1-1-2008

Individualised rating-scale procedure: a means of reducing response style contamination in survey data?

Elisa Chami-Castaldi
University of Bradford

Nina Reynolds
University of Wollongong, ninar@uow.edu.au

James Wallace
University of Bradford

Recommended Citation

Chami-Castaldi, Elisa; Reynolds, Nina; and Wallace, James, "Individualised rating-scale procedure: a means of reducing response style contamination in survey data?" (2008). *Faculty of Business - Papers (Archive)*. 535.
https://ro.uow.edu.au/buspapers/535

Individualised rating-scale procedure: a means of reducing response style contamination in survey data?

Abstract

Response style bias has been shown to seriously contaminate the substantive results drawn from survey data; particularly those conducted using cross-cultural samples. As a consequence. identification of response formats that suffer least from responst style bias has been called for. Previous studies show that respondents' personal characteristics, such as age, education level and culture, are connected with response style manifestation. Differences in the way respondents interpret and utilise researcher-defined fixed rating-scales (e.g. Likert formats), poses a problem for survey researchers. Techniques that are currently used to remove response bias from survey data are inadequate as they cannot accurately determine the level of contamination present and frequently blur true score variance. Inappropriate rating-scales can impact on the level of response style bias manifested, insofar as they may not represent respondents' cognitions. Rating-scale lengths that are too long present respondents with some response categories that are not 'meaningful', whereas rating-scales that are too short force respondents into compressing their cognitive rating-scales into the number of response categories provided (this can cause ERS contamination - extreme responding). We are therefore not able to guard against two respondents, who share the same cognitive position on a continuum, reporting their stance using different numbers on the rating-scale provided. This is especially problematic where a standard fixed rating-scale is used in cross-cultural surveys. This paper details the development of the Individualised Rating-Scale Procedure (IRSP), a means of extracting a respondent's 'ideal' rating-scale length, and as such 'designing out' response bias, for use as the measurement instrument in a survey. Whilst the fundamental ideas for self-anchoring rating-scales have been posited in the literature, the IRSP was developed using a series of qualitative interviews with participants. Finally, we discuss how the IRSP's reliability and validity can be quantitatively assessed and compared to typical fixed researcher-defined rating-scales, such as the Likert format.

Keywords

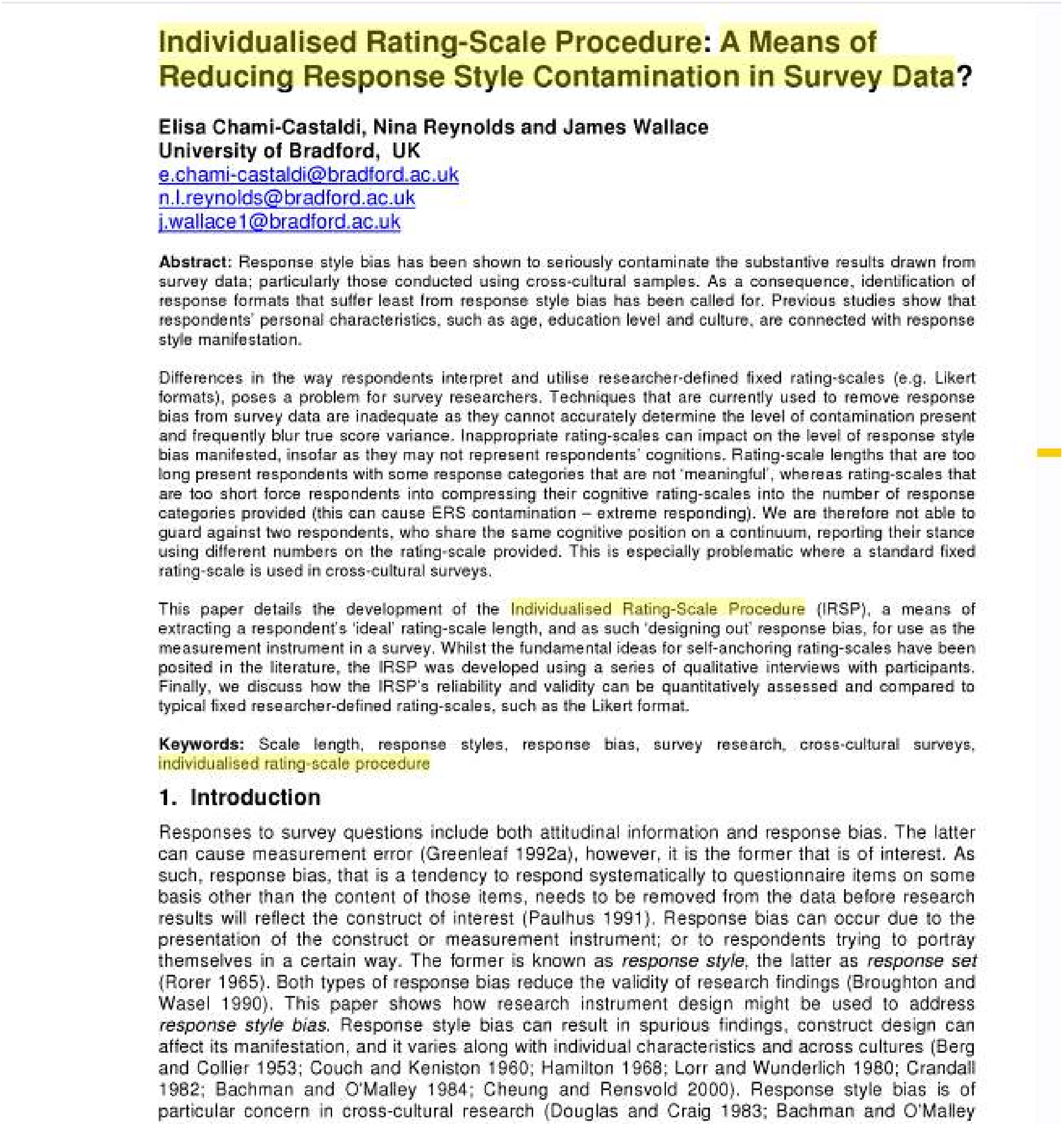
means, procedure, scale, rating, individualised, contamination, data, style, survey, response, reducing

