



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/175566/>

Article:

Summerfield, Quentin, Kitterick, Pádraig and Goman, Adele (2022) Development and critical evaluation of a condition-specific preference-based measure sensitive to binaural hearing in adults:the York Binaural Hearing-related Quality of Life System. *Ear and Hearing*. pp. 379-397. ISSN: 1538-4667

<https://doi.org/10.1097/AUD.0000000000001101>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Supplementary Digital Content 8

Tests of reproducibility

(This document is supplementary to the paper by Summerfield, Kitterick, and Goman entitled 'Development and critical evaluation of a condition-specific preference-based measure sensitive to binaural hearing in adults: the York Binaural Hearing-related Quality of Life System'.)

1. Introduction

1.1. In the paper, we report statistical tests of *reproducibility*. This supplementary digital content explains the rationale for these tests and describes how we implemented them.

2. Measures of Reproducibility: Agreement and Reliability

2.1. Introduction

2.1.1. Measures obtained at test and retest can be compared both for their *agreement* and for their *reliability*. Together, indices of agreement and reliability provide evidence of *reproducibility*.

2.1.2. de Vet et al. (2006, p. 1034) explained the difference between agreement and reliability as follows. "Indices of agreement address the question: 'How good is the agreement between repeated measurements?' This question concerns the measurement error and assesses exactly how close the scores for repeated measurements are. In comparison, indices of reliability address the question: 'How reliable is the measurement?' in other words, how well can patients be distinguished from each other, despite the measurement errors? In this case, the measurement error is related to the variability between study objects."

2.1.3. de Vet et al. refer to Guyatt et al. (1987) who noted that indices of agreement are required for instruments that are used for evaluative purposes, while indices of reliability are required for instruments that are used for discriminative purposes. We intended that the YBHRQL should demonstrate both types of reproducibility. Therefore, we report both types of index in the paper.

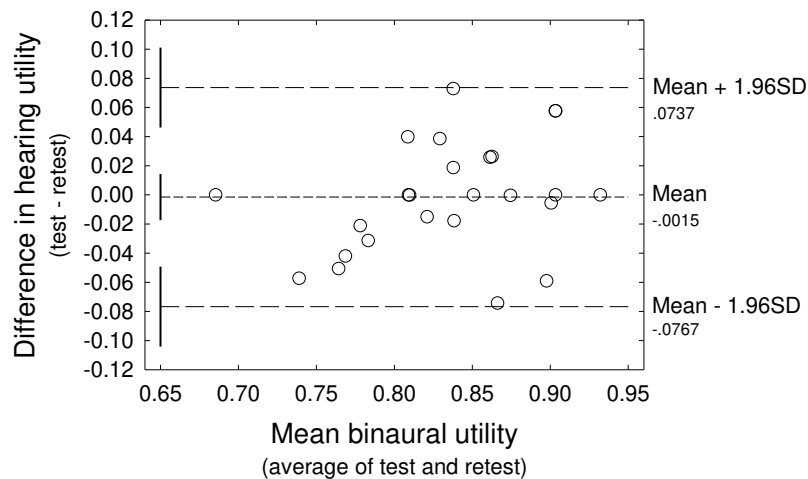
2.1.4. In this supplementary digital content, we first apply the method of Bland and Altman (1986) to estimate the limits of agreement between values of binaural utility from test and retest. We then describe the application of the methods of de Vet et al. (2006) to estimate an index of agreement as the standard error of measurement (Musselwhite & Wesolowski 2018) and an index of reliability as an intra-class correlation coefficient.

2.2. Limits of agreement: Bland and Altman

2.2.1. Bland and Altman (1986) recommended an approach to assessing agreement based on the differences between measures obtained in two test sessions. The sessions might involve two different test methods conducted more or less at the same time, or the same method used on two separate occasions.

2.2.2. The approach can be summarized in a graph of the difference between measures plotted against their mean in what is known as a 'Bland-Altman Plot' (e.g. Giavarina, 2015). Figure 2 is a Bland-Altman Plot for values of binaural utility from the YBHRQL at test and retest.

Figure 2 Bland-Altman plot of measures of binaural utility at test and retest.



2.2.3. The plot has three relevant properties.

2.2.3.1. The central horizontal line composed of short dashes plots the bias between test and retest, calculated as the mean difference between test and retest (-.0015). The heavy vertical line at the left-hand end of the central dashed line plots the 95% confidence interval of the mean (-.0173 to .0143). The interval includes zero. Thus there is no statistically significant evidence of a bias.

2.2.3.2. The upper and lower horizontal lines composed of long dashes plot the mean difference plus and minus 1.96 standard deviations. If the differences between test and retest distribute normally, then 95% of the differences would be expected to fall between these limits. It can be seen that this is the case. A Shapiro Wilk Test confirms the normality of the differences (statistic = .97, df=25, p=.65).

2.2.3.3. The limits themselves extend from -.0767 to .0737. This range is one measure of agreement between values of binaural utility at test and retest. The heavy vertical lines at the left-hand ends of the lines marking the limits plot the 95% confidence intervals of the limits. In the worst case, the limits might extend from about -.10 to .10.

2.2.3.4. We defer an assessment of the informativeness of this analysis to the Summary in Section 3.4.

2.3. Indices of agreement and reliability: de Vet et al.

2.3.1. The index of reproducibility that is most often reported in the literature is an intra-class correlation coefficient (ICC). Its generic formula is:

$$ICC = \frac{\text{Variability due to study objects}}{\text{Variability due to study objects} + \text{Measurement error}}$$

2.3.2. Some authors (e.g. McGraw & Wong, 1996) recommend a different, but related, generic formula in which the numerator and denominator are exchanged and the measurement error is subtracted from the variance due to study objects. In either case, the smaller the measurement error, the closer the ICC is to unity. Because the ICC depends on the variability among study objects – e.g. patients – it is an index of

reliability. It will only reproduce across studies if the heterogeneity of study objects is the same in the different studies.

