Deciding on the Scale Granularity of Response Categories of Likert type Scales: The Case of a 21-Point Scale

Noel Pearse
Rhodes Business School, Rhodes University, Grahamstown, South Africa
N.Pearse@ru.ac.za

Abstract: This research investigates the use of a 21-point Likert type scale in the design of a questionnaire that explores factors related to staff turnover and retention. The paper examines the notion of granularity in researcher-defined fixed rating scales, where granularity refers to the number of response categories or cut off points that are imposed on a scale (Smithson 2006). The aim of this research paper is to examine the usefulness of a scale with high granularity, from the perspectives of respondents and the researcher. The questionnaire was administered among employees in three different public sector organisations in South Africa, to create a combined data set of 178 respondents. Informing the formulation of the hypotheses was Parducci's (1965 cited in Tourangeau, Rips & Rasinski 2000) range-frequency model, which assumes that respondents make use of the various response categories available with equal frequency, if they are evenly spaced. It was therefore hypothesised that (1) there are no significant differences in the frequency of use of the 21 response categories, implying that all of the response categories are useful to respondents; (2) that there would be no difference in the response pattern of respondents when different scale types and lengths are used, implying that increasing the scale granularity did not lead to redundancy; and (3) that there are no significant differences in the variation of responses with ongoing use of the scale. That is, if the scale was useful to respondents, they would continue to use a wide range of the response options available. Chi-square tests were primarily used to test the hypotheses. It was concluded that the 21-point scale was useful to respondents and by implication to researchers as well. This was evident in the spread of responses across the 21 response categories of the scale, and that even with prolonged use, they continued to utilise a wide range of response options. It was recommended that researchers should give more explicit attention to scale granularity when designing a questionnaire and that further research is required to assess the value of various levels of scale granularity.

Keywords: questionnaire design; scale construction; likert scale; scale granularity

1. Introduction

Lee and Soutar (2010) maintain that while there are several ways to gather data of a quantitative nature, rating scales remain the most popular. There has been much interest recently in alternative scale formats such as graphic scales (Cook, Heath Thompson & Thompson 2001), single item scales (Lee, Douglas & Chewning 2007), tailoring scales for multicultural and/or multilingual settings (Arce-Ferrer & Ketherer 2003) and the development of individualised rating scale procedures as an alternative to researcher-defined fixed rating scales (Chami-Castaldi, Reynolds & Wallace 2008). Nevertheless, the popularity of researcher-defined scales seems to persist, and yet there seems to be relatively little attention paid by researchers to their decision to adopt a specific rating scale design.

This research investigates the use a particular researcher-defined fixed rating scale, namely a 21-point Likert type scale, in the design of a questionnaire that explores factors related to staff turnover and retention. It is argued that the properties of such a scale enhance statistical analysis. The aim of this research paper is therefore to examine the use of such a scale by respondents, and by extension, to infer the usefulness of the scale for the researcher.

2. Likert type scales

According to Rattray & Jones (2007), Likert type scales are one of a range of scale types that researchers can choose from, and they identify Frequency, Thurstone, Rasch, Guttman, Mokken and Multiple choice formats as alternatives. DeVellis (2003) refers to a Likert scale as a type of response format for a scale item, rather than a scale type. This distinction is helpful, as it serves to differentiate summated scale types from the characteristics or format of a single item. Summated scales such as the Guttman and Thurstone scales consist of a number of items making up the scale. In contrast typical response formats for a single item include the Likert, semantic differential, and visual analogue scales. This study's focus is on the response format of single items.

Likert type scales can be traced back to the work in the 1930's by their namesake, Rensis Likert, who experimented with a simpler response format for various Thurstone attitude scales (Likert 1932; Likert Roslow & Murphy 1934). A statement was provided and respondents were given one of five response

Reference this paper as: Pearse, N. "Deciding on the Scale Granularity of Response Categories of Likert type Scales: The Case of a 21-Point Scale *The Electronic Journal of Business Research Methods* Volume 9 Issue 2 2011 (pp 159-171), available online at www.ejbrm.com

options by which to describe their reaction to the statement. These options were: "Agree with the statement", "Strongly agree with the statement", "Disagree with the statement", "Strongly disagree with the statement", or "Undecided". Respondents indicated their reaction by writing down a symbol that corresponded to the option, rather than writing down a number. In addition, Likert's original instructions set out the options in the order that they are listed above, and not in a more intuitively logical sequence that would suggest a continuum of reactions ranging from "Strongly disagree" to "Strongly agree". This continuum of responses has certainly become the more popular format (Dawes 2008). Furthermore, a neutral response was not provided by Likert, but rather the option of being "undecided". It is interesting to note that there is no evidence of Likert personally providing a theoretical justification of his method (Roberts, Laughlin & Wedell 1999). Nevertheless the Likert scale is seen to be consistent with classical test theory, which was developed subsequently (Roberts, *et al*. 1999).

A Likert scale has several defining characteristics, namely a declarative statement, and a number of response categories that have distinct cut-off points and assume linearity and equal intervals between them. These characteristics are now discussed in more detail.

Firstly, Likert scales consist of a "declarative sentence, followed by response options that indicate varying degrees of agreement with or endorsement of the statement." (DeVellis, 2003 pp.78-79). Referring to attitudinal scales in particular, Roberts, *et al*. (1999) note that the declarative statement typically expresses a clearly positive or negative opinion and not a neutral one. This is designed to solicit more definitive responses from respondents, rather than eliciting muted, unvarying responses. Furthermore, some argue that a well designed questionnaire should have some of the items reversed so that response bias in the form of acquiescence is reduced (Churchill 1979). However, there are voices of caution that reversing an item may not create an exact opposite statement and can lead to a reduced internal consistency of the scale (Rodebaugh, Woods, & Heimberg 2007; Wong, Rindfleisch & Burroughs 2003).