Disciplines

Business

Publication Details

Chami-Castaldi, E., Reynolds, N. & Wallace, J. (2008). Individualised rating-scale procedure: a means of reducing response style contamination in survey data?. 7th European Conference on Research Methodology for Business and Management Studies (pp. 99-109). Reading, United Kingdom: Academic Publishing Limited.



Individualised Rating-Scale Procedure: A Means of Reducing Response Style Contamination in Survey Data?

Elisa Chami-Castaldi, Nina Reynolds and James Wallace
University of Bradford, UK
e.chami-castaldi@bradford.ac.uk
n.l.reynolds@bradford.ac.uk
j.wallace1@bradford.ac.uk

Abstract: Response style bias has been shown to seriously contaminate the substantive results drawn from survey data; particularly those conducted using cross-cultural samples. As a consequence, identification of response formats that suffer least from response style bias has been called for. Previous studies show that respondents' personal characteristics, such as age, education level and culture, are connected with response style manifestation.

Differences in the way respondents interpret and utilise researcher-defined fixed rating-scales (e.g. Likert formats), poses a problem for survey researchers. Techniques that are currently used to remove response bias from survey data are inadequate as they cannot accurately determine the level of contamination present and frequently blur true score variance. Inappropriate rating-scales can impact on the level of response style bias manifested, insofar as they may not represent respondents' cognitions. Rating-scale lengths that are too long present respondents with some response categories that are not 'meaningful', whereas rating-scales that are too short force respondents into compressing their cognitive rating-scales into the number of response categories provided (this can cause ERS contamination – extreme responding). We are therefore not able to guard against two respondents, who share the same cognitive position on a continuum, reporting their stance using different numbers on the rating-scale provided. This is especially problematic where a standard fixed rating-scale is used in cross-cultural surveys.

This paper details the development of the Individualised Rating-Scale Procedure (IRSP), a means of extracting a respondent's 'ideal' rating-scale length, and as such 'designing out' response bias, for use as the measurement instrument in a survey. Whilst the fundamental ideas for self-anchoring rating-scales have been posited in the literature, the IRSP was developed using a series of qualitative interviews with participants. Finally, we discuss how the IRSP's reliability and validity can be quantitatively assessed and compared to typical fixed researcher-defined rating-scales, such as the Likert format.

Keywords: Scale length, response styles, response bias, survey research, cross-cultural surveys, individualised rating-scale procedure

1. Introduction

Responses to survey questions include both attitudinal information and response bias. The latter can cause measurement error (Greenleaf 1992a), however, it is the former that is of interest. As such, response bias, that is a tendency to respond systematically to questionnaire items on some basis other than the content of those items, needs to be removed from the data before research results will reflect the construct of interest (Paulhus 1991). Response bias can occur due to the presentation of the construct or measurement instrument; or to respondents trying to portray themselves in a certain way. The former is known as *response style*, the latter as *response set* (Rorer 1965). Both types of response bias reduce the validity of research findings (Broughton and Wasel 1990). This paper shows how research instrument design might be used to address *response style bias*. Response style bias can result in spurious findings, construct design can affect its manifestation, and it varies along with individual characteristics and across cultures (Berg and Collier 1953; Couch and Keniston 1960; Hamilton 1968; Lorr and Wunderlich 1980; Crandall 1982; Bachman and O'Malley 1984; Cheung and Rensvold 2000). Response style bias is of particular concern in cross-cultural research (Douglas and Craig 1983; Bachman and O'Malley

1984; Ross and Mirowsky 1984; Farh and Dobbins 1991; Chen, Lee et al. 1995; Javeline 1999; Stening and Everett 2001; van Herk, Poortinga et al. 2004).

This paper briefly reviews the types of response style bias; their measurement, current techniques for reducing their contamination of survey data; and their connection to rating-scale length. Subsequently, the paper details the development of the Individualised Rating-Scale Procedure (IRSP), which seeks to address the problem of response styles by means of a dynamic respondent-centred method of attitude measurement. The paper concludes by outlining how the technique's reliability and validity can be tested.

