, 1995; Nevill and Atkinson, 1997) are simple, yet potentially shocking in their significance. In short, it could be claimed that the reliability of the many measurement tools used routinely in physical education is highly questionable, or even unknown. Moreover, unknown reliability means unknown validity. The problem identified relates primarily to the frequent mis-use of the Pearson correlation coefficient (*r*), and (to a lesser extent) the intraclass correlation (*R*) as measures of *absolute* reliability (Baumgartner, 1989), and the inference that a high correlation (say > 0.80) between repeated scores equates to good agreement. Clearly it is wrong since two sets of scores can be very highly correlated (> 0.95) but be very different. In the fictitious data below, 10 subjects are measured under identical conditions on two occasions, seven days apart. The variable being measured is unspecified, but has a lower limit of zero, and an unbounded upper limit (which, it could be argued, is the case for many physical performance indicators).

The scores for each subject are different for Test 1 and Test 2 (4 increased and 6 decreased), but the rank order within the group is stable. The smallest change is for subject 5 (4 units) and the largest for subject 3 (11 units).

The Pearson and intraclass correlations between the two sets of scores are 0.96 and 0.97, respectively, and, coupled with a small, non-significant ($p > .05$) difference between the mean values (0.9 units), this particular measurement could be interpreted as being highly reliable. Indeed, as a measure or *relative* reliability (Baumgartner, 1989), the high correlations do reflect well the stability of position or rank order within this particular sample. However, the scores do not agree (for any subject), and absolute reliability is not good.. In the worst case, the disagreement was 58% between tests, and in the best case 10%. The final decision on whether the extent of these differences is ‘acceptable’ is a matter for informed opinion - “informed” referring to the practical significance of the observed difference. At present, such a statement is rather vague in that it is unlikely that consensus exists in the PE domain as to what should be used to reflect acceptable

reliability. A consideration should be the magnitude of change (difference) in the dependent variable expected due to recognised independent factors, such as training, sex, maturation, or malaise. If the degree of test-retest variability is similar in size to the effects of any of these, then it must be deemed unacceptable. The following (genuine) data exemplifies this point.

The differences between Test 1 and Test 2 range from -10 to +6 cm, and in only two cases were the performance scores in agreement (difference = 0). However, the mean values were identical and the Pearson and intraclass correlations were high (0.93 and 0.96, respectively). Previous reports on the reliability of the vertical (or Sargent) jump test have yielded correlations of 0.93 to 0.99 (Adams, 1994, p. 56) and therefore favourable interpretations. Yet, the above data clearly show variable agreement; at worst a decrease in performance of 25%, and an overall mean change (regardless of sign) of 9.2% (3.5 cm) from Test 1 to Test 2. Is this amount of variability acceptable, or is it large enough to conclude that this particular test, often regarded as a measure of anaerobic power, is unreliable?

As indicated above, the answer lies in knowing how much such a performance is expected to differ between defined groups (for example, elite versus novice; skilled versus unskilled; old versus young), or change following some kind of controlled 'intervention', be it a period of power-related training, or de-training, and knowing the likely consequence (or value) of such a change. A study by Viitasalo (1985) sheds some light on the current data. Following 5 months' power training, Viitasalo reported that the vertical jump performances of elite male volleyball players increased on average by 9% - an amount almost equivalent to the test-retest difference in the above data. Now, whilst the subjects in Viitasalo's study may have been more skilled in performing the vertical jump than the mixed sample above, and therefore more consistent as performers, one could justifiably question whether such improvements were due to the training, or the apparent unreliability of the jump test. The situation is one in which the reliability of the test is having a marked impact on its ability (statistical power) to detect changes in performance due to the treatment (Schabort, Hopkins,

and Hawley, 1998). Knowing whether volleyball performance increased in parallel - the 'value' of the measured change - would add to the interpretation here.

Appropriate Reliability Analysis

British exercise scientists over the past three years have been advocating the use of a relatively simple statistical technique described in the mid-eighties by the medical statisticians Bland and Altman (1986). Atkinson (1995), Nevill (1996), and Nevill and Atkinson (1997) have provided a convincing argument for using Bland and Altman's *Limits of Agreement* technique to assess measurement reliability of data which is parametric in nature (measured on an interval or ratio scale - the kind of data typically generated by many assessment tools in PE). Assuming the data fulfils three important conditions (see below), Limits of Agreement (LoA) analysis will provide a more complete appraisal of repeatability than the other popular reliability statistics (Bailey, Sarmandal, and Grant, 1989; Sarmandal, Bailey, and Grant, 1989; Ottenbacher and Tomchek, 1994; Atkinson, 1995). In addition, unlike correlation coefficients, LoA analysis allows reliability to be expressed in the unit of the measurement, and it is not vulnerable to over-estimating reliability due to sample heterogeneity (wide variability between subjects).