Secondly, in adopting a closed-ended question format (Dillman, Smyth & Christian 2009), Likert scales make use of a number of response categories. Cook, *et al.* (2001) observed that Likert scales explicitly present their scoring metric to the respondents and expect them to limit their choice to one of the options provided. As noted above, the response categories of Likert's original scale were not placed on a continuum. However, there is an increased tendency to provide graded response scales, and typically in a "disagree-agree" format (Roberts, *et al*. 1999). With Likert type scales, the appropriate number of categories and the type of response options used is something that the researcher should be giving more attention to when designing a questionnaire.

There has been much debate on what should be regarded as an optimal number of response categories. Dillman, *et al.* (2009) recommend that only four or five categories should be used, while Fink (1995) recommends five to seven, and Foddy (1994) concludes that a minimum of seven categories is required to ensure scale validity and reliability. Some researchers have preferred to make use of a nine-point format instead (e.g. Almli, Naes, Enderli, Sulmont-Rossé, Issanchou & Hersleth 2011; Lee & Soutar 2010), and more rarely, a 15-point format (e.g. Chaiken & Eagly 1983). Nevertheless, the five- or seven-point formats would appear to be the most prevalent (Dawes 2008). The seven-point format typically provides the following response options: "1 = Very Strongly Disagree", "2 = Strongly Disagree", "3 = Disagree", "4 = Neutral", "5 = Agree", "6 = Strongly Agree", "7 = Very Strongly Agree". The five-point format typically does not have the two extreme options of "Very Strongly Disagree", or "Very Strongly Agree".

While a disagree-agree format is most typically associated with a Likert type scale, it should be noted that there are other less popular types of response options that are also used, that would adhere to the characteristics of the Likert type scale format. These formats include endorsement (true-false), frequency (always-often-sometimes-never), intensity (mild-moderate-severe), influence (e.g. size of the problem) and comparison (more-less than others) (Fink 1995). Furthermore, Dillman *et al.* (2009) propose that Likert type scales can be unipolar or bipolar, depending if a zero point is placed at the end of the scale or somewhere towards the middle, between opposite dimensions.

Thirdly, Likert scales have distinct cut-off points and often assume linearity and equal intervals between various response alternatives (Rattray & Jones 2007), thereby facilitating the statistical processing of (at a minimum) interval-level data, and using parametric statistics (Terre Blanche &

Durrheim, 1999). However there has been some debate as to whether Likert scale data can be assumed to be interval data, or to be only ordinal in nature (See Pedhazur & Schmelkin, 1991).

Likert scales seem to have become the most popular format used in scale design (Foddy, 1994:168). So much so that researchers rarely interrogate, or justify their adoption or use of this scale format. This paper examines the notion of granularity in researcher-defined fixed rating scales. Granularity refers to the number of response categories or cut off points that are imposed on a scale (Smithson 2006). Granularity is first examined from the perspective of the researcher and then from the respondent's viewpoint.

It is acknowledged that the value, quality and meaningfulness of a questionnaire is also affected by for example: the wording of each of the items, the characteristics of the sets of items that make up the scale or questionnaire, the sequence of the items, and the design and layout of the questionnaire (see for example DeVellis 2003; Dillman, *et al.* 2009; Fink 1995; Schuman & Presser 1996). Furthermore, all attempts to quantify responses have to deal with problems of representation, objectivity and correspondence (Terre Blanche & Durrheim, 1999). Representation is related to uncertainty about what characteristics the numbers actually represent, objectivity interrogates the rules related to the assigning of numbers, and correspondence is concerned with the extent to which differences between scores on a measure correspond with the actual differences that they represent. All of these issues are not explored in detail here, except where they may relate to granularity.

3. The researcher's perspective on scale design

It is argued that increased granularity can achieve three main interrelated objectives, namely to ensure more precise data is collected, to increase the reliability and validity of the data, and - from the perspective of statistical analysis - to ensure that more useful data is gathered.

Firstly, a researcher is concerned with the precision of the question responses offered to respondents in a researcher-defined fixed rating scale. This would include considerations of the inclusiveness, exhaustiveness and mutual exclusivity of categories (Bourque & Fielder 1995; Fink 1995). However, to achieve precision, a central concern of the researcher is to achieve an optimal level of granularity that combines linguistic differentiation with measurement precision (Smithson 2006). This implies that the categories need to be meaningful to respondents and not trivial or ambiguous (Dillman, *et al.* 2009; Fink 1995), while simultaneously providing as wide a range as possible of alternative, but significant responses. To achieve this ideal level of precision, the researcher also needs to take into account the cognitive ability and level of patience of respondents (Cook, Heath, Thompson & Thompson 2001; Oppenheim 1966). Clearly, if a scale has more options, it will be more difficult and take longer for respondents to make a choice among the alternatives and the therefore a questionnaire would take longer to complete. This concern about precision is therefore largely a matter of correspondence (Terre Blanche & Durrheim, 1999).

Secondly, the choice of granularity can affect the reliability and validity of the scale. There has been much debate related to the statistical properties of various levels of scale granularity or coarseness. For example Krieg (1999 p. 763) notes that scale coarseness can affect the bias of a scale as well as "the mean, variance, covariance, correlation coefficient, and the reliability of scores". Cook, Heath Thompson & Thompson (2001) note that in theory, increasing the number of response alternatives used, will automatically increase score variance, and that this has the potential to increase score reliability. However, the results of studies in this regard reveal mixed results and do not always support Symonds' (1924) assertion that using seven scale points would achieve an optimal level of reliability. For example, Cook and Beckman (2009) compared nine- versus five-point rating scales and concluded that nine-point scales appeared to provide more accurate scores. See Miller (1956), Molenaar (1982), Foddy (1994) and Coelho and Esteves (2006) for further discussion of this topic. Of relevance to and in support of this study though, is the finding of Andrews (1984) who made use of various rating scales, some of which had more than 20 categories, and concluded that the validity of the rating data improved with an increased number of categories.