2. Background

2.1 Response bias, data analysis and rating-scale length

Response styles can be divided into three main groups: those related to the use of particular scale points (e.g. extreme responding and mid-point responding), the spread of responses (e.g. index of dispersion), and the respondent's reaction to item direction (e.g. acquiescence) (Diamantopoulos, Reynolds et al. 2006). Response style bias can affect both level and structural comparability in cross-cultural research (van de Vijver and Leung 1997), and can result in the appearance of differences between groups when no differences actually exist and/or can hide real differences between groups (Heide and Gronhaug 1992; Baumgartner and Steenkamp 2001). They can increase the association between variables so that significant relationships appear, yet response style bias can also decrease associations resulting in no apparent relationships in the data (Chun, Campbell et al. 1974; Lorr and Wunderlich 1980; Bardo and Yeager 1982b; Heide and Gronhaug 1992). Various techniques can be used to reduce response style bias (e.g., partial correlations, ANCOVA, etc), however, these generally require the researcher to estimate the extent to which response style bias is present.

Three main methods exist to estimate response style bias; using uncorrelated items to estimate response styles (Greenleaf 1992b); collecting both attitudinal and behavioural information as response style bias can then be estimated due to its lesser impact on more concrete (behavioural) information (Greenleaf 1992a); and estimating response styles from existing questionnaire items (Baumgartner and Steenkamp 2001). However, all have flaws: The first requires a large bank of uncorrelated items from which to draw (and assumes the items are uncorrelated across cultures), the last assumes that existing items on a single questionnaire would not share common variation. The problems with the second method of measuring response styles is that it is sometimes difficult to develop behavioural measures that directly relate to attitudinal constructs, and that doing so could greatly increase the length of any questionnaire (Chami-Castaldi, Reynolds et al. 2006).

Researchers aim to choose a scale length such that it is long enough to maximise the amount of information that can be collected, yet short enough to get accurate responses (Cox III 1980). One of the concerns when deciding on rating-scale length is its effect on response style bias. While a longer rating-scale is likely to lower extreme responding (Hui and Triandis 1989), it is also likely to increase scale attenuation (Wyer 1969). As such a balance is needed and six or seven response categories have long been considered appropriate (Miller, 1956). Nevertheless, if an 'ideal' rating-scale length could be used (for each respondent), then response styles would be less likely to manifest. Theoretical antecedents of response styles are either dispositional (characteristics of the individual) or situational (context or stimulus related) (Snyder and Ickes 1985; Baumgartner and Steenkamp 2001). The effects of these are considered below.

2.2 Response bias - Dispositional and situational effects

Dispositional effects, of personality (Cronbach 1946; Cronbach 1950; Berg and Collier 1953; Lewis and Taylor 1955; Zax, Gardiner et al. 1964; Iwawaki and Zax 1969; Norman 1969; Merrens 1971; Crandall 1982); age (Osgood, Suci et al. 1957); education (Light, Zax et al. 1965); gender (Berg and Collier 1953; Lewis and Taylor 1955); culture (Smith 2004); and occupation and social class (see Hamilton (1968) for a summary of these studies) have been studied in relation to response style. The findings are varied and can appear contradictory. However, what *can* be said is that there appears to be a link between personal characteristics and response styles, and thus it can be hypothesised that:

> ***H_1:*** *A respondent's 'optimum' number of response categories will be related to the respondent's characteristics.*

Situational factors can discourage or encourage response style manifestation (Snyder and Ickes 1985; Baumgartner and Steenkamp 2001). As such the researcher's measurement choices can directly impact on the data collected from respondents, they should be considered carefully (Rossiter 2002). Factors such as rating-scale length and rating-scale format have been shown to have significant effects on response style bias (Bardo and Yeager 1982b). As such, it is hypothesised that:

> ***H_2:*** *Rating-scales with the number of response categories closest to the respondent's 'optimum' number, will be least affected by response style bias.*

Traditionally designed measurement instruments usually decide on a standardised rating-scale length for all respondents. Advances in scale development, however, indicate that this may not be necessary.

2.3 Respondent-defined rating-scales

There have been several researchers, in the past, that have argued the benefits of involving the respondent in the generation of more meaningful rating-scales (Kilpatrick and Cantril 1960; Battle, Imber et al. 1966; Donnelly and Carswell 2002; Nugent 2004).

Theoretically, it could be argued that there are three key ways in which a respondent could self-anchor a rating-scale:

- Verbally anchor the scale endpoints,
- Numerically anchor the rating-scale endpoints (i.e. defining the number of response categories they would like to use),
- Conceptually anchor the scale endpoints.

Previous studies have experimented with allowing respondents to conceptually anchor the scale endpoints, in that fixed numerical endpoints are shown to the respondent, before they are asked to anchor the two extreme endpoints with a meaningful scenario (specified by the researcher).

Kilpatrick and Cantril (1960) describe their self-anchoring scale approach as one in which each respondent is asked to describe, in terms of his/her own perceptions, the top and bottom of the dimension on which scale measurement is desired, and then to employ this self-defined continuum as a measuring device. Nugent (2004, p. 171) asked respondents "*to imagine a thermometer-type instrument that measures the magnitude of … depression, with higher scores indicating a greater intensity problem with depression and lower scores indicative of a lower magnitude problem.*" Respondents conceptualised their maximum and minimum depression intensities by imagining what, *for them,* would be the most/least depressing scenario they could picture. Bloom et al. (1999) defined this as an individualised rating scale. Measures from these conceptually-anchored rating-scales were found to be reliable (Battle, Imber et al. 1966; Morrison, Libow et al. 1978) and valid (Battle, Imber et al. 1966; Bond, Bloch et al. 1979; Mintz, Luborsky et al. 1979).

