Sports Med 2000 Nov; 30 (5): 375-381
0112-1642/00/0011-0375/$20.00/0

Typical Error Versus Limits of Agreement

Hopkins[1] provides a comprehensive narrative on the analysis of measurement error. One section of his paper is devoted to a comparison and contrast of two statistics that quantify error:

1. The standard error of measurement (SEM). This statistic is also known as the within-subject standard deviation. Hopkins[1] referred to it as 'typical error'.

2. The 95% limits of agreement (LOA), discussed most recently by Bland and Altman.[2]

Hopkins[1] (page 1) stated that LOA are 'biased' and 'more limited' than typical error, whilst he also maintained that 'they [LOA and typical error] are too closely related' to be cited together in a publication[1] (page 5). This is a contradiction that has resulted from a confusion over reference and confidence intervals, together with a reluctance to compare both statistics at the same level of abstraction. We will show that when typical error is defined properly, there is, in reality, little difference between it and LOA, since both statistics are *probabilistic reference intervals* for measurement error. We will also show that it is erroneous to think that LOA are biased in the way that was suggested.

Limits of Agreement Are Not Confidence Limits

The LOA represent a reference interval (also known as a 'normal' range) for the test-retest differences expected for 95% of individuals in a population. Clinicians use reference intervals routinely to make probability statements for expected values using the known relationship between the standard deviation and centiles of a normally distributed population. The most common coverage probability for reference intervals is 0.95. Full details on the calculation and application of reference intervals are given by Wright and Royston.[3] Confidence limits (CL) are, on the other hand, the upper and lower boundaries within which a *population parameter* is thought to lie, given a particular statistic derived from a random sample. The most common coverage confidence for CL is also 0.95. The width of CL is dependent on sample size and the t-statistic is employed for calculation of inference to the population parameter.

As Bland and Altman[2] (page 139) noted recently, '. . . despite the superficial similarity these [LOA] are not the same thing as CL, but like a reference interval.' Hopkins[1] made the mistake of thinking that LOA represent CL for a population parameter, and he applied the t-statistic to the calculation of the *sample* LOA. We stress that the LOA is derived from a sample standard deviation which should be left alone. In agreement with Hopkins,[1] Bland[4] and Altman[5] recommended sample sizes of at least 50 individuals in a study in order for the sample LOA to be a precise estimate of the population LOA. Bland and Altman[2,6] also showed how CL can be calculated to provide an indication of the *precision* of the sample LOA for the inference to a population LOA. It is in these calculations *only* that use of the t-statistic is appropriate. As well as the underlying logic for the application of a t-value to a sample statistic being erroneous, Hopkin's 'correction' of a sample LOA using the t-statistic would lead to a marked overestimation of the population LOA if it was employed by researchers.[7]

'Typical Error' is a 68% Reference Interval for the True Score

Hopkins[1] echoed the advice of Atkinson and Nevill[8] in stressing the need for researchers to discuss the implications of measurement error for real applications of the measurement tool. In this respect, it is important for a measurement error statistic to have practical, unambiguous meaning. The SEM, as a statistic used in classical test theory, allows one to make probability statements about the variability of measurements for individual subjects in a population.[9] Nevertheless, Hopkins[1] considered the SEM simply as typical error without consulting the wealth of literature in the sport and exercise sciences which provides a more informative definition of the statistic. For example, Thomas and Nelson[10] (page 232) provided the classical defini-

tion: 'Standard errors [of measurement] are assumed to be normally distributed and are interpreted in the same way as standard deviations. About two-thirds (68.26%) of all test scores will fall within plus or minus one standard error of measurement of their true scores'.

From the above definition, it is clear that typical error, just like LOA, represents a reference interval. The difference between the 2 statistics is that LOA represent a reference interval for test-retest variability whereas typical error represents a reference interval for true score error. It is also clear from the above definition of SEM that the 'typical' that is considered by Hopkins[1] actually means a reference interval with a coverage probability that approximates to 68% of all true scores in a population.[9]

Typical Error Covers Only 52% of Test-Retest Differences in a Population

Hopkins[1] (pages 2 to 3) stated that typical error represents 'the variation we would expect to see from trial to trial if *any* of these participants [the individuals tested] performed multiple trials'. This denotes a serious confusion over the underlying theory of SEM. Ironically, Hopkins[1] is interpreting the SEM as a LOA for test-retest differences here. The SEM is calculated from the standard deviation (SD) of differences divided by the square root of 2 (or $0.707 \times$ SD of differences). One standard deviation covers 68% of the differences. Seventenths (0.707) of a standard deviation covers about 52% of differences. If the typical error is said to represent 'test-retest error' or 'trial-to-trial variation' as Hopkins suggests, researchers should be aware that the statistic actually describes the test-retest variability for only 52% of individuals in a population. We doubt whether this coverage probability should be defined as 'typical'. The underlying theory of SEM is based on a meaningful description of true score error, not test-retest variability.

