

Physiolometrics and the puzzle of methodical acumen

'It is of the highest importance in the art of detection to be able to recognize, out of a number of facts, which are incidental and which vital!' These words, spoken by Sherlock Holmes in *The Adventure of the Reigate Squire* (later also known as *The Adventure of the Reigate Puzzle*) published in the June issue of *The Strand Magazine* in 1893, resonate with our quest as physiologists for methodological rigor to this day. It is tempting to speculate that John Newport Langley (1852–1925), who during that period assumed control of *The Journal of Physiology* and became its most notorious Editor-in-Chief to date (Bailey et al., 2023), much like his contemporaries in The Physiological Society, found inspiration in the unwavering pursuit of deductive excellence exemplified by Sherlock Holmes.

At the turn of the century, Langley's insistence on a meticulously structured approach to the analysis and presentation of data fundamentally altered physiology as an experimental science, and continues to do so to this day (Bailey et al., 2023). During those years, many prevailing theories such as the unitary view of brain function (i.e., absence of localised function), the theory of urine secretion in the kidney, and the concept of oxygen secretion in the lungs, to mention a few, were revisited. One by one, they were debunked by uncovering a flawed empirical basis due to methodological limitations or the development of more accurate experimental models and measurement techniques to help advance understanding (Krogh, 1910; Langley & Sherrington, 1884; Starling, 1899). Indeed, the discipline focused on the advancement and systematic evaluation of physiological measurement techniques, which we refer to as *physiolometrics*, is essential to modern physiology. Here, we will briefly define the main physiometric concepts of validity and reliability and describe how they may be used to systematically evaluate both existing and new physiological methods.

Validity pertains to the ability of a method to accurately reflect the true physiological variable. In physiometrics, it mainly encompasses three domains: logical validity, construct validity and criterion validity (George et al., 2000). *Logical validity* entails a qualitative evaluation of whether the measurement method, principles and assumptions align with providing meaningful measurements, based on established principles of physics, biochemistry, biology and physiology. However, while an assessment of logical validity can be used to infer whether the theoretical basis and measurement principle of a method is fundamentally flawed, it does not provide any quantitative measure of a method's validity.

In contrast to logical validity, construct and criterion validity are based on quantitative statistics-based assessments. These quantitative measures include sensitivity and specificity of the method for detecting a given change in a physiological parameter, and an assessment of systematic error – measurement bias. Measurement bias is defined as predictable deviations from the true value, arising from flaws or biases in the measurement process, such as faulty equipment, measurement instrument calibration issues, or biased measurement procedures. *Construct validity* examines how well a measurement (the 'construct') behaves in response to various physiological stimuli and in relation to other physiological parameters. Construct validity can be evaluated using simple paired tests to determine whether relevant changes in the measure to a given intervention occurred simply by chance alone, or for example, by correlational analyses to assess whether the measure behaves as expected in relation to other variables (Cronbach & Meehl, 1955). Discriminant validity is a subdomain of construct validity that explores the measure's ability to distinguish between individuals expected to differ. In the case of measures that have established thresholds, sensitivity and specificity can be employed, categorizing the measure as either normal or abnormal. For continuous measures without established thresholds, construct validity can be investigated using receiver operating characteristic analysis and calculating the area under the curve (Olsen et al., 2022).

Criterion validity involves the comparison of a given method to an established reference (or criterion) method. Importantly, the formal assessment of criterion validity should not rely on simple difference tests, as this approach obscures individual variability, or on correlation techniques as these are not suitable for repeated assessments of the same variable, and do not account for systematic error. The most suitable approach for assessing criterion validity is arguably the Bland–Altman plot, where the difference between each measurement obtained by the two methods is plotted against the mean of the two with the latter considered to represent the best estimate of the 'true' value (Bland & Altman, 1986; Bunce, 2009). This offers direct estimates of systematic error (bias) between the methods using the 'limits of agreement' which essentially represent a range of validity, given that the data do not demonstrate heteroscedasticity, that is, that the magnitude of error does not increase with higher measured values. A complementary approach is the use of the intraclass correlation coefficient (ICC), but this does have considerable limitations and pitfalls as outlined below in relation to the assessment of reliability. Furthermore, it must be noted that neither Bland–Altman plots nor

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Experimental Physiology* published by John Wiley & Sons Ltd on behalf of The Physiological Society.

ICCs provide any clues as to which method is superior when there is evidence of substantial systematic error. In such cases, it may be necessary to assess how the measure of interest obtained by the two methods predicts a given outcome.