These studies demonstrate that respondents can self-anchor a rating-scale, *conceptually.* Should respondents also be capable of verbally-anchoring *and* numerically-anchoring their own rating-scales, this would maximise the *meaningfulness* of the rating-scale. Allowing each respondent to individualise their own rating-scale should, theoretically, account for both dispositional and situational antecedents of response style, and as such result in measures that more accurately reflect respondents' *actual* opinion (Viswanathan, Sudman et al. 2004). As such, it is hypothesised that:

H_3: *Measuring constructs using respondent-defined rating-scales will produce more valid measures.*

The theoretical issues surrounding the use of individualised rating-scales to minimise measurement error has been considered above. However, before these hypotheses could be tested, it was necessary to determine the feasibility of such a technique and develop a working version.

3. Methodology – Development of the individualised rating-scale procedure (IRSP)

Existing methods of self-anchoring rating-scales do not use the three theoretically possible methods – conceptually, verbally and numerically anchoring – simultaneously. The proposed method, would have respondents independently anchor the numerical *as well* as verbal endpoints. As such it was necessary to determine the method's feasibility and develop suitable instructions so that respondents could successfully generate personally *meaningful* response categories.

When examining the feasibility of individualised rating-scales, the concept of agreement/disagreement was chosen to be the focus of measurement. This concept is frequently used in surveys; when measuring respondents' opinions towards products/services (i.e. to what extent they agree/disagree with items), when measuring respondents' psychological characteristics (i.e. Likert-type rating-scales measuring the extent of agreement with a statement representing the 'self'). If an individualised rating-scale procedure (IRSP) were feasible, it would be necessary to compare its performance against a typically-used measurement tool. Given Likert formats are frequently used to measure agreement/disagreement, it was deemed practical and useful that the substantive focus for the IRSP would be on its ability to measure respondents' agreement/disagreement with items.

3.1 Phase one – qualitative exploration & development

To assess the feasibility of the technique and aid its development, fifteen semi-structured in-depth interviews using different verbal and numerical anchoring methods, were conducted. These provided a rich qualitative ground for exploring the order of instructions, phrase structure, semantics and pictorial aids, that yielded the most meaningful rating-scales for respondents.

3.1.1 Main findings

In brief, the results of the interviews indicated that the technique was feasible by:

- Including a visual aid (horizontal line with neutral marked at its centre) to help respondents define their 'ideal' rating-scale.
- Giving respondents a conceptual definition of the agreement and disagreement endpoints.
- Having respondents verbally anchor their endpoints, resulting in them having personally-meaningful representations of the extreme positions.
- Having respondents decide on the number of response categories they want to use.
- Having respondents practice using their individualised rating-scale (IRS) by rating a small batch of uncorrelated items (Greenleaf's (1992b) 16 uncorrelated items were used, given their particular suitability), with the option to revise their numerical endpoints. This task allowed respondents to assess the *meaningfulness* and *distinctness* of their response categories before proceeding to the main survey.

3.1.2 The question of comparable data analysis

Prior to this preliminary testing, the issue of comparable data analysis was considered. Given that, even if it were theoretically possible for respondents to numerically (and verbally) anchor their own rating-scales, we considered an important question: 'How can researchers analyse responses obtained from a group of respondents that each have their own uniquely defined number of response categories?' This problem was deemed to be solvable if certain controls were implemented:

- That the concept of agreement/disagreement would be measured on a bipolar rating-scale with a neutral point at its centre,
- That the neutral point be numerically anchored at '0', with its conceptual definition defined to the respondent. This is to ensure that each respondent is given the same conceptual definition

of the neutral point (i.e. the same point of origin).

- That the endpoints of the scale represent the respondent's conceptual agreement/disagreement extremes (i.e. that the endpoints to the right/left of neutral represent 'the *most* they (the respondent) could possibly agree/disagree with a statement'. This is necessary in order to capture a respondent's entire continuum of cognitive agreement/disagreement.
- That it is assumed that the distance between adjacent intervals on the rating-scale is equal.

The implementation of these controls would allow scores from rating-scales of different lengths, to be converted into a common scores-index, enabling comparable data analysis. This is analogous to a researcher converting data recorded in different units of measurement into a common unit of measurement, such as imperial units (feet, inches) into metric units (metres, centimetres). Figure 1 illustrates this point

Notice that in this example, Respondent 1 has chosen verbal anchors that, for him/her, represent his/her conceptual extreme points on the agreement/disagreement continuum. Whilst Respondent 2 has chosen different verbal anchors, conceptually, they represent the same extreme positions. This is an important control in order for differing rating-scales to produce data that is comparable. Here, if Respondent 1 rated their particular agreement with a statement (using their IRS) as '3', this score would be converted into the IRSP score-index of '1'. Should they report a score of '-2', this would be converted to a '-0.67' in the IRSP score-index.