Finally, related to the reliability of the scale, is the fact that data derived from scales with higher levels of granularity are more likely to produce more meaningful results when subjected to statistical analysis. Many statistical procedures rest of the assumption of variability of data. When the data set lacks variance in respondents' scores, this is reflected in inconclusive statistical results. As DeVellis (2003 p. 75) states succinctly "A measure cannot covary if it does not vary".

This consideration is also important when researchers are looking for evidence of change over a period of time or between groups or individuals. Pejtersen Bjorner, & Hasle (2010) note that slight differences may prove to be statistically significant with large sample studies, but insignificant for small samples. Similarly, when making group comparisons a slight variation in scores may be more evident than when comparing individual scores (Pejtersen *et al.* 2010). The ability of the research to identify these minimally important differences (Pejtersen *et al.* 2010) is affected by measurement precision, and in the case of questionnaires - the granularity of the scale. In summary, it would appear that developing scales that have a high level of granularity can be of more value to the researcher by rendering accurate, reliable and valid data that is more suited to statistical analysis, as long as the level of granularity is also meaningful to respondents. In the next section, the respondent's perspective on scale construction is examined.

4. The respondent's perspective

As a type of closed-ended question (Dillman, *et al.* 2009), Likert type scales represent a forced choice format of questioning. Given the fixed and limited range of options provided, it is assumed that respondents would generally respond better to a wider range of meaningful choices. The question remains though, as to what can be regarded as an optimal number of response categories. It has previously been noted that the number of categories needs to be meaningful to respondents and not trivial. Properties of the response categories are explored in more detail here. It is acknowledged that having the items reviewed by experts (DeVellis 2003) and pretesting or piloting the questionnaire (Dillman, *et al.* 2009), will also facilitate the design of a questionnaire with meaningful response categories. However, this usually takes place towards the end of the design process, if it is given any purposeful and conscious attention at all. The focus of interest here is on a more considered design of the response categories themselves, while the questionnaire is being constructed.

It has already been established that having 20 or more response categories can be meaningful for the respondent (Andrews 1984), depending on the nature of the question items. It is therefore not simply a matter of simplistically setting predetermined limits on the minimum or maximum number of categories, but rather establishing what is required and how best to solicit an accurate and meaningful response, given the nature of the question that needs to be answered. However, items with different types of, and a different number of response categories are not responded to in the same way. This raises further issues about the type and number of response categories to use, as well as their numeric labelling.

Firstly, Rosch (1975) notes that numbers are not responded to equally, but that some (such as 1, 10, and 100) stand out as "cognitive reference points" and thereby draw a disproportionate share of responses. Rosch (1975) attributes this to rounding effects. It is further argued that these effects may be compounded if one of these reference points is also the largest numeric anchor, as Foddy (1994) cites evidence of a positive response bias towards the highest numeric value. On the other hand, there may also be a tendency for respondents to make finer discriminations closer to these reference anchors, and be less discerning when choosing responses that are further away from them (Tourangeau, Rips & Rasinski 2000). Yet other research indicates that some respondents have a tendency to select more extreme responses, partly as a result of their cultural or language background (Arce-Ferrer 2006; Arce-Ferrer & Ketherer 2003). That is, there is a higher frequency of responses at the upper and lower ends of the scale, regardless of the scale or item's content. Hui and Triandis (1989) reported significant extreme response patterns when a five-point scale was used, whereas this pattern was negligible when a ten-point scale was used. The net result of these various effects is that there may be a disproportionate accumulation of responses at the positive extreme of a scale, particularly if the end-point is, for example, 10 or 100. It is evident that this could affect the linearity of the scale, skew the distribution of responses, and generally complicate statistical analysis and compromise results (See Arce-Ferrer 2006 for a more detailed discussion in this regard). From the perspective of scale granularity, higher levels of granularity would be more prone to the distortion effects of cognitive reference points, simply because respondents are given a wider range of alternatives.

A second issue related to the type and number of response categories is the use of a central or neutral category. There is much debate regarding how respondents react to the provision of a middle or neutral response option that is offered as part of the scale continuum. Foddy (1994) notes that Likert type scales have been prone to a central tendency error, implying that this may be exacerbated by the provision of a mid-point. However, he also cites evidence of the number of "uncertain"

responses declining as the number of categories increases, which could suggest that a neutral or mid-point position was being used as a fallback position when respondents did not recognise a category that matches their ideal or desired response. In investigating the inclusion of a middle alternative in a forced-choice attitude item, Schuman and Presser (1996) concluded that this alternative attracted about 10 to 20% of responses, but that this tended to affect both polar positions proportionately, so that the overall distribution of opinion was not unduly affected if the middle alternative was to be excluded. As a rule of thumb, Fink (1995) recommends using a neutral response only if it is a valid response. That is, it serves as a neutral or midpoint, or a valid "no opinion"/"don't know" option that the researcher wishes to make available to respondents, but also recognising there is the risk that it may be used as an excuse for not answering the question. This also raises the possibility of including a "not applicable" option at the end of and separated from the list of response categories, either as an alternative to, or in addition to the neutral response (Dillman *et al*. 2009).