Reliability focuses on the consistency of measurement and aims to quantify both inherent and random errors of the measurement and encompasses aspects such as test–retest reliability and inter-rater reliability. Furthermore, reliability pertains to either the consistency of measurement when repeated under identical experimental conditions (repeatability) or its consistency across varying experimental conditions (reproducibility) (Bartlett & Frost, 2008). Random error refers to unpredictable fluctuations or variations in measurement that occur randomly and without any consistent pattern. It is typically caused by chance factors, such as human error, natural variability, or environmental influences that affect the measurement process. As for criterion validity, the use of Pearson's correlation to assess reliability can create an exaggerated impression of reliability and should be avoided. Paired tests between the mean values in the case of test–retest data do not directly inform reliability, since any random error will add to the variability within and between the repeated measure, and thus increase rather than decrease the *P*-value. Thus, paired tests should only be used as a 'gatekeeper method' to ensure internal consistency of the data, that is, to rule out systematic errors at the group level that will render the given dataset inappropriate for test–retest reliability assessment (Silbernagel et al., 2005). Reliability can be evaluated in terms of absolute and relative reliability, where absolute reliability focuses on the magnitude or extent of measurement error quantified in the same unit as the measure of interest, whereas relative reliability provides information about the relative contribution of measurement error to the overall variation in the measure and is usually provided as a percentage.

Absolute reliability can be assessed through a Bland–Altman plot as outlined above, as well as by calculation of the closely related smallest real difference (SRD). SRD is easily interpretable because it provides an estimate of the maximal difference there will be between two measurements on 95% of occasions (Vaz et al., 2013).

A commonly used relative reliability estimate is the coefficient of variation (CV) reflecting the relationship between the standard deviation within the group (also known as the standard error of measurement) and the mean (Searls, 1964), and in reliability studies it will furthermore often be relevant to calculate 95% confidence intervals of the CV, which is possible by a simulation procedure based on the distribution of the estimates of mean and residual variance from a linear mixed model (Liu, 2012). Caution should, however, be exercised when the mean is close to zero, as it can lead to inaccurate results (e.g., infinite CV regardless of standard error of measurement). Furthermore, if the CV is reported in studies measuring graded steps, as often done in exercise physiology, the CV will often decrease as intensity increases, because the mean value of the estimate increases relatively more than the standard error of measurement. As for many other physiometrics, the cut-off values for acceptable and unacceptable CVs are somewhat arbitrary and depend on the research field.

The ICC, which encompasses several subtypes depending on scope and type of data, is another widely used measure for relative reliability and can conveniently be used to classify the reliability of a method categorically as poor, moderate, good or excellent (Koo & Li, 2016). However, the ICC should be reported and interpreted cautiously due to its sensitivity to variations within and between groups (Madsen et al., 2023). This basically means that if the ICC is reported on a very heterogeneous population, a high within-group standard deviation may lead to a high ICC no matter how flawed the method is. Hence, the ICC alone does not offer a comprehensive assessment of reliability and should not be considered independently of other reliability measures. A more rarely used reliability measure is the root mean square deviation coefficient, which provides a measure of the overall deviation (or error) between the observed and predicted values. It can also be applied for the test–retest study set-up, but one should be cautious because the measure is very sensitive to outliers and does not differentiate between random and systematic error (Barchard, 2012). For the assessment of absolute and relative reliability, we have recently developed a publicly available *calcrel* function in the *clintools* package in R (<https://cran.r-project.org/web/packages/clintools/index.html>).

When considering (relative) reliability, physiological biomarkers are dynamic metrics subject to natural variation. This variation encompasses both analytical and, more prominently, biological components that collectively contribute to what is known as the *critical difference* (Fraser & Fogarty, 1989). This describes random variation around a homeostatic point, indicative of the change that must occur before a true physiologically or clinically relevant difference can be claimed. While this concept emanates from the field of clinical biochemistry applied to metabolic biomarkers of exercise stress (Davison et al., 2012), it is less well known in clinical and other fields of physiology where it can optimise the interpretation and stratification of responses (Rose et al., 2018). Furthermore, the critical difference may be an important consideration in interventional studies with responder analyses where claimed 'responders' may simply be artefacts within a defined zone of natural variation.

As was the case 130 years ago when Langley may have read the works of Sir Arthur Conan Doyle, the distinction between incidental and vital facts, that is, between measurement error and the true underlying value, remains a defining feature of physiology as an experimental science. So, for us as physiologists to appropriately design and conduct studies and provide results that uncover the fundamental mechanisms of life, both in the healthy and diseased state, we need to embrace physiometrics. This will help ensure that rather than being discarded due to methodological and/or statistical flaws, our findings may remain relevant for future generations of physiologists, even as they develop and overturn today's major concepts and theories as part of the self-correcting nature of science (Whipp, 2010). Thus, the systematic and critical assessment of the validity and reliability of our methods is a *sine qua non* if our results are to stand the test of time. Or as Sherlock Holmes would put it himself: 'It is elementary!'

AUTHOR CONTRIBUTIONS

All authors have read and approved the final version of this manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All persons designated as authors qualify for authorship, and all those who qualify for authorship are listed.

FUNDING INFORMATION

The Centre for Physical Activity Research is supported by TrygFonden Grants ID 101390, ID 20045, and ID 125132. J.P.H. is funded by HelseFonden and Rigshospitalet. D.M.B. has received funding from a Royal Society Wolfson Research Fellowship (#WM170007) and the Higher Education Funding Council for Wales.


CONFLICT OF INTEREST

None of the authors have any competing interests to disclose.