As the name suggests, LoA analysis addresses the amount of *agreement* between repeated measurements of the same variable (and not the relationship). For a study sample, LoA analysis quantifies how close (agreeable) the repeated measures are for most subjects in that sample. Ideally, there should be perfect agreement between scores for all subjects, giving an average difference of zero units. As this situation seldom occurs when measuring human attributes, Bland and Altman (1986) indicated that the data of 95% of subjects should be regarded in the analysis, allowing for extreme or unusual measurements to be ignored, and the level of agreement found amongst their data being that which represented its reliability. Thereafter, the interpretation of this level of agreement should be left to the practitioner, whom is assumed to be familiar with the variable of interest and be able to assess its practical significance.

LoA statistics for test-retest data take the form of the mean (or *bias*) difference between repeated measurements, and the standard deviation (SD) of these differences multiplied by 1.96. This latter calculation provides the level or *limits* of agreement for 95% of the sample, and is expressed as a plus or minus value (since some subjects have a difference above the mean value, and others below). Both bias and 95% limits can be expressed graphically (see Figure 1) as horizontal lines on a plot of subjects' differences (y-axis) against the mean of their repeated measures (x-axis). In the unlikely situation of perfect agreement for test-retest data, the LoA analysis would yield a bias of 0.0 and 95% limits of +/- 0.0, with the lines overlapping. Be aware, however, that a zero bias can be accompanied by 'large' 95% limits since large positive differences can cancel out large negative differences. These statistics are not difficult to calculate (or plot) with commonly available statistical software, such as SPSS, Excel, or Minitab. However, this particular LoA analysis should only be performed if:

(i) the test-retest differences among the subjects are *normally* distributed, and if

(ii) the test and retest means are not significantly different, and if

(iii) there is no significant relationship between the test-retest differences (expressed without sign) and the test-retest means.

Without condition (i) being satisfied, the calculation of 1.96 x SD does not give a value that accounts for the differences of 95% of the sample. Again, most statistical software can test for the normality of data, and remove the need to interpret visually the shape of a histogram. If the data is not normal, it can be transformed (usually to a log scale), and artificially normalised. However, the LoA analysis on such data now takes a different form and is somewhat more complex. At this point I would urge the interested reader to refer to the article by Nevill and Atkinson (1997) for guidance.

With regard to condition (ii), Bland and Altman (1986) consider that a significant ($p < .05$) difference between test means is indicative of the occurrence of either systematic measurement error or a learning process. This implies a situation in which the test and retest have not been performed under identical conditions, making an assessment of reliability inappropriate. A paired (dependent) *t*-test will clarify if this condition is satisfied.

Condition (iii) ironically requires the Pearson correlation to be computed to assess whether the test-retest differences have a tendency to either increase or decrease in magnitude as the variable of interest increases. That is, with the above example of vertical jump performance, do those subjects with high scores tend to have larger test-retest differences than those with low scores, or vice versa? If a significant relationship is found, then any computed limits of agreement would not show this so-called heteroscedasticity and therefore not truly reflect the reliability of the measure. As with violations of condition (i), it is likely that log-transforming the data will eradicate the relationship (make it non-significant) and enable a LoA analysis to be performed.

The above 'impositions' may paradoxically make the LoA analysis seem rather involved, but nothing described above cannot be carried out by a computer literate investigator. The worth of such analysis emerges if one applies it to the vertical jump data presented above. The Shapiro-Wilks test revealed the test-retest differences to be normally distributed (S-W = .09; $p > .05$) and the paired *t*-test on the mean values yielded a non-significant result ($t = .32$; $p > .05$). The correlation between the differences and the mean scores was positive ($r = .17$), but non-significant ($p > .05$). Consequently, having a bias of .003 m and SD of .042 m, the limits of agreement were .003 m $\pm$ 1.96(.042), or .003 m $\pm$ .082 m. Unlike the Pearson or intraclass correlations ($r = 0.93$ and $R = 0.96$, respectively), this information does not warrant a favourable interpretation. That is, whilst the jump performances changed on average by less than half a centimetre, individual performances changed by as much as eight centimetres. Expressed as a proportion of the average performance

over the two tests, this represents a change of 21.6%. To most exercise scientists, this statistic would not be indicative of satisfactory reliability.