Thirdly, a related issue is the assignment of the number zero to the neutral position and by implication, the use of a numeric continuum that has negative and positive values. Dillman, *et al.* (2009) have cautioned the researcher to carefully evaluate the use and possible impact of numeric labels. In particular, in some cases the use of a zero and negative values can have negative connotations for respondents and has shown to affect their responses by, for example, not using these values in questions where they had to rate themselves. Similarly, Kubovy and Psotka (1976) reported a tendency of respondents to avoid the numbers zero and one, surmising that these values were not seen as random, and furthermore that the value zero was interpreted as meaning the complete absence of the characteristic being investigated. It is further argued here, that the familiarity of the scale granularity will determine how meaningful the response categories are to respondents, and hence their use of the various categories on the scale. In particular it is proposed that the metric system is well ingrained and therefore provides a useful framework for creating response categories. As Dawes (2008: 63) notes "many people are familiar with the notion of rating 'out of 10'." Cummins and Gullone (2000) are of the view that a decile scale is the most instinctive of formats and easy for people to relate to; having learnt to count as children by counting ten fingers or toes. A word of caution is required though, in recognising that there may be cultural differences (c.f. Huynh, Howell & Benet-Martinez 2009; Lee & Soutar 2010) in response to a decile scale, especially in countries where the metric system is not well embedded.

Table 1 provides a synopsis of the advantages and disadvantages of high versus low levels of scale granularity. Given the arguments presented above, it is therefore proposed that it is more than reasonable to design a questionnaire that makes use of a 21-point bipolar scale, ranging from a score of minus ten to score of plus ten, with a middle point score of zero for a neutral response. Such a scale will capitalise on the advantages of high levels of granularity as set out in Table 1. In addition, the use of the metric system in the design of the questionnaire should address some of the disadvantages that scales of high granularity typically display, by providing more familiar scale categories. Furthermore, as explained below, the use of a zero midpoint and positive and negative numbers for the scale is not arbitrary, but is aligned to the intent of the scale and the underlying concepts being measured. Concerns related to category redundancy (e.g. linguistic differentiation, complexity, cognitive ability, distortion and acquiescence) remain, but are to be examined through the hypothesis testing. In the next section the questionnaire that formed the basis of this study is described and the theoretical rationale for developing the 21-point scale is presented.

Table 1: Scale granularity advantages and disadvantages

	Advantages	Disadvantages
Low granularity	Quicker to answer.	Scale exhibits more bias. Respondents become frustrated if their option is not represented in the options made available.
High granularity	More likely to have inclusive, exhaustive and mutually exclusive categories. More precise data. Higher reliability and validity. Increase score variance. More meaningful statistical results. Fewer neutral and "uncertain" responses.	Linguistic differentiation of categories more complex. More difficult to differentiate categories and to make a choice. Cognitive ability of respondents may hinder the proper use of the scale. Respondents may become impatient. Categories may become trivial. More prone to the distortion effects of cognitive reference points.

5. The scale

The literature on staff turnover and intention to quit has a long history, with Mobley and his associates (Mobley Horner & Hollingsworth 1978; Mobley Griffeth Hand & Meglino 1979; Mobley 1982a; 1982b) making a significant early contribution. The literature on staff retention has a more recent history. What is apparent when examining these two bodies of literature, is that while the concepts are closely related – and could be regarded as two sides of the same coin – the authors on staff turnover generally do not cite literature on staff retention, and vice versa. However, there are cases where the same factor is identified by both bodies of literature, albeit in different forms. See for example the factors related to the existence, or lack of staff support that contribute to retention or turnover, respectively (Hatton Emerson Rivers Mason Swarbrick Mason Kiernan Reeves & Alborz 2001**;** Riggs & Rantz 2001).

In the light of this observation, a questionnaire was designed to explore whether in fact employees would differentiate between turnover and retention variables as two distinct sets, or (as was anticipated) that a variable could be considered as contributing either to the employee's intention to stay or leave, depending on its level or characteristic in the organisation. Questions were therefore stated neutrally to allow for both of these possibilities. This formed the second section of the questionnaire that is briefly described below, and is the focus of this paper.

The questionnaire that has been designed consists of 124 questions. The five sections of the questionnaire are: (1) a 19 item biographic section which gathered background information on the respondent; (2) 67 items on turnover and retention factors, where the respondents was asked to rate the importance of a wide range of internal push factors (perceived to increase their intention to leave their employer) and/or pull factors (perceived to increase their intention to stay on); (3) 22 items describing management factors where the respondent was asked to rate the perceived value of a number of possible management interventions aimed at retaining them; (4) seven items describing external factors that may influence their decision to leave the organisation, but were generally regarded as outside of the employer's direct control and (5) nine items on intention to quit, where the respondents was asked to rate their likelihood of leaving in the near future.

6. Research hypotheses

To test the utility of the 21-point scale, the following hypotheses were generated:

Ho1: There are no significant differences in the frequency of use of the 21 response categories.

Ha1: There are significant differences in the frequency of use of the 21 response categories.

The idea that the 21 response categories are used with more or less equal frequency suggests that all of the response categories are useful to respondents. This notion is derived from Parducci's range-frequency model. According to Parducci's (1965, 1974, cited in Tourangeau, Rips & Rasinski 2000) model, it is assumed that respondents make use of the various response categories available with equal frequency, if they are evenly spaced. On the basis of this model it could then be assumed that an approximately uniform frequency distribution of responses could be expected across all response categories, or at least that all response categories would be used with regular frequency.

It should be noted that this assumption of a uniform distribution that underpins the range frequency model does not represent the natural distribution of a Likert scale. That is, the selection of response category for an item can be described as a dominant response process, which is more accurately represented by a cumulative model (Roberts *et al*. 1999). What this means is that for a positively worded item, there is an increased likelihood that the individual will agree with the statement rather than disagree. In contrast, negatively worded items are more likely to have lower levels of agreement than disagreement. By implication, there is therefore a higher probability that the null hypothesis will be rejected, simply because the response required is in the format of a Likert scale. This hypothesis was tested by making use of the Chi square test for independence.

