University of York Department of Health Sciences

M.Sc. in Evidence Based Practice

Measurement in Health and Disease

Assessment, June 2006

Specimen Answers

Question 1.

(a) *What is a kappa statistic? How could we interpret k = 0.90 and k = 0.58?* Kappa is a measure of agreement between two assessments or measurements using a categorical variable. It uses the proportion of cases for which there is agreement and the proportion we would expect to agree if there were the agreement we would expect by chance in the absence of any relationship between the two assessments. We take the amount by which the observed agreement exceeds chance, agreement minus expected agreement, divided by the maximum value this could have, one minus expected agreement. The maximum value is 1.00, for perfect agreement. A kappa = 0.00 means no more agreement than would be expected by chance. 0.90 is usually taken as indicating very good agreement, as is any kappa value above 0.80. 0.58 is taken to mean moderate agreement, as is any kappa between 0.40 and 0.60.

(b) *What is a weighted kappa statistic and how does it differ from kappa? Why are the weighted kappa statistics all larger than the corresponding kappa statistics and what does this tell us?* The ordinary kappa statistic does not take into account any ordering of the categories. All disagreements are treated the same. If we have ordered categories, such as 'poor', 'fair', 'good', 'excellent', we might want to regard a disagreement between two observations where one is 'fair' and the other 'good' as less important than a disagreement where one observations is 'poor' and the other 'excellent'. Weighted kappa does this, by attaching arbitrary weights to pairs of categories. The weights should be chosen in advance to reflect the meaning of the categories. Weighted kappa can only be used when there are more than two categories. It is usually the case that most disagreements are for pairs of categories which have been given a low weight for disagreement and pairs of categories given a high weight occur only rarely. This increases kappa from the unweighted version which treats all disagreements as of equal value.

(c) *Percentage agreement is given for each of the measures used. Why might this be a misleading statistic and why is it always greater than the corresponding kappa?* Percentage agreement can be misleading because when one category is chosen much more often than others, the percentage agreement will always be high. This is because just by chance many subjects will get the frequent category from both observers. For example, we might have the following

	Observer A	
Observer B	Yes	No
Yes	1	9
No	9	81

Here whether B says 'yes' is unrelated to whether B says 'yes', but the agreement is 82%. The corresponding kappa is proportion agreeing minus expected proportion divided by one minus expected proportion. It is easy to show that that this must be less than observed agreement unless this is equal to 1.00.

Question 2.

(a) *What is an intraclass correlation coefficient? Why is this described as measuring 'relative' reliability?* An intraclass correlation arises when we have several observations on each of a group of subjects and which we regard as equivalent. It is an average of all the correlations which we could produce by selecting two measurements from each subject and finding the correlation between them. It is the ratio of the variance of the true (i.e. average over all possible observers) quantity being measured to the variance of the measured value. It measures relative reliability because it is depends on the variation of the true value of the quantity between the subjects. If we choose a sample which has little variation we will get a lower value of the ICC than if we select a sample with high variation. ICC is a pure number with no units and is a measure of how good the measurement is at distinguishing between individuals.

(b) *What are limits of agreement and how could they be used to investigate test-retest reliability? Why is this described as measuring absolute reliability?* Limits of agreement are a pair of numbers within which we estimate 95% of differences between pairs of observations will lie. They are calculated from the mean and standard deviation of these differences, which here are between the first and second measurement. The limits are estimated from the mean difference minus 1.96 (or 2) standard deviations to mean plus 1.96 standard deviations. The limits are in the same units as the observations, so they tell us the range of values which a second measurement could be in if the true value for the subject does not change. Thus they enable us to interpret an individual measurement. They are not affected by the variation of the true quantity in the sample, which is why they are absolute rather than relative.

(c) *The authors report that "the 95% limits of agreement of Bland-Altman plots contained most data points for both methods". Why doesn't this tell us anything useful?* We estimate these limits to contain 95% of pairs of observations. Hence we expect 95% of points to be between them, particularly when these observations are the data used to estimate the limits. We have what we would expect to get given the way the limits were calculated, so we learn nothing useful.

Question 3.

(a) *What does the authors mean by 'Exploratory . . . factor analysis of the CMRS-P indicated that the scale was unidimensional.'?* Factor analysis ask whether a set of variables are explained by a smaller number of underlying factors. We judge how many factors we need from the eigenvalues which the factor analysis calculations produce. This may be done by looking how many eigenvalues exceed 1, or by a scree plot which uses a graphical method to find how many factors we need. One of these methods has been used here and it has been concluded that one factor is needed to explain all the variables, so the resulting scale is unidimensional.

(b) *The internal consistency (alpha) and test-retest reliability were both 0.96. What do these terms mean and how would we interpret 0.96?* The internal consistency of a scale can be measured by Cronbach's alpha. This looks at how closely the variables are correlated and provides and estimate of how closely the measured scale will be correlated with the underlying quantity being measured. It does this by treating the observed variables as a representative sample of the variables which could be used. The test-retest reliability is the correlation coefficient between the scale measured on two different occasions, Usually an intraclass correlation coefficient is used. Alpha = 0.96 represents a very high degree of internal consistency between the items in the scale and shows that this makes a coherent scale which measures something. Test-retest reliability = 0.96 show that the variation from occasion to occasion is small compared to the variation of the quantity being measured. It also provides good evidence that the scale discriminate well between different individuals and measures something.

(c) *What is 'construct validity' and how could the analysis described demonstrate it?* Construct validity means that the measurement behaves in the way w would expect a measurement of this construct to behave, having the relationships and lack of relationship to other variables which we would expect. Here construct validity is to tested by examining the relationships between the mania scale and a range of existing psychological measures. We are looking for strong relationships with constructs where mania might be observed and weaker relationships where we would be less likely to see mania.