Hypothetical Retests and True Scores

Hopkins[1] criticised the use of LOA for encouraging the researcher to conceptualise a hypothetical test-retest situation in the discussion of the impact of measurement error. We stress that the incorporation of typical error, if defined properly, into such a discussion is just as hypothetical, that is the researcher considers a hypothetical and practically unobtainable 'true score' to help with his/her reliability decisions.

Choice of Level of Probability for Measurement Error Statistics

Despite Hopkins'[1] dislike for 95% probability intervals, he should be aware that several other sport and exercise statisticians have hinted that the SEM might not be 'typical' enough.[10,11] These authors detailed how one can multiply the SEM by 1.96 to represent the 95% reference interval for the true score (a 95% SEM). Differences in interpretation of what is 'typical' illustrate the importance of considering the SEM as a probabilistic reference interval with defined probability rather than just typical error, which is open to interpretation and encourages researchers not to appreciate what an error statistic actually means. Interested readers should consult Wyrwich and Wollinsky[12] for the latest discussion on probability level and the application of measurement error statistics to 'true' changes in individuals. We agree that the conventional coverage probability of 0.95 might not be appropriate in all situations. Nevertheless, there are cases in which researchers need to be more (not less) confident in their measurements (e.g. drug testing of athletes).

Error Statistics and 'Monitoring an Individual'

Harvill[9] and Eliasziw et al.[13] discussed how the SEM is multiplied by the square root of 2 to examine whether measurements from 2 individuals are really different from each other, or whether a change in measurement within the same person is likely to be caused by measurement error or not. Interestingly, such a calculation effectively converts

the SEM to the standard deviation of the differences. This conversion makes sense, since the situations here involve 2 measurements which are *both* measured with error. For a comparison of a measurement to a cut-off value, Harvill[9] also showed how the SEM per se was not appropriate. Conversely, Mathews[14] showed how LOA can be used to calculate the probability of classifying individuals wrongly on the basis of comparing a measurement to a cut-off value. Hopkins[1] does not discuss these important issues for which LOA are appropriate. Importantly, it is the researchers who have provided tutorials on the use of SEM who show how the statistic is converted to the LOA in order for differences between and within individual subjects to be investigated.[9,13]

Summary

We have shown that the SEM (typical error) and LOA are very similar when defined at the same level of abstraction. The calculation of these sample statistics does not depend on sample size, but the precision of their estimate for the population parameter does. Only the latter concept involves the t-statistic. They *do* differ in the type of measurement error that is described (true score error versus test-retest error) and the coverage probability of the reference interval (0.68 versus 0.95). Bland and Altman[2] and Atkinson and Nevill[8] have promoted the citation of *either* the SEM or the LOA to help researchers in their discussion of the impact of error to real uses of the measurement tool. What is vital in this discussion is the researcher having a thorough understanding of the underlying theory behind the measurement error statistic(s) that is/are employed, especially the definition of error and the coverage probability that is selected. Such issues have been built into the title of the 95% LOA statistic, and are also an inherent part of SEM. Whilst the concept of typical error appears to be easy to understand and teach, this is only because the underlying theory and definition of what the statistic actually represents is not communicated. Only if all researchers adopt one single statistic (e.g. typical error) for measurement error is it remotely possible to push underlying theory into the background, since there would be a baseline of comparison for all. Because this scenario is highly unlikely, it is important that any measurement error statistic is well defined and understood by researchers.

Greg Atkinson
Research Institute for Sport and
Exercise Sciences
Liverpool John Moores University
Liverpool, England
Alan Nevill
School of Performing Arts and Leisure
University of Wolverhampton
Walsall, England

References

1. Hopkins W. Measures of reliability in sports medicine and science. Sports Med 2000; 30: 1-15
2. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999; 8: 135-60
3. Wright EM, Royston P. Calculating reference intervals for laboratory measurements. Stat Methods Med Res 1999; 8: 93-112
4. Bland M. An introduction to medical statistics. Oxford: University Press, 1995
5. Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991
6. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; I: 307-10
7. Royston P, Matthews JNS. Estimation of reference ranges from normal samples. Stat Med 1991; 10: 691-5
8. Atkinson G, Nevill AM. Statistical methods in assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998; 26: 217-38
9. Harvill LM. An NCME instructional module on standard error of measurement. Educ Meas Iss Pract 1991; 10 (2): 33-41
10. Thomas JR, Nelson JK. Research methods in physical activity. Champaign (IL): Human Kinetics Books, 1996
11. Morrow JR, Jackson AW, Disch JG, et al. Measurement and evaluation in human performance. Champaign (IL): Human Kinetics, 1995
12. Wyrwich KW, Wollinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. J Eval Clin Pract 2000; 6: 39-49
13. Eliasziw M, Young SL, Woodbury MG, et al. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther 1994; 74: 8: 777-88
14. Mathews JN. A formula for the probability of discordant classification in method comparison studies. Stat Med 1997; 16 (6): 705-10