Analysing questionnaires and ratings data

There are a couple of traps in the analysis of ratings data which do not seem to be commonly understood.

1. The first is what the quantisation of the responses implies about the accuracy of the results. To illustrate this let us assume that we are using a ten point integer scale from 1–10. Say someone wanted to respond with a 5.5, then because of the quantisation we are forcing them to use, they could either answer it as a 6 or a 5. The error induced by the quantisation implicit in the scale we are using, means that a 6 could mean anything between 5.5 and 6.5. Say a second person responded with a 3, it means that their true answer could be anywhere between 2.5 and 3.5. If we then take an average of the two responses, the value could be anywhere between (2.5 + 5.5)/2 and (3.5 + 6.5)/2, that is somewhere between 4–5. Simply averaging the responses will give us 4.5, but our analysis indicates that the true answer lies somewhere in the range 4–5, or to put it another way the answer is 4.5 ±0.5. The interesting thing about this is no matter how many responses you get, the error of the average is the same as the size of the quantisation you have used. On a ten point scale this is ±5%, while on a five point scale it is ±10%. There is some evidence that people are able to differentiate answers using a twenty point scale (Guilford, 1954), in which the error would be ±2.5%. Obviously this demonstrates that the more points you have in the scale the smaller the quantisation error in your answers. This analysis of quantisation indicates that someone using a five point scale and attempting to suggest that differences which are less than 10% apart are significant needs to be taken with considerable scepticism. Of course we could always apply the central limit theorem to our answers and assume that the error is normally distributed, but then we need to estimate the standard deviation of the error and apply a confidence limit to the error – but I have not seen this type of analysis done in practice.
2. The second is that the best you can get for the responses of people to ratings questionnaires is to assume that the scale they are using is somewhere between the ordinal and interval scales as defined by Stevens (1951). That is they are likely to be monotonic (i.e. a higher number means a higher level of response) but are not likely to have equal intervals between the numbers, nor be based on the same absolute zero. Also different people will have different intervals in their scales which are likely to be based on different absolute zeros. This is because people use scales that are reliant on their different experiences, conceptions and emotional states. In this case analysing the resulting data as if they were numbers on a perfect ratio scale is likely to lead to very dubious conclusions.

If we compare this with how we measure, say, the distance of a metre, we find that people do not share a universal measurement standard for the each of the items under consideration. The acceptance of a standard is a vital precondition before any set of measurements can be accurately compared. In the metric system of measuring length, a platinum rod in Paris used to be the primary standard metre (it now is defined as a precise number of wavelengths of a particular radiation). All measurements rely on this standard to define the exact length of a metre and because of this length measurements are based on a perfect ratio scale.

In attempting to measure psychological characteristics in a quantitative way, there is no convenient standard that can be defined. Ideally we would like to calibrate a person's measurement scales prior to getting them to respond to a ratings questionnaire. This could be done by getting them to assign numbers to a set of tightly defined items to determine the parameters of the scale they use. In practice this would not be feasible because it would be impossible to verify that the scale was likely to be applied consistently to all the items in the questionnaire due to their different experiences and emotional states. Indeed it is likely that they could use different scales for different items, and even the attempt at calibrating their scales may well affect their emotional state and hence their subsequent rating scale.

A second alternative method would be to actually define what the steps in the scale are, by each step having a description of the item. For example, an item on employee empowerment might have a number of descriptions defining a scale from totally disempowered to completely empowered. The respondee would simply select the description which they felt was the closest to their view of the item. This way we have defined the scale and all we are assuming is that all the respondees will have similar interpretations of the descriptions.

A third alternative is to try and transform the responses by using analytical techniques that make the analysis of the responses more robust. An approach we have often used, is to apply a standardisation transformation to the responses of each individual based on the set of responses they have provided to all items. The assumption here is that the respondee has used a consistent, though unknown, scale across all items being measured. That is their answer about empowerment used a scale which is consistent with their answer about their levels of trust in the organisation. By making this assumption we can estimate what the mean and standard deviation of their 'scale' is and apply a standardising transform to all of their ratings based on these statistics.