The reasoning of the range-frequency model can be extended to scales of varying length. This extension would imply that it would be reasonable to assume that there would be no significant difference in the response patterns when different scale types and lengths are used. In other words, respondents are making regular use of the increased variety of response categories that are provided

in scales of higher granularity. In this study, this would suggest that even increasing the scale granularity to 21 categories has not led to redundancies or meaningless categories. This leads to the second hypothesis.

Ho2: There are no significant differences in the distribution of responses between two different scales (that the same set of respondents has used in response to a questionnaire).

Ha2: There are significant differences in the distribution of responses between two different scales.

To examine this hypotheses, responses to Section 2 of the questionnaire (which makes use of the 21-point scale), were compared to the responses to Section 3 of the questionnaire, which makes use of an 11-point scale, with values ranging from zero (meaning the item is "not important at all") to ten (meaning the item is "extremely important"). The similarity in construction of the two scales is apparent, the main difference being the use of a negative set of response categories in the 21-point scale. This hypothesis is tested by making use of the Chi Square test of association, or a contingency table Chi square. Prior to testing this hypothesis, the comparative 11-point scale will be tested to see if the 11 response categories are used with equal frequency, and thereby can serve as a suitable comparative scale.

Finally, it is assumed that if the scale is useful to respondents, they would continue to use the range of response options throughout the questionnaire. That is, there would not be evidence of "fatigue" or acquiescence evident in a reduced range of options being used later in the 67-item section of the questionnaire, compared to earlier responses. This leads to the third hypothesis that is tested by making use of the Chi Square test of association:

Ho3: There are no significant differences in the variation of responses between the first and second halves of Section 2 of the questionnaire.

Ha3: There are significant differences in the variation of responses between the two halves.

7. Research procedure

Equivalent versions of the questionnaire were administered among employees in three different public sector organisations in South Africa, to create a usable combined data set from 178 respondents. The only differences between the three versions of the questionnaire were in the adaptation of some of the items of the biographic section's response categories to suit the particular organisation context. Furthermore, due to a printing error in one of the versions, the last item of the section was omitted and so only 66 items appeared in the second section and not 67.

The data was captured in Microsoft Excel and various quality checks applied, including descriptive statistics (such as calculating minimum and maximum values), to check the accuracy and completeness of data capture. Further descriptive and inferential data analysis was conducted in Microsoft Excel following the guidelines of Remenyi, Onofrei and English (2010). The response categories used by respondents were firstly subjected to descriptive statistical analysis to examine the underlying response patterns. Thereafter, inferential statistics in the form of Chi-square tests were conducted to test the three hypotheses.

8. Results

As described above, Section 2 of the questionnaire was designed to measure respondent's responses to neutrally stated items describing factors related to turnover or retention, and which were assessed by respondents in terms of their importance as push or pull factors affecting their intention to stay with or leave the organisation. This scale therefore combined two decisions, namely (1) deciding if the factor was a push or pull factor and then (2) deciding on the importance of the factor. Given the nature of these question items, the distribution of responses between positive (scores ranging from +1 to +10) and negative (scores ranging from -1 to -10) categories for all 67 items combined, was first examined, and is illustrated in Table 2. The results indicate a reasonable balance between positive (58.3%) and negative (31.8%) responses to warrant further analysis of respondent's use of the full scale. The table also illustrates that the neutral response was not overused, reflecting just under 9% of all responses. The missing data category was also omitted from the Chi square analyses. Of the 127 blank or missing items, 40 were for the missing item 67 in one of the versions of

the questionnaire. The rest of the missing items seemed to be quite widely distributed across question items and respondents.

Excluding the missing values as a category, a uniform distribution would provide a mean percentage frequency of approximately 4.76% for each of the 21 categories. The results of the Chi square tests are summarised in Table 3. The results indicate that the first null hypothesis cannot be rejected, implying that the frequency distribution of responses amongst the 21 categories are sufficiently uniform.

Table 2: Frequencies of responses by aggregated category

	Negative	Neutral	Positive	Blank	Total
Actual	3788	1056	6955	127	11926
Percentage	31.8	8.9	58.3	1.1	100.0

Table 3: Chi square results to test for uniformity of distribution

Degrees of freedom	20
Significance level	5%
Test calculated	25.8473
Critical value	31.41
Decision	As the Calculated Chi Square value is less than the Critical value, the null hypothesis cannot be rejected

The nature of this distribution was explored further using descriptive statistics. Table 4 illustrates the combined percentage frequencies for all 67 items for all 21 response categories. The table has been truncated and the remaining categories added below so that it can fit within the page width. In addition, the sequence of the positive categories (i.e. +1 to +10) has been reversed so as to facilitate comparison with the negative equivalents. The distribution of responses is further illustrated graphically in Figure 1.

The results indicate that all categories are being utilised reasonably often, the minimum frequency being for category -2 which was only used 2% of the time. It should be noted though, that this category is part of the negative set of categories that received a lower number of responses than the positive set of categories. It is further evident that some categories are used more frequently, notably the categories 0, -5 and +5, -8 and +8, and -10 and +10, suggesting that they may be serving as cognitive reference points (Rosch 1975). This is more clearly illustrated in Figure 2 which reflects the categories in absolute values. That is, category -1 is combined with category +1, -2 with +2, and so forth. The zero category and blank responses were omitted from this Figure. In addition, the results indicate that there is a trend of higher frequencies towards the extremes of the scale and lower frequencies towards the central point.

