Reliability and Validity of Speech Evaluation in Adductor Spasmodic Dysphonia: Common Mistake and Statistical Issues

To the Editor:

I was interested to read the paper by Yanagida et al published in *Journal of Voice* on August 2017.[1] The purposes of the authors were to evaluate speech in patients with adductor spasmodic dysphonia (ADSD) by perceptual evaluations and acoustic measures, and to examine the reliability and validity of these measures.[1] Twenty-four patients with ADSD and 24 healthy volunteers were included in the study. Speech materials consisted of three sentences constructed from serial voiced syllables to elicit abductor voice breaks. Three otolaryngologists rated the degree of voice symptoms using a visual analog scale (VAS). VAS sheets with five 100-mm horizontal lines were given to each rater.[1] To evaluate the intra- and inter-rater and intermeasurer reliabilities of the VAS scores or acoustic measures, Pearson r correlations were calculated. To examine the validity of perceptual evaluations and acoustic measures, the sensitivity and the specificity were calculated.[1] Based on their results, Pearson r correlation coefficients for overall severity showed the highest intra- and inter-rater reliabilities. For acoustic events, the intermeasurer reliabilities were r = 0.64 (frequency shifts), r = 0.96 (aperiodic segments), and r = 1.0 (phonation breaks), and the intermeasurer reliability ranged from r = 0.10 to r = 1.0 Perceptual evaluation showed high sensitivity (91.7%) and specificity (100%), whereas acoustic analysis showed low sensitivity (70.8%) and high specificity (100%).[1]

Reliability (precision, repeatability, or reproducibility) is being assessed by different statistical tests such as Pearson r, which is one of the common mistakes in reliability analysis.[2] Pearson r just assesses the linearity and we can have a linear correlation with no reliability at all! Figure 1 shows that any shift in location and scale of the regression line, which indicates no reliability, cannot be detected by Pearson r. Moreover, in reliability analysis, our approach should be individual based instead of global average, and Pearson r cannot cover this approach.[2–11] Therefore, for quantitative variable, intraclass correlation coefficient single measure should be used. Regarding validity, depending on the type of the variable (qualitative or quantitative), well-known statistical test can be applied.[3–11] Although they correctly applied sensitivity, specificity, positive predictive value, and negative predictive value, they could consider other complementary estimates to assess validity. Likelihood ratio positive (ranging from 1 to infinity; the higher the LR+, the more accurate the test), likelihood ratio negative (ranging from 0 to 1; the lower the LR–, the more accurate the test), and odds ratio (ratio of true to false results) are the most appropriate estimates to evaluate validity of a test compared with a gold standard.[2,8,10] In case of quantitative variable, depending on the distribution of the variable, Pearson r or Spearman rho can be applied. In conclusion, to assess reliability and validity, appropriate test and correct interpretation should be considered. Otherwise, misdiagnosis and mismanagement of the patients cannot be avoided.