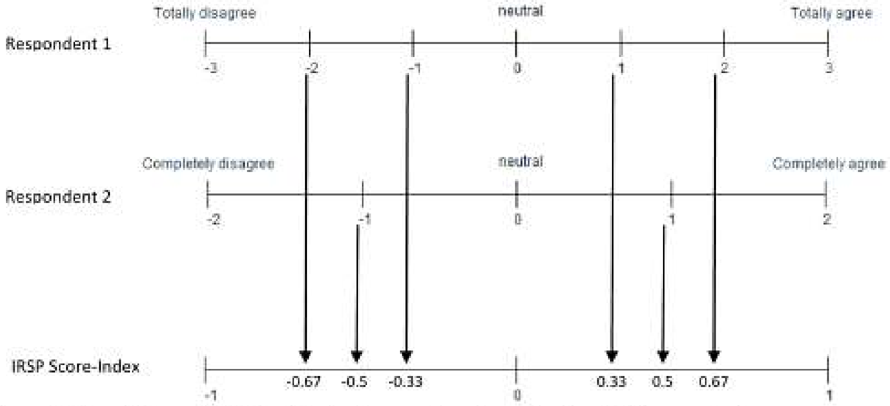


Figure 1: Translation of individualised rating-scale ratings into the IRSP score-Index

When considering data comparability in this context, it is useful to reflect upon an observation made by Kilpatrick and Cantril (1960: 4),

> *"One may question the method by asking, are the data comparable from person to person, or group to group, when one allows each individual to anchor the scale on his own terms? Our position is that they are psychologically directly comparable, that is, the scale level selected by one person or group (average of selections) can be specifically and meaningfully said to be higher, lower, or equal to the scale level of some other individual or group, because the frames of reference of the replies are in fact similar psychologically [...] Meaningful numerical comparisons are then possible."*

Given this method of measurement is dynamic and respondent-specific, it was clear that a computer program was required, to further develop the technique. This would provide respondents with the facility to use interactive visual aids, with their individualised rating-scales being presented to them on screen, to be used to rate survey statements. There was enough data from phase one, to inform the creation of the IRSP software.

3.2 Phase two – Creation of the IRSP software and further development

Using bespoke online survey software, phase two aimed to (a) determine whether more meaningful response categories could be developed using IRSPversion1 or IRSPversion2, (b) ensure that respondents found the resultant survey user-friendly, and (c) ensure data capture was accurate. Sixteen in-depth interviews were conducted using verbal protocols followed by retrospective debriefings (Taylor and Dionne 2000). Respondents were asked to use an online survey to create an individualised rating-scale (using either IRSPv1 or IRSPv2) and rate items using their individualised rating-scale (IRS) whilst 'thinking aloud' (verbal protocol), and were subsequently interviewed about their experiences with the exercise (retrospective debrief). Given that response styles have been linked with dispositional characteristics, these surveys used several psychological scales, and provided the opportunity to test the suitability of these scales for future IRSP development.

3.2.1 Main findings

Listed are some of the findings that were particularly pertinent or interesting:

- The IRSP survey was viable.
- Respondents carrying out IRSPv2 seemed to produce more personally-meaningful rating-scales, than respondents carrying out IRSPv1.
- E.g. The mysterious attraction to '±10' numerical endpoints (mentioned in Chami-Castaldi, Reynolds et al. 2006), only occurred with respondents that carried out IRSPv1.
- Insights were gained, where respondents chose different verbal anchors for the endpoints of agreement and disagreement.
- Insights were gained, with respondents who chose not to use an equal number of response categories for the positive (agree) and negative (disagree) sides of the neutral point.
- Allowing respondents to practice using their IRS on Greenleaf's 16 uncorrelated items, proved a valuable part of the process for several reasons;
- Respondents could ascertain the ease-of-use of their IRS.
- Respondents could reflect on their responses to the items, and give second thought as to whether their response categories are *distinctly* meaningful to them, before proceeding to the rest of the survey. This stage allowed respondents to 'reduce' or 'increase' their rating-scale lengths.
- The verbal protocols showed that respondents who had a tendency to acquiesce, for example, became conscious of it when reflecting on their use of their IRS on Greenleaf's items. In some of the retrospective debriefs, these respondents said that after realising this, they subsequently tried to be more honest and accurate about their opinions when rating items in the main survey.
- Qualitative insights gained after each protocol-debrief interview, frequently led to immediate improvements/modifications for both IRSPv1 and IRSPv2. This meant that IRSP development and data collection was done concurrently.

Due to word limit constraints, only some of these findings will be discussed in more detail.

The quantitative data collected from this small sample of 16 respondents, together with the protocol-debrief qualitative insights, appeared to suggest that the IRSPv2 was more effective than the IRSPv1 (in having respondents produce more *meaningful* response categories). The mean number of response categories for those that used the IRSPv1 was 11.38, with a standard deviation of 7.05. Whereas the mean for those that used the IRSPv2 was 8.88, with a standard deviation of 5.35. Figure 2 helps to illustrate this point.