Table 4: Frequency of responses to all categories

Category	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
Frequency	923	363	411	298	340	422	285	263	234	249
Percentage Frequency	7.7	3.0	3.4	2.5	2.9	3.5	2.4	2.2	2.0	2.1
Category	+10	+9	+8	+7	+6	+5	+4	+3	+2	+1
Frequency	815	642	942	812	819	1143	628	482	403	269
Percentage Frequency	6.8	5.4	7.9	6.8	6.9	9.6	5.3	4.0	3.4	2.3

Category	0	Blank	Total
Frequency	1056	123	11859
Percentage Frequency	8.9	1.1	100.0

Before testing the second hypothesis, where the 21 point scale is compared to an 11 point scale, the latter scale was tested for uniformity to assess if it was a suitable for comparative purposes. As the results displayed in Table 5 indicate, the distribution of responses to this scale was not sufficiently uniform. Consequently, the second hypothesis could not be tested. Figure 3 illustrates that all categories are not being utilised often enough and that there is a bias towards the midpoint and extremes, with the category "10" accounting for more than 20% of responses.

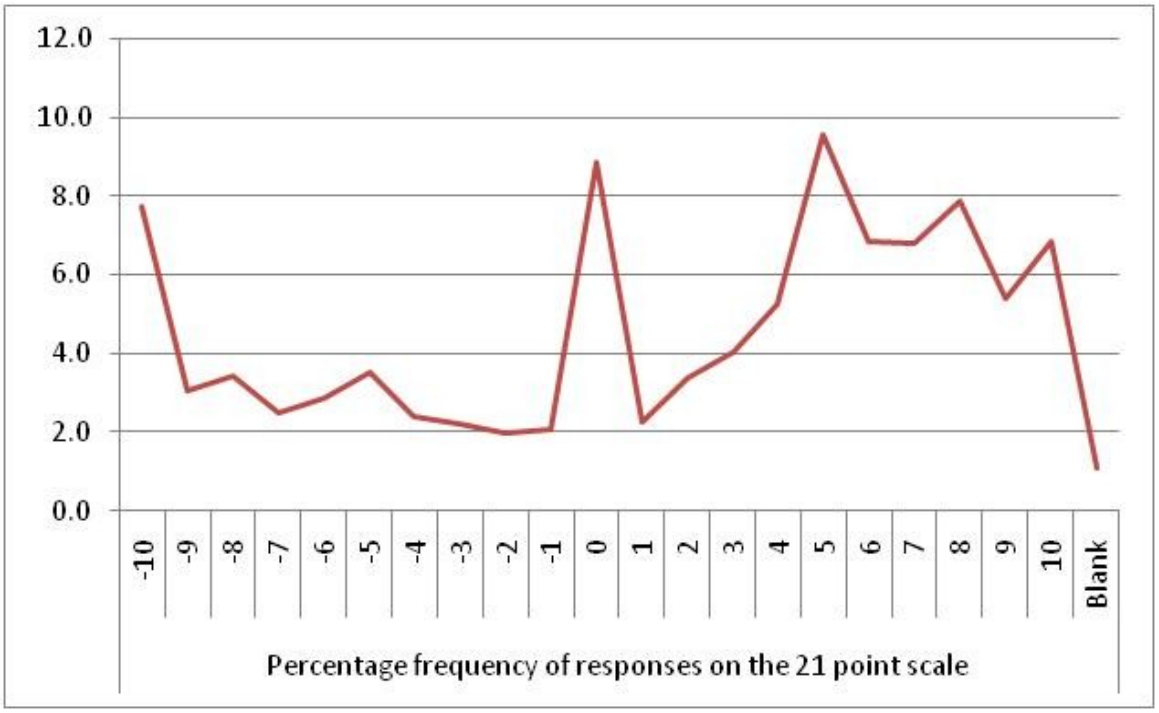

Figure 1: Distribution of percentage frequency responses for the 21 categories

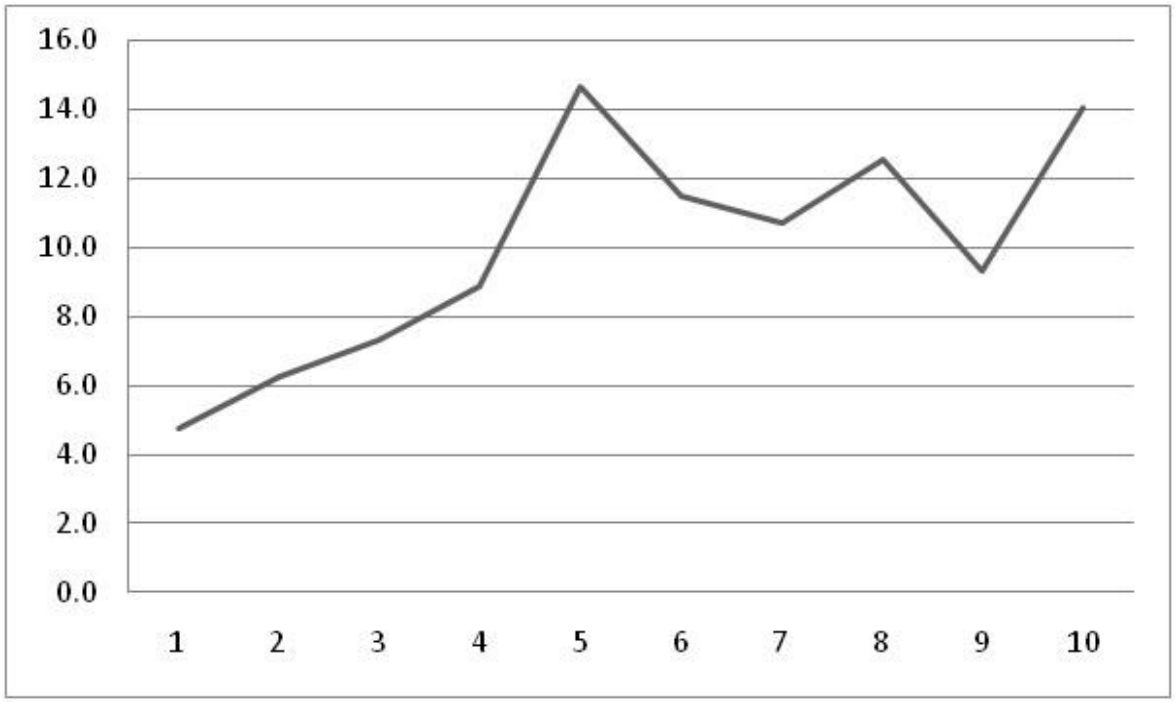

Figure 2: Distribution of percentage frequency responses for absolute value categories