To date, LoA statistics for the test-retest reliability of commonly used exercise science and PE measures do not exist. Even in recent editions of the text books referred to above, where reliability is considered, the Pearson correlation is quoted. The protagonists of the LoA technique in exercise science have advocated the establishment of a data base of LoA statistics for the variety of measurement tools routinely used, and have begun the process by publishing the results from 23 separate studies performed at the Centre for Sport sand Exercise Sciences at Liverpool John Moores University, England (Nevill and Atkinson, 1997). These fairly small-scale studies (average n = 18) dealt with the reliability of 13 different popular measures of cardio-respiratory endurance, muscular strength, and anaerobic power. Now, whilst the authors proceeded to apply log-transformations to all the studies' data on account that some violated condition (i) and others condition (iii), their *initial* 95% LoA analysis revealed considerable variability between the test and retest scores of most variables. For instance, the Astrand-Ryhming cycle ergometer and Fitech step-tests gave estimates of maximal aerobic capacity that varied by as much as 52% and 42%, respectively, over two repeated trials. For the Fitech test, which had a test-retest correlation of r = 0.80, this meant that a subject with an estimate of 30 ml/kg/min in the first test could have a score as low as 15.5 ml/kg/min, or as high as 41.5 ml/kg/min in the second test (Nevill and Atkinson, 1997). Clearly, such uncertainty undermines the utility of this and similar estimates of aerobic capacity when one considers that the trainability of this particular physiological attribute is typically 15-20% for a previously sedentary person (Wilmore and Costill, 1994, p. 228).

Conclusion

This commentary has been written due to my concern that both researchers and consumers of PE-related research may not be aware of the recent and notable realisation regarding a previously unquestioned method of quantifying measurement reliability. The correlation coefficient should no

longer be used to assess absolute (agreement) reliability, especially as a more legitimate method is readily available. It may take time for the limits of agreement technique to become the routine method by which this pre-requisite of validity is judged, but hopefully the reasoning for its preference over the correlation coefficient has been made lucid. It is a bold step to 'go against the tide' of established practice, but in this case it is a move that is wholly justifiable. Even if the existing reliability statistics of measures happen to be merely corroborated by the LoA method, it would be comforting to know that the correct form of analysis had been used.

References

Adams, G.A. (1994) *Exercise Physiology Laboratory Manual*. Dubuque: Brown and Benchmark.

Atkinson, G. (1995) A comparison of statistical methods for assessing measurement repeatability in ergonomics research. In G. Atkinson & T. Reilly (Eds.), *Sport, Leisure and Ergonomics* (pp. 218-222). London: E. & F.N. Spon.

Bailey, S.M., Sarmandal, P., and Grant, J.M. (1989) A comparison of three methods of assessing inter-observer variation applied to measurement of the symphysis-fundal height. *British Journal of Obstetrics and Gynaecology*, 96, 1266-1271.

Baugartner, T.A. (1989) Norm-referenced measurement reliability. In M.J. Safrit & T.M. Wood (Eds.), *Measurement Concepts in Physical Education and Exercise Science* (pp. 45-72). Champaign: Human Kinetics Books.

Bland, J.M. and Altman, D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, i, 307-310.

Bartlett, R. (1997) The use and abuse of statistics in sport and exercise sciences. *Journal of Sports Sciences*, 15, 1-2.

Baumgartner, T.A. and Jackson, A.S. (1987) *Measurement for Evaluation in Education and Exercise Science*. Dubuque: Brown.

Bosco, J.S. and Gustafson, W.F. (1983) *Measurement and Evaluation in Physical Education, Fitness and Sport*. Englewood Cliffs: Prentice-Hall.

Cohen, L. and Holliday, M. (1979). *Statistics for Education and Physical Education*. London: Paul Chapman Publishing.

Morehouse, C.A., and G.A. Stull. (1975) *Statistical Principles and Procedures with Applications for Physical Education.* Philadelphia: Lea & Febiger.