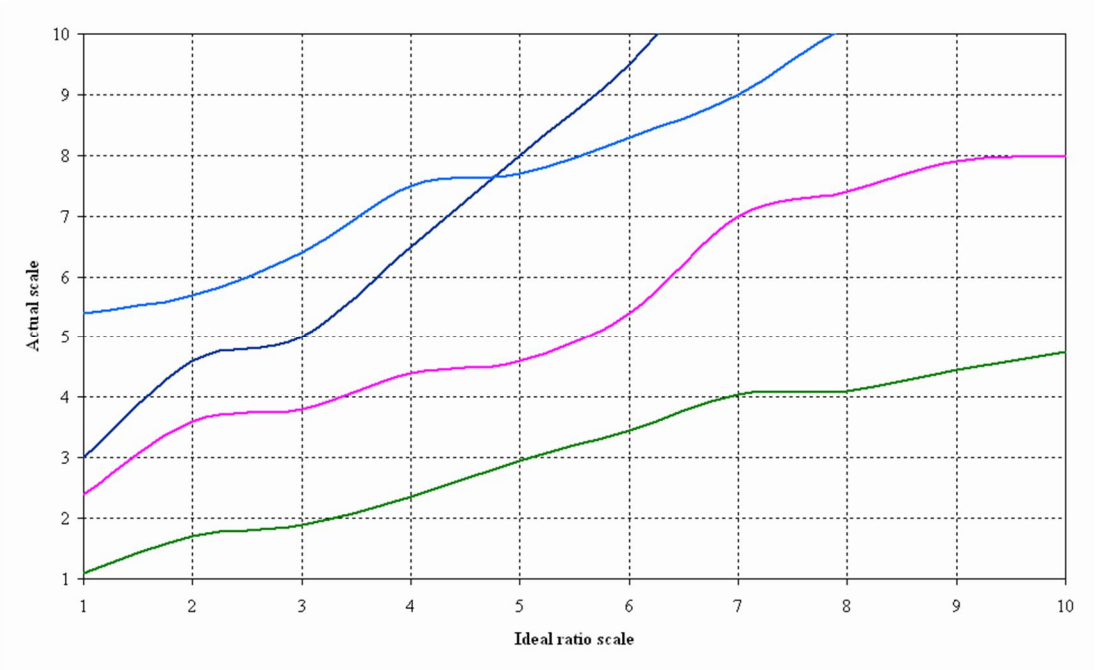
We can illustrate the effect of performing this standardisation by plotting lines representing the actual scales used by respondents against the ideal ratio scale. Figure 1 illustrates four such possible scales. Note that these examples are all monotonic, but they are curved to represent non-interval scales (i.e. they are non-linear scales), do not pass through a common point (i.e. do not share a common zero) and have different levels of steepness to represent the differences between intervals used. Blindly attempting to average responses across these scales and then make comparisons based on the differences between the averages is likely to produce fairly meaningless results.

Standardisation of data

If we assume that people share the same perception of the phenomenon we are trying to measure, but simply use different scales in assigning a value to the phenomenon, then we can correct for the range of the scale they use by standardising their results to a zero mean and unit standard deviation. We can use the transformation:

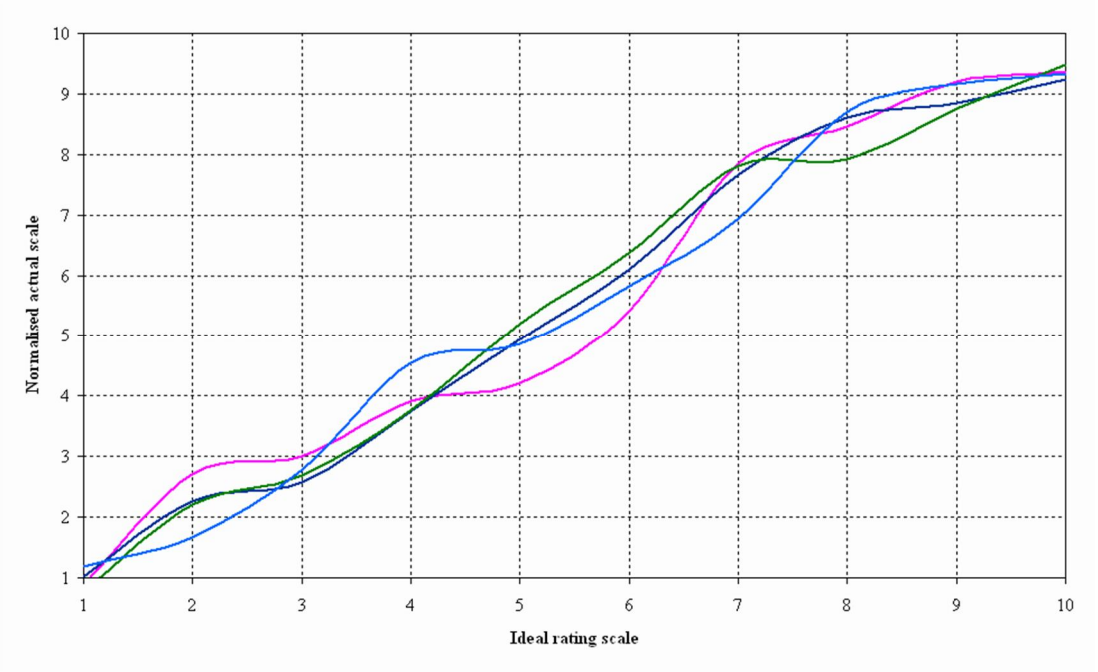
$z_{yi} = \frac{x_{yi} - \bar{x}_y}{\sigma_y}$, where $\bar{x}_y, \sigma_y$ are the mean and standard deviation of all of the responses of person **y** respectively and x_{yi}, z_{yi} are the original response and the standardised response to the ith item for person **y**.

Figure 1: Plot of possible ratings scales used by respondents



If we apply this transformation to the scales in Figure 1 and, then for comparison, adjust the z-values to have a mean of 5 and standard deviation of 3.03 (i.e. transforming them back to the initial ten point ratio scale) leads to the scales in Figure 2.

Figure 2: Standardised scales of respondents



Note that now the standardised values are a lot closer together as you look up vertically from the ideal scale axis. This indicates that it is now a lot more meaningful to average responses than was the case in Figure 1. Also the standardised scales span the full range (1-10) and they are also more closely based on the same zero (the point (1,1)) and therefore are can be more correctly treated as if they were a ratio scale. Note also that if, by chance, they were originally linear, they would now be a set of perfect ratio scales with equal intervals (Thurstone).

A commonly expressed objection to this approach is that you lose the absolute value of responses. For example, if I give a 9 as a response it now no longer becomes a 9. But this is exactly what is required, because your 9 is not the same as anybody else's 9 and treating it as if it were will introduce an important systemic error in the measurement process. Your 9 will still be one of your most positive responses under the standardisation transformation (depending on your other responses) and will still contribute to raising the overall average of responses to that item. Similarly, if all respondents used the highest rating of their scale on this question, it would end up having the most positive average result of all the items.

Conclusions

It is not a sound analytical approach to treat raw responses that lie somewhere between ordinal and interval measures as if they were ratio scales, as is commonly done. Possibly the best way to design items in a questionnaire is to use a set of defined points in the scale by having descriptions for each response level. This approach avoids the assumptions which would be otherwise necessary in order to analyse the responses from the people surveyed. It also allows comparisons between the results from different organisations to be made, as the scale being used is identical.

Another common issue is that there is often no realisation of the impact the choice of scale has on the errors due to quantisation of the responses. It can easily be shown that the commonly used five point scale has an error range of ±10% when averaged across any number of responses. This implies that trying to suggest that small differences between items are valid is quite misleading. If the scale is quantised by assigning the numbers 1–5 to the scale, then the best one can do, without making an assumption about the applicability of the central limit theorem, is to round the average of the responses to each item to the nearest integer.

Another technique is to apply a standardisation transform to the ratings scores which helps overcome some of the differences between scales used by people in their responses. The standardisation technique still requires a major assumption to be made about how consistently a person uses a common scale when rating different items. However if that assumption is a valid one, or approximately so, the standardisation technique converts peoples' response scales into a far closer approximation of a set of ratio scales. Once this is done then standard analytical techniques can be applied with far more certainty of producing useful results.