Table 5: Chi square results to test for uniformity of distribution of comparative scale

Degrees of freedom	10
Significance level	5%
Test calculated	38.8542
Critical value	18.31
Decision	As Calculated value is more than the Critical value, the null hypothesis can be rejected

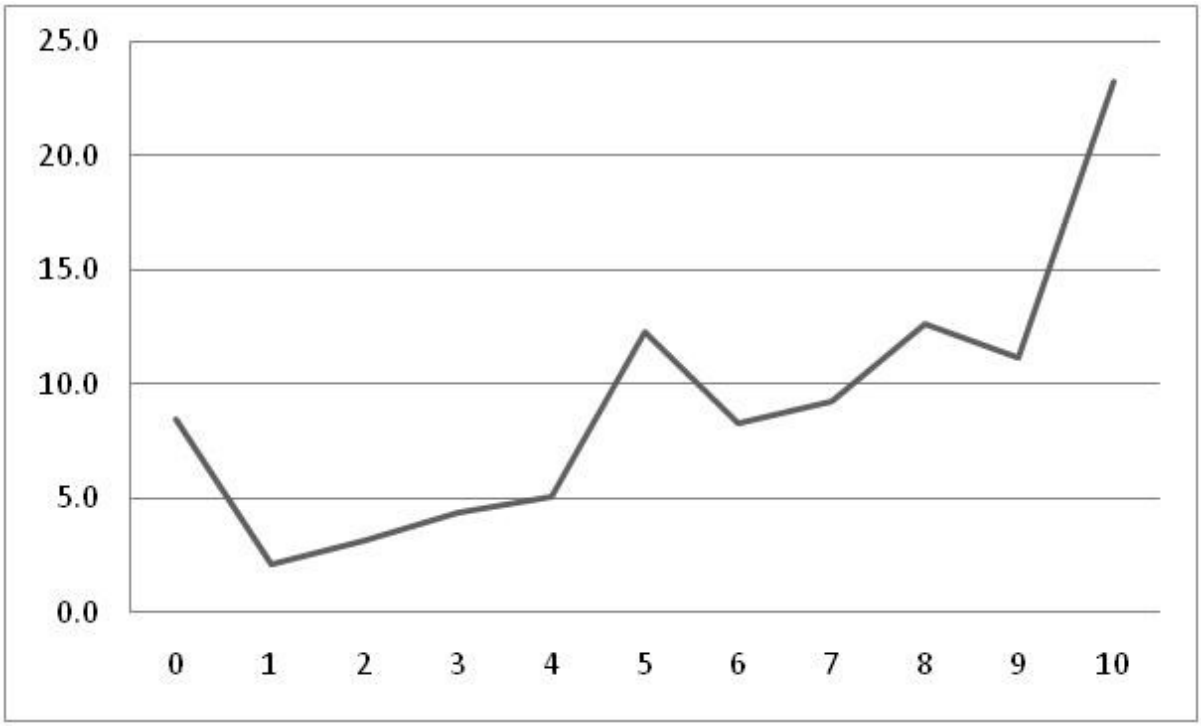

Figure 3: Distribution of percentage frequency responses for the 11 categories of the comparative scale

The third hypothesis compared the variation in responses to the first half of Section 2 to the second half. The first 33 items were compared to the next 33. The last item was omitted from the comparison, given that it was not included as an item in one of the three sets of data that were combined. Table 6 displays the Chi square test results and indicates that the null hypothesis cannot be rejected, suggesting that the distribution of the frequency responses of the two halves are equivalent. Figure 4 provides a graphical illustration of the similarity of the distribution of the two halves.

Table 6: Chi square results of split halves comparison

Degrees of freedom	41
Significance level	5%
Test calculated	3.053594
Critical value	56.94
Decision	As the Calculated value is less than the Critical value, the null hypothesis cannot be rejected

