FORUM

More scales than a fish?

Michael Cramphorn
Swinburne University

Verbal scales are intrinsic to attitude measurement. One approach is to employ agreement–disagreement scales. However, when inter-country comparisons will be made, it cannot be presumed that results will be directly comparable. Different verbal usages often prevail, even where the language is the same. Variations in the patterns of response across cultures and languages are non-systemic, the consequence of which is that single overall country adjustments should not be used. Focusing on meaningful measurement can yield scales that provide real comparability.

Introduction

Measuring people’s feelings can be simply like or dislike – a binary scale; it can be more sensitive, seeking to measure the degree to which something was liked, using some sort of multi-point scaling approach, but, implicitly or explicitly, scales are an inevitable part of the measurement process.

However, using scales is far from straightforward. For example, should a scale be balanced, how many intervals should be included, are the ‘distances’ between them equivalent; does the questionnaire have sufficient interest to sustain quality input (Sudman & Bradburn 1982; Albaum & Murphy 1988; Garland 1991; Albaum 1997; Lee et al. 2002; Hansen & Smith 2012)? But these questions relate to developing a scale for a homogeneous population. What happens when a successfully developed scale is used in a new environment?

It cannot be assumed that, because people have attitudes, a phrase such as ‘I dislike it a little’, used in English, has the exactly equivalent meaning in German, or Urdu, or Xhosa. As will be shown, it is not even true in...
different countries using the same language. Fortunately, there is increasing awareness of the potential pitfalls of not adapting research methods to deal with other environments (Mouncey 2012).

**Using scales**

In verifying a scale, an outcome where there is a mild reaction to non-contentious issues is helpful as it is likely to generate frequencies for each attitudinal option that approximates a normal curve, with most responses falling around the centre and a diminishing number out at the extremes. This can be verified further by qualitatively checking that the rating option chosen accurately reflects actual feelings.

But response lop-sidedness is not uncommon. When such reasonable ‘normality’ does not apply, it is not the fault of the respondents – either the stimulus material is extreme, or else the response options presented do not enable the respondents to sensitively reflect how they feel. For example, often in evaluating confectionery, the *Not good* options are seemingly redundant, and extra positives such as *Extremely good* and *Superlative* need to be included to introduce enough sensitivity to obtain degrees of feeling and a non-lopsided (but positive) pattern of overall response. Conversely in some countries, where there is an aversion to eating food that sticks to one’s teeth, the negative end of the confectionery scale may need to be extended to allow for descriptors such as *Horrible* and *Disgusting*. Arguably, all these options, both positive and negative, should be there all the time. However this can make a scale very cumbersome. Also, the ‘normality’ assumption can break down entirely in places where people are culturally unwilling to say anything negative (the so-called ‘courtesy bias’) so that, in practice, only the positive half of the scale gets used. In such cases, the measurement is not interpretable.

Yet in everyday speech, the same people can verbally describe the intensity of their positive and negative feelings. So the problem is to reflect the level of intensity of these feelings using a scaling process. Sometimes this can be catered for by adding new descriptors as described above. But others require the researcher to devise new ways to obtain measurements that relate proportionately to these feelings.

**Dealing with systemic unbalanced patterns**

International experience over the past two decades has provided instances of this sort of response pattern. In the 1990s in India, it proved almost
impossible to get anything other than positive responses on a balanced
scale, even though people would be verbally critical of the material
to which they were assigning a positive rating. After several months,
experimenting with numerous approaches and then comparing what was
said with what was rated, this challenge was dealt with by administering
a numeric scale out of 100 and developing an algorithm where 100
meant *strong agreement*, 98–99 meant *agreement*, over 95 meant *partial
agreement*, and so on. But this example is not a long-term answer because
it presumes that all respondents are at the same level of sophistication in
their response patterns, and also that the population will remain fixed at
this level in the future. Another approach, less susceptible to chronological
constraints, was employed to deal with ‘courtesy bias’ in the Philippines.
The ‘scale’ used different-sized visual squares to denote the level of
agreement with statements. Again a qualitative confirmation of the scaling
process was required to validate its use. It is also possible to ask: ‘Do you
agree or disagree when I say …? Now is that a very big “yes”, a tiny “yes”
(or “no”), or somewhere in between?’ This sequence of questions yields a
de-facto verbal six-point scale.