Morrow, J.R., Jackson, A.W., Disch, J.G., and Mood, D.P. (1995). *Measurement and Evaluation in Human Performance*. Champaign: Human Kinetics Books.

Nevill, A.M. (1996) Validity and measurement agreement in sports performance. *Journal of Sports Sciences*, 14, 199.

Nevill, A.M. and Atkinson, G. (1997). Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *British Journal of Sports Medicine*, 31, 314-318.

Ottenbacher, K.J. and Tomchek, S.D. (1994) Measurement variation in method comparison studies: an empirical examination. *Archives of Physical Medicine and Rehabilitation*, 75, 505-512.

Safrit, M.J., and T.M. Wood (1989) *Measurement Concepts in Physical Education and Exercise Science*. Champaign: Human Kinetics Books.

Sarmandal, P., Bailey, S.M., and Grant, J.M. (1989) A comparison of three methods of assessing inter-observer variation applied to ultrasonic fetal measurement in the third trimester. *British Journal of Obstetrics and Gynaecology*, 96, 1261-1265.

Schabort, E.-J., Hopkins, W.G., and Hawley, J.A. (1988) Reproducibility of self-paced treadmill performance of trained endurance runners. *International Journal of Sports Medicine*, 19, 49-51.

Thomas, J.R. and Nelson, J.K. (1996). *Research Methods in Physical Activity*. Champaign:Human Kinetics Books.

Viitasalo, J.T. (1985) Effects of training on force-velocity characteristics for sportsmen in field conditions. In D.A. Winter, R.W. Norman, R.P. Wells, K.C. Hayes, & A.E. Patla (Eds.), *Biomechanics IX-A* (pp. 96-101). Champaign: Human Kinetics.

Wilmore, J.H. and Costill, D.L. (1994) *Physiology of Sport and Exercise*. Champaign: Human Kinetics.

Table 1.

Subject		Test 1	(Rank)	Test 2	(Rank)	Difference
1	24		(8)	33	(8)	+9
2	42		(5)	48	(5)	+6
3	19		(9)	30	(9)	+11
4	67		(3)	59	(3)	-8
5	40		(6)	36	(6)	-4
6	57		(4)	49	(4)	-8
7	85		(1)	75	(1)	-10
8	29		(7)	35	(7)	+6
9	15		(10)	10	(10)	-5
10	80		(2)	74	(2)	-6
Mean	45.8			44.9		

Table 2. Test-retest reliability of the vertical (Sargent) jump test amongst adults[1] (*n* = 27).

Subject	Test 1 (m)	(Rank)	Test 2 (m)	(Rank)	Difference
1	0.27	(23)	0.29	(22.5)	+.02
2	0.28	(22)	0.26	(24)	-.02
3	0.22	(25)	0.20	(26.5)	-.02
4	0.19	(27)	0.20	(26.5)	+.01
5	0.20	(26)	0.25	(25)	+.05
6	0.40	(11.5)	0.30	(20)	-.10
7	0.29	(20.5)	0.32	(17)	+.03
8	0.37	(16)	0.41	(10.5)	+.04
9	0.48	(6.5)	0.44	(8)	-.04
10	0.24	(24)	0.30	(20)	+.06
11	0.46	(8)	0.40	(12)	-.06
12	0.56	(1)	0.62	(1)	+.06
13	0.32	(18.5)	0.34	(15)	+.02
14	0.45	(9)	0.50	(6)	+.05
15	0.53	(3.5)	0.56	(2)	+.03
16	0.53	(3.5)	0.47	(7)	-.06
17	0.32	(18.5)	0.31	(18)	-.01
18	0.55	(2)	0.53	(3)	-.02
19	0.40	(11.5)	0.42	(9)	+.02
20	0.39	(14.5)	0.33	(16)	-.06
21	0.29	(20.5)	0.29	(22.5)	.00
22	0.40	(11.5)	0.36	(13)	-.04
23	0.48	(6.5)	0.51	(5)	+.03
24	0.39	(14.5)	0.41	(10.5)	+.02
25	0.52	(5)	0.52	(4)	.00
26	0.40	(11.5)	0.35	(14)	-.05
27	0.33	(17)	0.30	(20)	-.03
Mean	0.38		0.38		.035*

[1]Unpublished data
*Absolute mean difference (disregarding sign)

Figure 1. Bland and Altman plot of vertical jump data