Letters to the Editor

Clinical measures: reliable or not?

SIR—We were interested to read the paper by Gregson and colleagues [1] in the May edition of Age and Ageing, not just because we have a common interest in improving the measurement of post-stroke impairment but also because their findings appear to contradict our own [2]. Although the findings of this study influence our interpretation of the published reliability data for the Modified Ashworth Scale (MAS), the problem of the measurement of low tone, or flaccidity, remains. Moreover it is interesting that the ‘unweighted’ k values were similar to those we found for the three-category scale, spastic/normal/flaccid, which we tested, and to the values found in other studies of the MAS [3, 4]. Gregson and colleagues [1] employed the weighted k on the justification that “a difference of one point on each of the scales would not be considered clinically significant”. This statement and their findings suggest that, although the MAS may be reliable when a standardized assessment procedure is used, it might still be a rather blunt instrument to measure change in response to an intervention.

The use of the weighted k statistic by Gregson and colleagues [1] is statistically valid, but we are concerned that it may have given an inflated impression of the real clinical reliability of the MAS. In their seminal paper on reliability, Bland and Altman [5] argue cogently that agreement is the central issue. More recently they have pointed out that reliability is not a statistical issue but a clinical decision [6]; statistical methods should be used to inform but not direct those clinical decisions. We agree with the sentiment and are concerned that the weighted k statistic, despite its excellent statistical pedigree, can act contrary to it, converting poor clinical agreement into high statistical agreement.

For example, the two assessors rated the muscle tone of the knee on the MAS differently for 47% of patients, but the weighted k was good (0.73) [1]. If all or most of the disagreements were only by one scale point, clinical reliability probably exists, and this weighted k value reflects that. However, if a reasonable proportion were by two points or more, the weighted k value does not reflect the clinical reality.

Valerie M. Pomeroy, Brian Faragher, Ray C. Tallis
The Stroke Association’s Therapy Research Unit,
Clinical Sciences Building, Hope Hospital,
Eccles Old Road, Salford M6 8HD, UK
Fax: (+44) 161 787 5722
Email: vpmoyer@fsi.ho.manch.ac.uk


SIR—We were interested to see the juxtaposition of our article with that of Pomeroy et al. in Age and Ageing, volume 29, number 3 [1]. We were further interested to read this group’s letter and the editorial by Ward [2]. We were pleased that we had a positive influence on Pomeroy and colleagues’ interpretation of published data. It is clear that they and we have read the same articles published up to March 1999 on measurement of tone. However, we have drawn different conclusions and have gone down a different experimental path, leading to several publications which support the reliability of clinical measures [3, 4].

While we agree that the modified Ashworth scale is of no value in measuring low tone, we do not agree that previous empirical evidence is unequivocal in showing it to be unreliable in measuring normal and increased tone. This is most evident in the work of Bohannon and Smith [5], dismissed by Pomeroy et al. as having used the wrong statistics. The article contains the raw data, and if the appropriate statistics are used (i.e. k [6] or weighted k [7]), we see very good agreement, k = 0.83 and $k_w = 0.98$.

Pomeroy et al. question the use of k and weighted k statistics, quoting an example of our work in which muscle tone at the knee was rated differently by the two raters in 47% of patients but with a k of 0.73, and arguing that these differences may be of two or more points. However, use of k with quadratic weights, whilst giving partial credit for ratings which differ by one point on the scale, gives only minimal credit for ratings which differ by two points or more. Thus, a high k value implicitly reflects that most differences were of only one point. We would be happy to provide our raw data to anyone who may be interested.

Furthermore, we believe that a difference of one point in a given patient would not immediately be considered as a definite and clinically relevant change, but would rather be noted as a possible change and reassessed on another occasion. Thus, we are not considering reliability as merely a statistical issue, but rather we would use it as a guide to our clinical practice.

We are puzzled why Pomeroy et al. did not train their raters in use of the tools or indeed why they suggest training should decrease reliability. It is a tenet of research that standardization of methods will reduce
disagreement [8]. Haas et al. [9] suggest that their poor agreement may be due to lack of training, while Bohannon and Smith [5] had training and showed very good agreement, as indeed did we [3, 4, 10].

Even more puzzling for us was the development and evaluation of two apparently new scales not used in clinical practice. Also, the format of the scales begs some questions. For the three-point scale of “flaccid, normal and spastic”, it is unlikely that good reliability could be demonstrated, since as the number of items of a scale decreases, reliability also decreases, as more of the observed agreement is accounted for by chance [6]. Equally, there is no rationale for use of a visual analogue scale with a potentially infinite number of items, and it is not surprising that agreement was not evident.

Furthermore, Pomeroy et al. used intra-class correlations to analyse this continuous data, rather than the method recommended by Bland and Altman [11]. Appropriateness of the statistics used is further called into question by the use of individual pairwise comparisons between raters, rather than the more appropriate $\kappa$ for multiple raters [12]. Not only is it important to understand the spread of scores by different raters recorded in the same patient at the same time, but also the score by a single rater at different times (intra-rater reliability), which has not been considered by Pomeroy et al. This is critical in measuring agreement, as within-observer disagreement may explain between-observer disagreement [8].

We acknowledge that the modified Ashworth scale does not measure low tone and the development of such a measure would be useful. However, by recognizing and building upon the strengths of previous empirical work, we have been able to show the modified Ashworth scale to be reliable in measuring normal and increased tone in those with stroke. This forms the basis of a rigorous approach to the development of other clinical measures.

In Ward’s editorial [2], we were concerned that the author appeared to think that we had compared the modified Ashworth and Medical Research Council (MRC) scales with each other, and further suggested a more appropriate comparison between the original Ashworth and MRC scales. We did not make any comparison between the scales. Indeed, to try and so would make no sense since they do not measure the same thing—the modified Ashworth scale measures tone and the MRC measures power.

**Letters to the Editor**


JANINE M. GREGSON, MICHAEL J. LEATHLEY, A. PETER MOORE, TUDOR L. SMITH, ANIL K. SHARMA, CAROLINE L. WATKINS

Stroke Team for Audit and Research, Department of Medicine for the Elderly, University Hospital Aintree, Lower Lane, Liverpool L9 7AL, UK

Fax: (+44) 151 529 3239

Email: aksharma@aintreestar.i-net.com