The intent should be to try to gain an overall response pattern that, for a
typical range of stimuli, approximates a normal curve, which demonstrates
that the entire scale is capable of being used meaningfully by the population
of respondents, so that extreme outcomes represent real extremes of
feeling, not just polarised dichotomies. This seems, to the author, to be the
core of any practical approach to developing scales to measure degrees of
feelings. In some instances, pragmatically reducing the scale range may be
appropriate for a homogeneous population (Bendixen & Yurova 2012). But
where there is potentially a wide range of degrees of feeling across many
ideas, such a pre-emptive approach at the data-capture stage would appear
to limit capacity to obtain richer data, prior to having an opportunity for
analysis. However, it could be justified if it were already established that, in
all cultures of interest, this was indeed how the scale was used.

**Data**

In all cases, the stimulus material is advertising. In each case, the
respondent has been exposed to the advertising and then been asked to
complete a questionnaire that rates both the ad and the brand on many
aspects. The data presented in this paper are drawn from an archive of
ratings of questions, used virtually unchanged for more than 20 years in
more than 50 countries. In order to arrive at the statements that reflect
these aspects, an initial study was undertaken in the early 1990s using more than 50 ads and more than 80 candidate statements in order to establish which statements provided an insight into the way in which people respond to advertising. Following factor analysis, two major factors emerged that correspond to the feeling towards the brand, and the level of interest elicited in the ad itself. Since then, as more and more data became available, and as more ideas for measurement emerged through discussion and exposure to new cultures, more candidate statements were tested for possible inclusion. The set of approximately 65 statements stabilised in the mid-1990s and, after that, new candidate statements, when tested, did not add new insights. The statements are couched in both positive and negative formats, and these are interwoven to break up response patterns. The statements about the ad and the brand are also mixed up and randomised in order to minimise recall of previous ratings and reduce order effect. Quality control includes internal consistency checks on each respondent, and the requirement that all key questions be answered. The archive is a by-product of a regular commercial process, and was created specifically to retain detailed information that would enable the exploration of areas of interest that might arise in the future.

For this paper, in order to gain an overall pattern of scale use within each country, all the results of all the ads tested for all of the positive statements have been combined to obtain the pattern of usage of each of the six scale options. Thus, in the case of New Zealand, 6,550 people have rated one of 133 ads using 43 positive statements and 22 negative statements relating to either the ad or the brand. This gives a bundled total of 281,650 individual ratings of positive statements and 144,100 individual ratings of negative statements. It follows that any effects attributable to individual statements are negligible, and that the distribution of responses across the six rating options is stable.

The ratings for the countries discussed are summarised in Table 1.

The ads tested were intended for use in everyday advertising, and so represent typical advertising stimuli for each country. Each ad, and the subject brand, was evaluated on multiple criteria using a six-point agreement scale. That is, the criteria discussed earlier are met:

- to have a large enough pool of varied stimuli to enable the whole scale to be used meaningfully
- to have a sufficient number of people undertaking the ratings to be confident that the entire scale is being used appropriately in response to a wide spectrum of possible stimuli.
Yet cultural response tendencies may still be evident. This is demonstrated when reviewing ratings in different cultures where there is no translation. For example, consider the results from the UK, the US, New Zealand and Australia. The language is common; all these countries have similar living standards and lifestyles, and use the same sorts of services and consumer goods, often the same brands and even sometimes the same ads. So when the response tendencies of respondents in these countries are compared, one would not expect to detect any significant differences in how they use the scale.

The levels-of-agreement response patterns for the four English-speaking countries, averaged across 43 positive statements, are shown in Figure 1. Similar charts for negative statements were prepared, but have been omitted in the interests of brevity. It will be apparent that, with the number of responses involved, that the statistical significance of these results will be in the order of 99.9% or more, i.e. close to absolute certainty.

Figure 1 demonstrates that all of the scale is being used with greatest usage of the central points. Thus, although skewed positively, the requirement, across a wide range of stimuli, for the scale to be used most often near the centre (normal), is met. Perhaps the reason that people are generally somewhat more likely to agree than to disagree with positive statements is because ads that are tested are likely to have had the real ‘duds’ pruned out earlier. So, while not quite normal, this probably reflects the stimulus material and that, importantly, the whole scale is regularly being used to convey degrees of feeling. As discussed, if the responses were heavily