Jacob Peter Hartmann^{1,2}

Markus Harboe Olsen³

George Rose⁴ 

Damian M. Bailey⁴ 

Ronan M. G. Berg^{1,2,4,5} 

¹Centre for Physical Activity Research, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark

²Department of Clinical Physiology and Nuclear Medicine, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark

³Department of Neuroanaesthesia, The Neuroscience Centre, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark

⁴Neurovascular Research Laboratory, Faculty of Life Sciences and Education, University of South Wales, Pontypridd, UK

⁵Department of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Correspondence

Ronan M. G. Berg, Neurovascular Research Laboratory, Faculty of Life Sciences and Education, University of South Wales, Alfred Russel Wallace Building, Pontypridd CF37 4AT, UK.

Email: ronan@sund.ku.dk

Jacob Peter Hartmann and Markus Harboe Olsen contributed equally as first authors.

Handling Editor: David Poole

ORCID

George Rose  <https://orcid.org/0000-0002-9598-6372>

Damian M. Bailey  <https://orcid.org/0000-0003-0498-7095>

Ronan M. G. Berg  <https://orcid.org/0000-0002-5757-9506>

REFERENCES

Bailey, D. M., Berg, R. M. G., Stewart, A., Adams, J. C., & Kohl, P. (2023). Sharpey-Schafer, Langley and Sherrington: 'Swordsmen' of physiology. A historical look to the future. *Experimental Physiology*, 108(5), 655–658.

- Barchard, K. A. (2012). Examining the reliability of interval level data using root mean square differences and concordance correlation coefficients. *Psychological Methods*, 17(2), 294–308.
- Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. *Ultrasound in Obstetrics & Gynecology*, 31(4), 466–475.
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327(8476), 307–310.
- Bunce, C. (2009). Correlation, agreement, and Bland-Altman analysis: Statistical analysis of method comparison studies. *American Journal of Ophthalmology*, 148(1), 4–6.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Davison, G. W., Ashton, T., McEneny, J., Young, I. S., Davies, B., & Bailey, D. M. (2012). Critical difference applied to exercise-induced oxidative stress: The dilemma of distinguishing biological from statistical change. *Journal of Physiology and Biochemistry*, 68(3), 377–384.
- Fraser, C. G., & Fogarty, Y. (1989). Interpreting laboratory results. *British Medical Journal*, 298(6689), 1659–1660.
- George, K., Batterham, A., & Sullivan, I. (2000). Validity in clinical research: A review of basic concepts and definitions. *Physical Therapy in Sport*, 1(1), 19–27.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- Krogh, A. (1910). On the mechanism of the gas-exchange in the lungs. *Skandinavisches Archiv Für Physiologie*, 23(1), 248–278.
- Langley, J. N., & Sherrington, C. S. (1884). Secondary degeneration of nerve tracts following removal of the cortex of the cerebrum in the dog. *The Journal of Physiology*, 5(2), 49–126.
- Liu, S. (2012). *Confidence interval estimation for coefficient of variation (thesis)*. Georgia State University. https://scholarworks.gsu.edu/math_theses/124
- Madsen, A. C., Thomsen, R. S., Nymand, S. B., Hartmann, J. P., Rasmussen, I. E., Mohammad, M., Skovgaard, L. T., Hanel, B., Jønk, S., Iepsen, U. W., Chistensen, R. H., Mortensen, J., & Berg, R. M. G. (2023). Pulmonary diffusing capacity to nitric oxide and carbon monoxide during exercise and in the supine position: A test-retest reliability study. *Experimental Physiology*, 108(2), 307–317.
- Olsen, M. H., Riberholt, C., Plovsing, R. R., Berg, R. M. G., & Møller, K. (2022). Diagnostic and prognostic performance of Mxa and transfer function analysis-based dynamic cerebral autoregulation metrics. *Journal of Cerebral Blood Flow and Metabolism*, 42(11), 2164–2172.
- Rose, G. A., Davies, R. G., Davison, G. W., Adams, R. A., Williams, I. M., Lewis, M. H., Appadurai, I. R., & Bailey, D. M. (2018). The cardio-pulmonary exercise test grey zone; optimising fitness stratification by application of critical difference. *British Journal of Anaesthesia*, 120(6), 1187–1194.
- Searls, D. T. (1964). The utilization of a known coefficient of variation in the estimation procedure. *Journal of the American Statistical Association*, 59(308), 1225–1226.
- Silbernagel, K. G., Thomeé, R., & Karlsson, J. (2005). Cross-cultural adaptation of the VISA-A questionnaire, an index of clinical severity for patients with Achilles tendinopathy, with reliability, validity and structure evaluations. *BMC Musculoskeletal Disorders*, 6(1), 12.
- Starling, E. H. (1899). The glomerular functions of the kidney. *The Journal of Physiology*, 24(3-4), 317–330.
- Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test-retest reliability. *PLoS ONE*, 8(9), e73990.
- Whipp, B. J. (2010). D.B. Dill Historical Lecture: The self-correcting nature of science. In *57th Annual Meeting of The American College of Sports Medicine*.