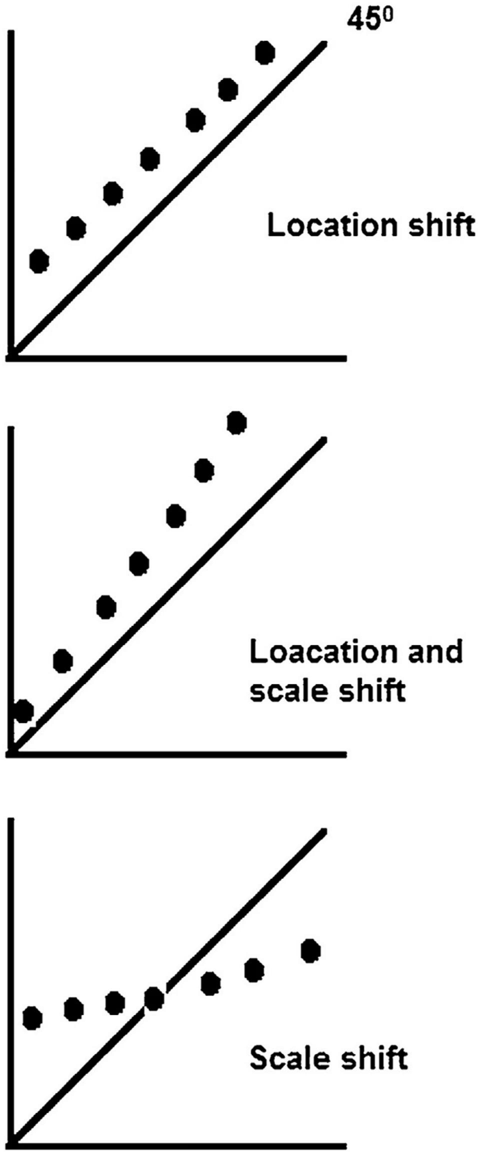

FIGURE 1. Cases when Pearson correlation coefficient fails to detect nonreproducibility.

Siamak Sabour
Department of Clinical Epidemiology, School of Health,
Shahid Beheshti University of Medical Sciences, Tehran, Iran
Safety Promotions and Injury Prevention Research Center,
Shahid Beheshti University of Medical Sciences, Tehran, Iran
E-mail address: s.sabour@sbmu.ac.ir

Journal of Voice, Vol. 33, No. 1, pp. 49–50
0892-1997

https://doi.org/10.1016/j.jvoice.2017.10.007

REFERENCES

1. Yanagida S, Nishizawa N, Hashimoto R, et al. Reliability and validity of speech evaluation in adductor spasmodic dysphonia. *J Voice*. 2017;pii: S0892-1997:30095–30104. doi:10.1016/j.jvoice.2017.06.022. [Epub ahead of print].
2. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45:255–268.
3. Szklo M, Nieto FJ. *Epidemiology Beyond the Basics*. 2nd ed. Manhattan, NY: Jones and Bartlett Publisher; 2007.
4. Sabour S. Adherence to guidelines strongly improves reproducibility of brachial artery flow-mediated dilation. Common mistakes and methodological issue. *Atherosclerosis*. 2016;251:490–491. doi:10.1016/j.atherosclerosis.2016.05.035. Epub 2016 May 20.
5. Sabour S. Reliability of a new modified tear breakup time method: methodological and statistical issues. *Graefes Arch Clin Exp Ophthalmol*. 2016;254:595–596. doi:10.1007/s00417-015-3138-4. Epub 2015 Aug 28.
6. Sabour S. Reproducibility of dynamic Scheimpflug-based pneumotonometer and its correlation with a dynamic bidirectional pneumotonometry device: methodological issues. *Cornea*. 2015;34:e14–e15. doi:10.1097/ICO.0000000000000401.
7. Sabour S. Spinal instability neoplastic scale: methodologic issues to avoid misinterpretation. *AJR Am J Roentgenol*. 2015;204:W493. doi:10.2214/AJR.14.13870.
8. Sabour S. Validity and reliability of the new Canadian nutrition screening tool in the “real-world” hospital setting: methodological issues. *Eur J Clin Nutr*. 2015;69:864. doi:10.1038/ejcn.2015.69. Epub 2015 Apr 29.
9. Sabour S. Reliability of automatic vibratory equipment for ultrasonic strain measurement of the median nerve: common mistake. *Ultrasound Med Biol*. 2015;41:1119–1120. doi:10.1016/j.ultrasmedbio.2014.10.017. Epub 2015 Jan 16.
10. Sabour S. Validity and reliability of the 13C-methionine breath test for the detection of moderate hyperhomocysteinemia in Mexican adults; statistical issues in validity and reliability analysis. *Clin Chem Lab Med*. 2014;52:e295–e296. doi:10.1515/cclm-2014-0453.
11. Sabour S, Dastjerdi EV. Reliability of four different computerized cephalometric analysis programs: a methodological error. *Eur J Orthod*. 2013;35:848. doi:10.1093/ejo/cjs074. Epub 2013 Oct 16.