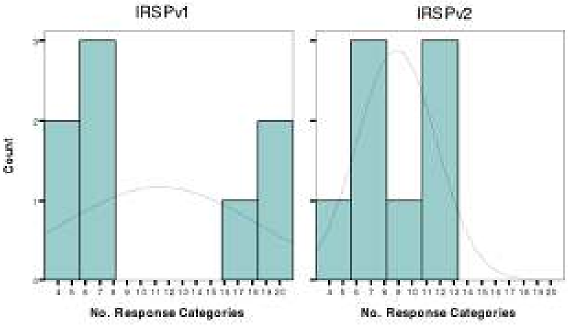


Figure 2: Respondents' chosen no. categories on versions 1 and 2 of the IRSP

On the whole, respondents were able to easily anchor their own verbal endpoints on the agree/disagree continuum in both versions of the IRSP. There were only three respondents who chose peculiar verbal endpoints, shown in the table below (respondents 13, 5 and 11). In their retrospective debriefs, respondents 13 and 5 felt that they had misunderstood the verbal anchoring instructions as a result reading the instructions improperly and possibly the language barrier (given that their first languages are Polish and Chinese, respectively). Respondent 11 apologised for his

seemingly unusual "agree" verbal endpoint, explaining that he had rushed through the exercise and had not read through all the instructions. Nonetheless, these three respondents (together with all the other respondents) indicated that when rating statements, they treated the endpoints on each side of the continuum as 'the most they could possibly agree/disagree with a statement'.

Insights into how some of the respondents conceptually anchored their endpoints before choosing their verbal anchors, proved interesting. Our prompts were worded carefully, so as to encourage respondents to choose personally meaningful verbal labels (i.e. that they would most naturally use). Several respondents talked about the specific scenarios they had pictured when choosing their verbal anchors. For example, Respondent 6 was asked what she was picturing when she chose the word 'absolutely', and whether she pictured herself talking to somebody. She said "Talking to my mum! [Laughs]." She confirmed that she pictured herself saying "I absolutely agree" to something her mum might say. She felt she would equally use the word 'absolutely' when expressing her extreme disagreement in a conversation with her mother.

Table 1: Verbal endpoints inputted by respondents (phase two)

						IRSPv1	
Resp.	**Age**	**Gender**	**First Language**	**Ethnicity**	**National Identity**	**Verbal Endpoint (agree)**	**Verbal Endpoint (disagree)**
1	22	Male	French	Other	French	totally	totally
4	24	Male	English	White	English	definatley	completley
6	22	Female	English	White	English	absolutely	Absolutely
8	21	Male	English	White	English	strongly	Strongly
9	25	Male	English	White	English	Fully	Fully
13	20	Female	Polish	White	Polish	minimaly	slightly [a]
15	21	Female	Guajarati	Asian - Indian	British	totally	highly
16	19	Female	English	Black - African	English	totally	completely

						IRSPv2	
Resp.	**Age**	**Gender**	**First Language**	**Ethnicity**	**National Identity**	**Verbal Endpoint (agree)**	**Verbal Endpoint (disagree)**
2	21	Male	Punjabi	Asian - Pakistani	British	totally	fully
3	24	Male	English	White	British	totally	totally
5	26	Male	Chinese	Asian - Chinese	British	ok	not ok [b]
7	18	Female	English	White	English	totaly	totaly
10	22	Male	German	White	English German	Fully	fully
11	22	Male	English	White	English	Agree	Strongly [c]
12	29	Female	Swedish	White	Swedish	defenetly	totally
14	20	Female	English	White	English	definitely	completely

[a] This respondent indicated that she misunderstood the verbal anchoring instructions, perhaps due to the language barrier.

[b] This respondent indicated that he misunderstood the verbal anchoring instructions, perhaps due to the language barrier.

[c] This respondent indicated that he had misunderstood the first verbal anchoring instruction, partly as a result of trying to rush through the exercise, and not reading all the instructions. This explains why his verbal endpoint for 'agree' did not appear to meet the conditions set.

Where respondents chose to use different verbal anchors to represent their agreement/disagreement extremes (respondents 4, 15, 16, 2, 12, 14 in Table 1), their reasons for doing so were explored in the retrospective debriefs. It was clear that the verbal anchors they generated were personally meaningful to them, when you consider some of the reasons they gave;

Respondent 16

- ""I completely agree" doesn't sound like something I would say, whereas, "I completely disagree" is something I *would* say." She indicated that both her chosen verbal endpoints represented the most she could possibly agree/disagree with something, however, she felt that she would naturally place different adverbs before 'agree' and 'disagree'.

Respondent 4

- "When I saw 'agree', - I just - 'definitely' sprung to mind. I think it's the way I talk…maybe I associate 'definitely' with more positive things. And then, um, when I saw the 'disagree' side of things – I just – thought of another word really. I suppose 'completely' just sprung to mind, I don't know if I associate that with being more *firm* and *disagreeing*…it was just *my opposite*."

It is worth noting that *all* respondents who chose different verbal endpoints indicated that both of their endpoints (although different) represented, for them, their extremes on the agreement/disagreement cognitive continuum. This echoed some of the findings from phase one of our research study (Chami-Castaldi, Reynolds et al. 2006).