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of raters</th>
<th>Number of stimuli (ads)</th>
<th>'Bundled' 43 positive ratings</th>
<th>'Bundled' 22 negative ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>92,076</td>
<td>1,440</td>
<td>3,959,268</td>
<td>2,025,672</td>
</tr>
<tr>
<td>Denmark</td>
<td>2,256</td>
<td>33</td>
<td>97,008</td>
<td>49,632</td>
</tr>
<tr>
<td>France</td>
<td>11,925</td>
<td>126</td>
<td>512,775</td>
<td>262,350</td>
</tr>
<tr>
<td>Greece</td>
<td>12,065</td>
<td>186</td>
<td>518,795</td>
<td>265,430</td>
</tr>
<tr>
<td>Italy</td>
<td>10,511</td>
<td>134</td>
<td>451,973</td>
<td>231,242</td>
</tr>
<tr>
<td>Netherlands</td>
<td>43,015</td>
<td>384</td>
<td>1,849,645</td>
<td>946,330</td>
</tr>
<tr>
<td>New Zealand</td>
<td>6,550</td>
<td>133</td>
<td>281,650</td>
<td>144,100</td>
</tr>
<tr>
<td>Norway</td>
<td>2,053</td>
<td>36</td>
<td>88,279</td>
<td>45,166</td>
</tr>
<tr>
<td>Spain</td>
<td>36,309</td>
<td>367</td>
<td>1,561,287</td>
<td>798,798</td>
</tr>
<tr>
<td>Sweden</td>
<td>4,261</td>
<td>74</td>
<td>183,223</td>
<td>93,742</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>18,701</td>
<td>234</td>
<td>804,143</td>
<td>411,422</td>
</tr>
<tr>
<td>United States</td>
<td>27,747</td>
<td>279</td>
<td>1,193,121</td>
<td>610,434</td>
</tr>
</tbody>
</table>

Table 1  Number of raters and of stimuli
weighted to one end of the scale, it would imply either a faulty scale with inadequate options, or ads being tested that were consistently abnormal to the regular advertising environment. There is some slight degree of greater negativity (in fact, closer to a normal distribution) from New Zealand. But, clearly, the patterns in the other three countries are consistent, and one could potentially conclude that, with minor adjustments, results in these countries are directly comparable.

However, when results in some European countries, with similar cultures and economies, are examined (Figure 2), the correspondence of response patterns is far smaller.

Although the full scale is used in each country, there are quite large differences between them. Over half the people in France, Italy, Greece and Spain (all abutting the Mediterranean Sea) will more typically agree with a positive statement. But, as one goes further north in Europe (Holland, Denmark, Sweden and Norway), over half the people will more typically disagree with a positive statement. Indeed, a case could be made to have an extra ‘Very strongly disagree’ option in those countries. These differences could be explained if all the ads presented in the south were more inclined to gain positives, and those presented in the north more likely to gain the negatives. But, with so many different and varied stimuli, the probability of the ads tested in northern countries all being more negative and the ads exposed to southerners being more positive, as a random event, is extremely small. As noted, such differences cannot be caused by translation of statements because these are the averages of response to all 43 positive statements taken together, and any individual translation effect would be minute. This implies that the differences are cultural.
It appears that there is a systemic range of response distributions and that respondents in those countries most inclined to be negative are, like New Zealand, located nearer the earth’s poles. Interestingly, there is supporting literature demonstrating the positive effect of sunlight on the emotions (Armstrong et al. n.d.; Benedetti et al. 2001; Sumaya et al. 2001; Lambert et al. 2002; Kent et al. 2009), but while it would be interesting to explore the effect of seasonal variation within countries, related to the position of the sun, the purpose of this paper is to discuss how to measure ads meaningfully, not to understand why these differences arise.

What Figures 1 and 2 of overall average response patterns appear to demonstrate is that, irrespective of any translation or stimuli effects, there is a systemic difference in the distribution of responses in apparently similar countries. Thus, the means of the responses will be affected, sometimes to a quite large degree, by the cultural systemic use of the scale, and, without making some adjustment, the results cannot be directly comparable. (The means are not shown in this paper because it is the underlying distribution that is of interest, not the mean itself.)

So if one could assume that the variation in distribution were consistent across all statements, then a simple way to resolve this would be to identify the differences in the means, and simply add or subtract appropriate adjustment amounts to bring the means in line. But such an approach would presume uniformity for each statement in the variance from country to country. However, if one statement does not conform, then it follows that others may not do so, and hence each statement in each country potentially has its own pattern of agreement and disagreement.

