# Experimental Determination of Optimal Scales for Usability Questionnaire Design

**Robert Bierton and Richard Bates**

HCI Research Group, Department of Computing Science
De Montfort University, The Gateway, Leicester, UK

Email: [rbierton, rbates]@dmu.ac.uk

## ABSTRACT
This paper presents experimentally derived recommendations for the optimal number of scale intervals that should be used for subjective usability and workload questionnaires and shows the results of preliminary work investigating the effects of language understanding on fully labelled or partially labelled questionnaire designs.

**KEYWORDS:** questionnaire design, usability testing, workload assessment, NASA-tlx

## INTRODUCTION
When designing a scale there are two main design decisions to be considered: should the scale be fully labelled or partially labelled and how many intervals should be used? A survey of usability and workload questionnaires used in HCI work has shown that there appears to be a large variation in these designs with both fully labelled and partially labelled formats and scale intervals ranging from 5 to 20 points. When choosing the number of intervals there is a trade-off between using too few, resulting in a loss of fine resolution due to "coarseness of grouping", and too many, thus exceeding the rater's ability to discriminate between the intervals, Hancock et al. (1991), Symonds (1924). The choice of using a fully labelled or a partially labelled scale has been investigated, Frisbie et al. (1979), with no clear preference found.

## DETERMINING AN OPTIMAL SCALE
Nine questionnaire scales were produced: 5, 7, 9, 11 interval fully labelled (all points) and 5, 7, 9, 11 interval partially labelled (end and centre anchor points only). A 20 interval partially labelled scale (end anchor points only, no centre anchor point) was also included to test the validity of the NASA-tlx workload questionnaire scale, Hart (1988). The labels used were determined from previous work. A simple Fitts' Law type target acquisition experiment was conducted with 4 levels of difficulty produced by adding jitter to the cursor position. To test the full range of the questionnaire scales, level 1 was set to be easy and level 4 to be difficult. To test the discrimination within the scales, levels 2 and 3 were set to be extremely close in difficulty. Levels 2 and 3 were determined by a pilot study that found the smallest difference in test difficulty that 90% of test subjects could discriminate. A fully randomised design was used with 14 postgraduate student subjects (who all spoke English as their first language) presented with each of the questionnaires and asked to rate the difficulty of the 4 test levels. The results were scaled to 20 points to allow inter-scale comparison and the $25^{th}$ and $75^{th}$ percentiles, giving the inter-quartile ranges, were plotted (figure 1).

All of the questionnaires correctly identified levels 1 and 4. Both of the 5 interval questionnaires exhibited coarseness of grouping and could not discriminate between levels 2 and 3. Both of the 7 interval questionnaires and the fully labelled 9 interval questionnaire gave good results. The 9, 11 and 20 interval partially labelled and the 11 interval fully labelled questionnaires had too many intervals and exceeded the subjects' ability to discriminate between the intervals, resulting in a wide distribution of results.

Overall, the fully labelled 7 interval questionnaire gave the best discrimination and was the only scale that produced a clear distinction between difficulty levels 2 and 3.

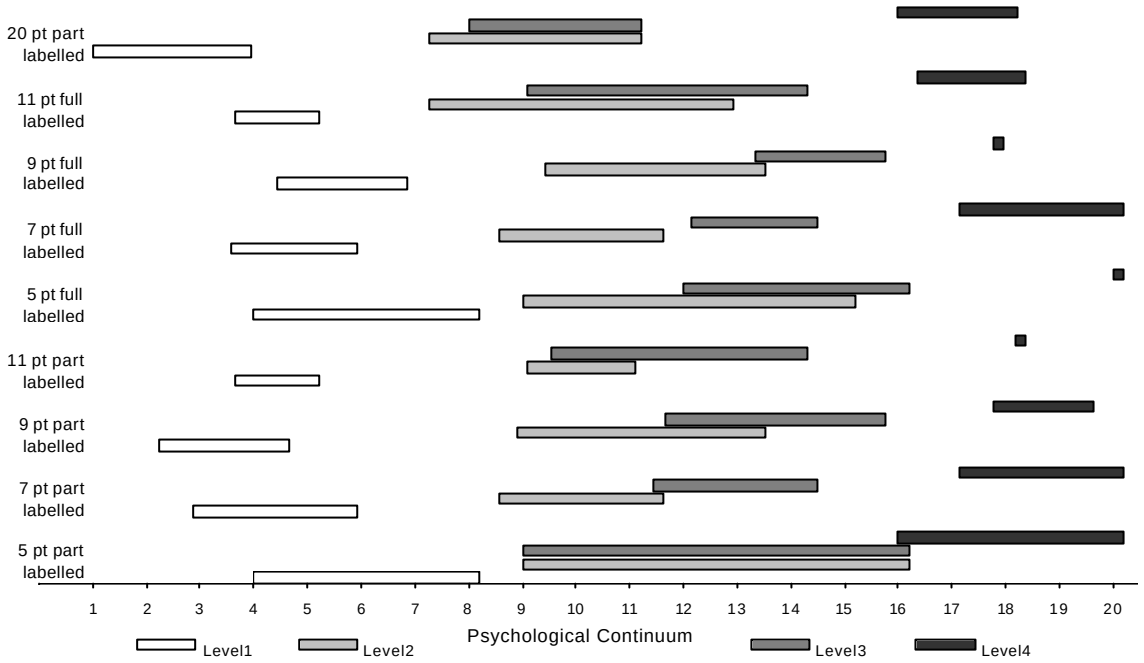


Figure 1. Distributions of Questionnaire Responses

## EFFECTS OF LANGUAGE

A follow-up test was conducted with 4 subjects whose first language was not English but who could fully understand spoken English. The preliminary results showed a marked deterioration in the discrimination of the fully labelled questionnaires when compared to the partially labelled questionnaires such that the partially labelled 7, 9 and 11 interval scales gave better discrimination than the fully labelled questionnaires. In this case the partially labelled 7 interval questionnaire now gave the best discrimination.

## CONCLUSIONS

Provided that the test subjects fully understand the meanings of the labels, 7 interval fully labelled questionnaires give the best discrimination. When this is not the case, 7 interval partially labelled questionnaires should be used. Using too few (less than 7) or too many (greater than 9) intervals results in loss of discrimination. The use of a long 20 interval scale in the NASA-tlx is questionable. Further work will be carried out to increase the number of test subjects and to further investigate the effects of language understanding on questionnaire responses.

## REFERENCES


Frisbie, D. A., Brandenburg, D. C. (1979) Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement* **16**(1), 43-48.

Hancock, G. R., Klockars, A. J. (1991) The effect of scale manipulations on validity: targeting frequency rating scales for anticipated performance levels. *Applied Ergonomics* **22**(3), 147-154.

Hart, S. G., Staveland, L. E. (1988) Development of the NASA-tlx (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock, N. Meshkati, Eds., *Human Mental Workload*. Elsevier, 139-183.

Symonds, P. M. (1924) On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology* **l7**, 456-461.