This is a particularly interesting finding, given that respondents have been shown to interpret standardised verbal anchors in different ways; intensity, quality (Rohrmann 2003). Rohrmann (2003) highlighted the disadvantages of using standardised (i.e. for all respondents) verbal anchors; inferior measurement quality and proneness to cultural biases. He emphasised the need to create rating-scales using verbal anchors which *reflect the cognitions of respondents.*

In phase one, it was discovered that some respondents desired numerically-imbalanced rating-scales, in that either the positive (agree) or negative (disagree) side of the continuum had more intervals than on the other side. We ensured that the IRSP software, allowed such respondents to individualise their own rating-scales in this manner. See respondents 4, 16 and 14 in Table 2.

Interestingly, respondents 14 and 16 defined a numerically-imbalanced IRS right from the start, whereas respondent 4 initially defined his extreme disagree at '-3' and his extreme agree at '3', providing him with a numerically-balanced IRS with seven categories (a typical length with Likert formats). Respondent 4, after completing Greenleaf's items, realised that whilst he was using all of his varying levels of *agreement*, he didn't *think* any 'finer' than *two* stages when it came to rating his level of *disagreement* with something. Respondent 14 also chose to modify her IRS (from [-5←0→6] to [-4←0→5]), before proceeding to the main survey. The process of practicing the use of her IRS on Greenleaf's items, led her to evaluate that whilst she did not need quite so many response categories, she still desired more "shades of grey" when 'agreeing' than when 'disagreeing'. This highlights another potential problem with having respondents use researcher-defined fixed rating-scales to rate statements. It is clear that some respondents gradate their levels of agreement to a greater/lesser extent than with disagreement. This has implications for the instrument validity of researcher-defined fixed rating-scales, and lends further support to the argument for individualised rating-scales

Some respondents became aware of their tendency to adopt a particular response style, after using their IRS to rate Greenleaf's 16 uncorrelated items, and being prompted to reflect upon their responses (a pictorial aid with instructions assisted them with this reflection). One particularly interesting case was respondent 4. When prompted to reflect on the meaningfulness of his [-3←0→3] IRS (after using it to rate Greenleaf's statements), not only did he modify his numerical endpoints (feeling "more comfortable" with a numerically-imbalanced [-2←0→3] IRS), he stated that prior to this point (in his retrospective debrief) that "I didn't realise I was such an agreeing person." When asked whether his realisation of this affected the way he completed the rest of the survey, he responded "I looked at the statements [psychological items] like I did with the first set [referring to Greenleaf's items] and just thought, "*do* I completely agree with this?" Um, I didn't want it to be as neutral as before. I wanted to be a bit more assertive with my answers."

Table 2: Numerical endpoints inputted by respondents (phase two)

	IRSPv1					
Resp .	**Verbal Endpoint (agree)**	**Verbal Endpoint (disagree)**	**Numerical Endpoint (agree)**	**Numerical Endpoint (disagree)**	**Modified Numerical Endpoint (agree)**	**Modified Numerical Endpoint (disagree)**
1	totally	totally	8	-8		
4	definatley	completle y	3	-3	3	-2
6	absolutely	Absolutel y	2	-2		

	IRSPv1					
Resp .	**Verbal Endpoint (agree)**	**Verbal Endpoint (disagree)**	**Numerical Endpoint (agree)**	**Numerical Endpoint (disagree)**	**Modified Numerical Endpoint (agree)**	**Modified Numerical Endpoint (disagree)**
8	strongly	Strongly	2	-2		
9	Fully	Fully	10	-10		
13	minimally*	slightly*	3	-3		
15	totally	highly	10	-10		
16	totally	completel y	3	-4		

	IRSPv2					
Resp .	**Verbal Endpoint (agree)**	**Verbal Endpoint (disagree)**	**Numerical Endpoint (agree)**	**Numerical Endpoint (disagree)**	**Modified Numerical Endpoint (agree)**	**Modified Numerical Endpoint (disagree)**
2	totally	fully	3	-3	2	-2
3	totally	totally	3	-3		
5	ok*	not ok*	6	-6		
7	totaly	totaly	5	-5		
10	Fully	fully	3	-3		
11	Agree*	Strongly	2	-2		
12	defenetly	totally	4	-4		

14	definitely	completel y	6	-5	5	-4	b

[a] This respondent opted for a numerically-imbalanced scale, in that they desired more response intervals for disagreeing then for agreeing.

[b] This respondent opted for a numerically-imbalanced scale, in that they desired more response intervals for agreeing then for disagreeing.

[c] This respondent initially defined their IRS as [-3←0→3], which is numerically balanced. He subsequently modified it, after rating Greenleaf's items, to a numerically-imbalanced scale [-2←0→3].

[-] The reason for these peculiar verbal anchors was explained in Table 1.

When comparing his spread of responses on Greenleaf's uncorrelated items using his [-3←0→3] IRS (Figure 3), to his spread of responses on the subsequent items using his modified [-2←0→3] IRS (Figure 4), it would seem that his tendency to acquiesce was reduced.