**Figure 2 Mean agreement tendency**
Individual statement variations

In the ‘English common language’ countries where there is no translation, the distributions to the statement ‘I enjoyed watching the ad’ are more likely to be disagreed with (a positive outcome) in New Zealand, the US and Australia, while UK viewers are quite a lot less likely to feel this way (Figure 3). New Zealand raters are the most positive in using this statement, which is in direct contrast to the bundled average shown in Figure 1 where, overall, they were most negative. Inspection of other statements confirms that they do not conform in a systemic way, but are everywhere idiosyncratic. This is probably due both to nuance variances in meaning and to the differing usage of English in the different countries.

This indicates that, even in countries using a common language, a single simple overall ‘country adjustment’ cannot be used since it will overcompensate some statements, while exacerbating others.

Figure 3 ‘I enjoyed watching the ad’ – English-speaking countries

This inconsistency of individual statement variation is confirmed by inspection of Figure 4; the country distributions of responses to this statement in other (European) countries are also quite variable, both to one another and as compared to the corresponding country bundled norms.

Indeed, although (for brevity) not illustrated in this paper, evaluation of the data for the other 42 positive statements and a further 22 negative statements for which there is similar data confirms that this is the case. The reader may find it somewhat disconcerting to find that the cultural comparability of rating results is so fraught with difficulty.
Managing the situation

In effect, this means that every statement in every country has its own idiosyncratic response pattern and thus, for comparative purposes, it is necessary to identify these in order to draw meaningful conclusions. Bundling them all together is a useful way of demonstrating differences in the overall response tendency between countries, but the bundled averages cannot be applied as a blanket rule. However, provided that a grass-roots, imaginative approach, such as those described earlier in the paper, is adopted, then much can be done to overcome this.

As a first step, translation should be carefully checked using back-translation, and even employing descriptors of what the acceptance or non-acceptance of the statement means. Following this, a series of tests should be undertaken using a range of stimuli to generate responses likely to utilise the full extent of the scale. The results from all these individual statement responses should be examined and, where possible, correlated with the verbal feelings of the respondents. At that stage, it becomes possible to determine how to calibrate each statement in that country to allow for the nuances of meaning of the statement in that country. The author’s experience with this process in more than 50 countries over the past two decades demonstrates that, once the calibrations are complete, they are robust. However, it remains advisable to review the results from time to time as words sometimes change their meaning and recalibration of individual statements may be required.

One may wonder whether it would be possible to create an industry-wide pooling of information in order to establish a ‘standardised’ set of
comparable scale wordings across countries, which researchers could access. It would be of obvious value to those preparing surveys and in looking for guidance in how to interpret results in different countries. Perhaps this could be an initiative undertaken under the auspices of an international industry body. It would require companies with such multi-country information to make this freely available, something they may perceive as being counter to their commercial interests. But, even if it were done, it may be of limited value, as it would require the precise same wording of the questions/statements and use of the same scale options. Thus a scale adjustment on a six-point scale may not be exactly applicable for a seven-point scale, even if the question used is identical.

**Conclusion**

There are at least two quite distinct factors affecting the way people respond to the scales employed in pretesting ads. One is an overall systemic cultural effect; the other is simply an individual statement nuance arising from the near impossibility of gaining precise language equivalence in translation between different languages.

The difference between online and offline surveys can introduce a further, sometimes significant, effect. So with both online and offline, 65 statements and more than 50 countries where testing has been undertaken, in excess of $2 \times 50 \times 65$ (6,500-plus separate scales) were required.

Provided that care is taken to ensure that full usage of the whole scale will occur as appropriate, then, as described above, it is possible to develop values to calibrate each individual statement on a country-by-country basis.

Without proper calibration of responses, there is no meaningful way to directly compare the responses obtained in different countries from the same questions. Of course, not all tests explore the advertising response in such detail and the task is simpler where there are fewer statements to consider.

Perhaps, in those cases, a fish will still have a few more scales.

**References**


**About the author**

Prior to retiring in order to devote more time to investigating the advertising response process, Michael (Spike) Cramphorn spent over 40 years both as a practitioner and an academic, studying the different ways and conditions in which people respond to advertising. The desire to apply his findings led him to establish his own advertising research business, using a unique methodology, in 1991. Adjunct Professor of Marketing at Swinburne University, Melbourne, Australia, he has published numerous papers in journals and magazines such as *International Journal of Market Research, Journal of Advertising, Admap* and others, and has presented his ideas at conferences in the US, the UK and Europe, Australia and Asia. He will continue to research the database of over 550,000 responses, amassed over the past 20 years and he looks forward in future papers to hopefully providing further insights about how human beings respond to advertising.

Address correspondence to Michael Cramphorn at cekips@gmail.com