2.3.3. de Vet et al. recommend the standard error of measurement (SEm) should be used as a measure of agreement and that the appropriate intraclass correlation coefficient (ICC) should be used as a measure of reliability. de Vet et al. note that the terms “agreement” and “consistency” are used (highly confusingly) in the literature to characterize variants of the ICC which, they emphasise, is always a measure of reliability, not agreement.

2.3.4. de Vet et al. described the analysis of a situation where two physiotherapists each made measurements of the flexibility of the shoulders of a group of patients. De Vet et al. advocated the following formulae for calculating an SEM and an ICC.

$$SEM = \sqrt{(\sigma_{pt}^2 + \sigma_{residual}^2)} \quad \text{Eqn 1}$$

$$ICC = \frac{\sigma_p^2}{(\sigma_p^2 + \sigma_{pt}^2 + \sigma_{residual}^2)} \quad \text{Eqn 2}$$

The components of the equations are measures of variance. σ_p^2 is the variance due to patients, σ_{pt}^2 is the variance due to the physiotherapists, and $\sigma_{residual}^2$ is the variance due to the interaction between patients and physiotherapists.

2.3.5. The equations can be re-expressed in terms of mean squares with a generic naming convention as follows:

$$SEm = \sqrt{(MS_{conditions} + MS_{error})} \quad \text{Eqn 3}$$

$$ICC = \frac{MS_{subjects}}{(MS_{subjects} + MS_{conditions} + MS_{error})} \quad \text{Eqn 4}$$

Table 3 lists the key values in the calculation of the indices of agreement and reliability. Applying Eqn 3, the SEM is estimated to be .028. Applying Eqn 4, the ICC is estimated to be .903.

Table 3 Key values in the calculation of indices of agreement and reliability

Component of sum of squares	Sum of squares	Degrees of freedom	Mean square	From de Vet et al.
Total	0.18787410941059	(N-1)	49	
Conditions	0.00002796928289	(k-1)	1	0.0000279692829 σ_{pt}^2
Within	0.18784614012770	(N-k)	48	0.0039134612527
Subjects	0.17019161031153	(n-1)	24	0.0070913170963 σ_p^2
Error	0.01765452981616	(k-1)(n-1)	24	0.0007356054090 $\sigma_{residual}^2$

Component of degrees of freedom	Value
Conditions (k)	2
Subjects (n)	25
Data points (N)	50

2.3.6. Readers may have noted that a different value of the ICC is reported in the paper. The difference arises because the formula for the ICC illustrated by de Vet et al. is generic, whereas specific formulae are advocated depending on the nature of the measures

that are to be compared and on the ways in which they were obtained. de Vet et al. noted that this was the case and recommended the paper by McGraw and Wong (1996) as a source of guidance on specific formula. For comparison of scores from a questionnaire that is completed on two occasions by the same group of patients, the recommended model is “2-way mixed effects, absolute agreement, single measurement”, denoted ICC(A,1) by McGraw and Wong. The formula for this version of the ICC is:

$$ICC(A, 1) = \frac{MS_{subjects} - MS_{error}}{MS_{subjects} + (k-1)MS_{error} + \left(\frac{k}{n}\right)(MS_{conditions} - MS_{error})} \quad \text{Eqn 5}$$

Substituting values from Table 3 into Eqn 5 gives the ICC(A,1) as 0.818 which is the value reported in the paper.

2.4. Reproducibility: Summary

2.4.1. The approach to assessing agreement set out by Bland and Altman is particularly relevant where a new measurement method is compared with an established gold standard. Knowledge of the limits of agreement can guide decisions about whether the new test is sufficiently accurate to be allowed to guide clinical judgement. We included a Bland-Altman plot in this supplementary digital content for completeness but judge that the indices of agreement and reliability described by de Vet et al. are more relevant when assessing the reproducibility at re-test of a self-report measure. As such, the SEM of the YBHRQL, at .03, is similar in size to values of the MCID reported for the HUI3 and the EQ-5D-3L. An ICC of 0.82 with a 95% confidence interval extending from .63 to .92 would mean that test-retest reliability was good, with a confidence interval ranging from moderate to excellent according to the criteria set out by Koo and Li (2016).

References

- Bland, J.M., Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* i, 307-310.
- Coretti, S., Ruggeri, M., McNamee, P. (2014). The minimum clinically important difference for EQ-5D Index: a critical review. *Expert Rev Pharmacoecon Outcomes Res.* 14, 221-233.
- Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica* 25, 141-151.
- Guyatt, G., Walter, S., Norman, G. (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis.* 40, 171-178.
- Horsman, J., Furlong, W., Feeny, D., Torrance, G. (2003). The Health Utilities Index (HUI®): concepts, measurement properties and applications. *Health Qual Life outcomes* 1, 1-13.
- Koo, T.K., Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15, 155-163.
- Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* 8, 335-362.
- McGraw, K.O., Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1, 30-46.
- Musselwhite, D.J., Wesolowski, B.C. (2018). Standard error of measurement. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, B.B. Frey (Ed.). Thousand Oaks: SAGE Publications Ltd. pp. 1588-1590.

de Vet, H.C.W., Terwee, C.B., Knol, D.L., Buter, L.M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology* 59, 1033-1039.