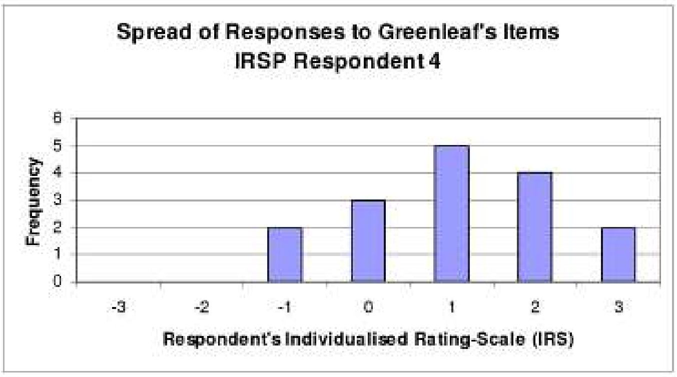
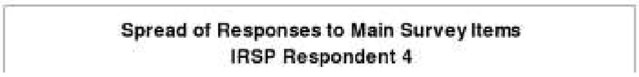


Figure 3: Responses to Greenleaf's Items



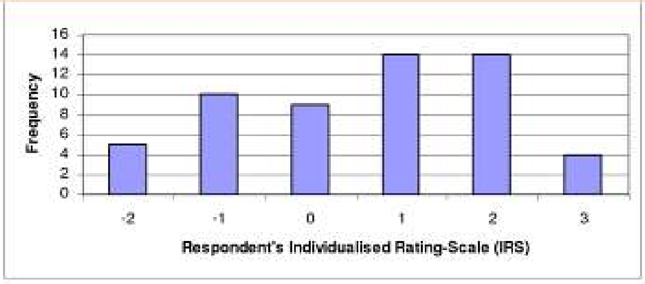


Figure 4: responses to main survey items

3.3 Phase three – IRSP software development and pilot test

A pilot survey with 51 respondents, 24 using IRSPv1 and 27 using IRSPv2, was conducted. This phase was necessary in order to;

- provide a live test for the robustness of the IRSP survey software (i.e. respondents simultaneously completing the survey).
- provide additional data to decide whether IRSPv1 or IRSPv2 would be chosen for the reliability and validity assessment.
- examine how effectively respondents were able to independently carry out the online survey.

Unfortunately, these findings cannot be discussed within the scope of this paper.

4. Future development

In response to Baumgartner and Steenkamp's (2001) call for forms of measurement that are resistant to response style bias, this paper has considered the impact of rating-scale length on response styles. The literature affirms that there is no single 'optimum' rating-scale length for all situations, as it is dependent on the respondent in question (Bonarius 1971; Hui and Triandis 1989; Si and Cullen 1998). This work has established the feasibility of respondents defining their own individualised rating-scales. However, while it may be *possible* to have individualised rating-scales, the *reliability* and *validity* of this measurement method has not yet been established.

Further quantitative research is necessary in order to test the hypotheses presented earlier in this paper (H_1 to H_3). The use of a multi-group experimental design is planned, together with the inclusion of previously validated psychological constructs in the online survey. This would help determine test-retest reliability, internal consistency, discriminant, convergent and nomological validity (Bagozzi 1994).

To conclude, if individualised rating-scales can be shown to reduce response style bias, then it would be possible to greatly improve research instrument design. Proving this method is legitimate, would be particularly advantageous for cross-cultural researchers where response style bias is especially problematic.

Acknowledgements

We would like to thank Patrick Oladimeji for building the IRSP Survey Software without which we could not have developed the IRSP technique further. His impressive programming skills and hard work have provided us with the tools necessary to continue developing the IRSP and to test it in a forthcoming online survey.

References

Baumgartner, H. and Steenkamp, J.-B. E. M. (2001), "Response Styles in Marketing Research: A Cross-National Investigation." *Journal of Marketing Research*, Vol. 38, No. 2, pp 143-156.

Chami-Castaldi, E., Reynolds, N. L. and Cockrill, A. (2006). "Respondent-Defined Scale Length: A means of Overcoming Response Style Contamination?" ANZMAC Conference 2006: Advancing Theory, Maintaining Relevance., Queensland University of Technology, Gardens Point Campus, Brisbane, Australian and New Zealand Marketing Academy (ANZMAC).

Couch, A. and Keniston, K. (1960), "Yeasayers and naysayers: Agreeing response set as a personality variable." *Journal of abnormal and social psychology*, Vol. 60, No. 2, pp 151-174.

Diamantopoulos, A., Reynolds, N. L. and Simintiras, A. C. (2006), "The impact of response styles on the stability of cross-national comparisons." *Journal of Business Research*, Vol. 59, No. 8, pp 925-935.

Greenleaf, E. A. (1992a), "Improving rating scale measures by detecting and correcting bias components in some response styles." *Journal of Marketing Research*, Vol. 29, No. 2, pp 176-188.

Greenleaf, E. A. (1992b), "Measuring extreme response style." *Public Opinion Quarterly*, Vol. 56, No. 3, pp 328-351.

Hui, C. H. and Triandis, H. C. (1989), "Effects of culture and response format on extreme response style." *Journal of Cross-Cultural Psychology*, Vol. 20, No. 3, pp 296-309.

Kilpatrick, F. P. and Cantril, H. (1960). *Self-Anchoring Scaling: A Measure of Individuals' Unique Reality Worlds* Washington DC: The Brookings Institution.