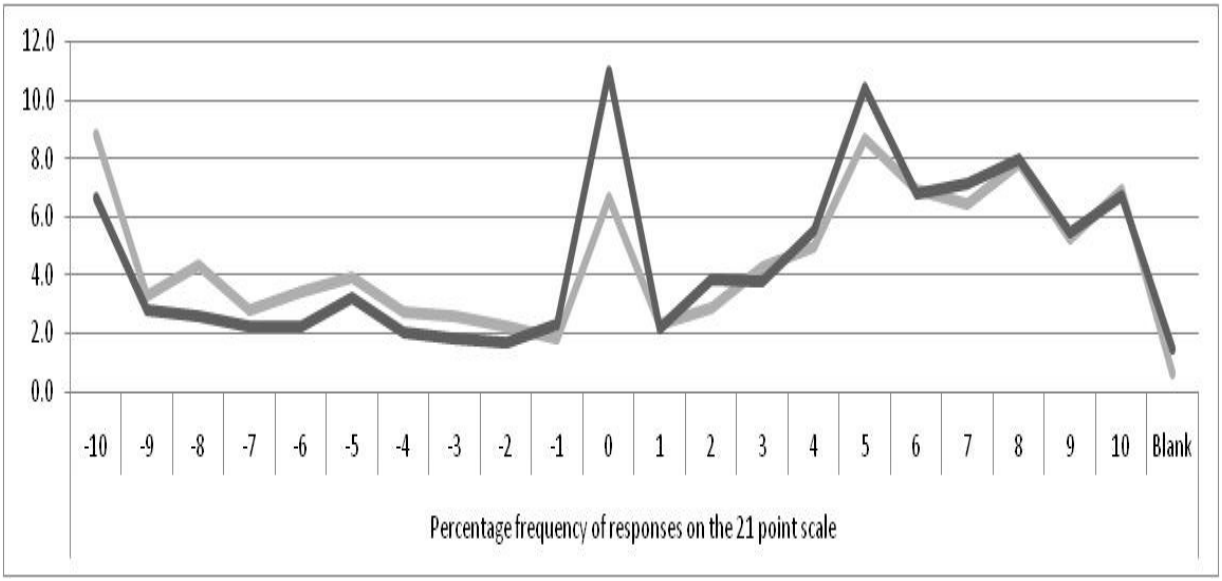

Figure 4: Split half distribution of percentage frequency responses for the 21 categories

9. Discussion and conclusion

Two of the three hypotheses for this study were tested. The first hypothesis concluded that the null hypothesis cannot be rejected. This conclusion implied that the frequency distribution of responses amongst the 21 categories are sufficiently uniform, thereby conforming to the assumptions of Parducci's range-frequency model and confirming the utility of the 21-category scale for respondents. Given that the scale is useful to respondents, it follows that the increased scale granularity should also benefit the researcher in the form of enhanced results in statistical analysis.

Unfortunately the second hypothesis could not be tested as the distribution of the comparative 11-point scale was not uniform. This was unfortunate, especially in the light of the evidence of possible cognitive reference points (Rosch 1975) in the 21-point scale. Further comparative research of scales making use of a different number of response categories is therefore required. A comparison of a 21-point and a 5-point or 7-point Likert scale would be particularly valuable.

In testing the third hypothesis, it was concluded that the null hypothesis cannot be rejected, implying that the two halves of the section of the questionnaire that was being investigated had equivalent distributions. Building on the results of the first hypothesis, this conclusion implies that respondents continued to use the full spectrum of responses throughout the Section and were not acquiescing or becoming "fatigued", in spite of the length of the section (i.e. 67 items).

It can therefore be concluded that the 21-point scale was of value to respondents. Consequently, it also has benefits for the researcher by producing more accurate data that is also more suited to statistical analysis, given its increased variability. The result of this research challenges researchers to give more explicit attention to an aspect of questionnaire design that is often taken for granted, and to consider the merits and demerits of design alternatives, including the type of response options and the appropriate number of categories. In particular, researchers should pay more attention to the scale granularity that they use when designing a questionnaire, rather than simply applying conventional wisdom. Researchers should also consider more carefully their choice of scale in the light of the statistical analysis that is to be conducted.

When piloting a questionnaire, the views of respondents about the range of response options that were provided should be solicited to find out if they were comprehensive enough and exhaustive, or if the level of granularity was too high and therefore some response options were either not meaningful, or were too difficult for respondents to differentiate from other choices. More holistically emotional reactions can also be gauged to determine if respondents were frustrated or satisfied with the level of granularity in relation to the length of the questionnaire. Of course the granularity of these pilot results could also be statistically assessed in the same way that was demonstrated above when testing the first hypothesis.

Finally, further research is required to assess the value of various levels of scale granularity in various types of questionnaires, and their application in various contexts. In particular, a limitation of this study was that it was only carried out in a South African public sector context. Other research suggests that gender, ethnicity and cultural background may affect an individual's responses to question items (e.g. Arce-Ferrer 2006; Arce-Ferrer & Ketherer 2003; Huynh *et al*. 2009; Lee & Soutar 2010). Therefore, exploring responses across a range of levels of scale granularity could make a valuable contribution to cross-cultural research.

References

Almli, V. L., Naes, T., Enderli, G, Sulmont-Rossé, C., Issanchou, S. & Hersleth, M.(2011) "Consumers' acceptance of innovations in traditional cheese. A comparative study in France and Norway". *Appetit*e, Vol 57, No.1, pp 110-120.

Arce-Ferrer, A. (2006) "An Investigation Into the Factors Influencing Extreme-Response Style: Improving Meaning of Translated and Culturally Adapted Rating Scales". *Educational and Psychological Measurement*, Vol 66, No. 3, pp 374-392.

Arce-Ferrer, A., & Ketherer, J. (2003) "The effect of scale tailoring for cross-cultural application on scale reliability and construct validity". *Educational and Psychological Measurement*, Vol 63, pp 484-501.

Bourque, L. B. & Fielder, E. P. (1995) *How to Conduct Self-administered and Mail Surveys*, Sage, Thousand Oaks.

Chaiken, S., & Eagly, A. H. (1983) "Communication Modality as a Determinant of Persuasion: The Role of Communicator Salience". *Journal of Personality and Social Psychology, Vol* 45, No. 2, pp 241-256.

Chami-Castaldi, E., Reynolds, N. & Wallace, J. (2008) "Individualised rating-scale procedure: A means of reducing response style contamination in survey data?" *The Electronic Journal of Business Research Methods*, Vol 6, No. 1, pp 9 – 20.

Churchill, G. A. (1979) "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research*, Vol 16, No. 2, pp 64-73.

Coelho, P.S. & Esteves, S.P. (2006) "The choice between a five-point and a ten-point scale in the framework of customer satisfaction measurement". *International Journal of Market Research*, Vol 49, No. 3, pp 313-339.

Cook, D. A. & Beckman, T.J. (2009) "Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX". *Advances in Health Science Education*, Vol 14, pp 655–664.

Cook, C.F. Heath, R., Thompson, L. & Thompson, B. (2001) "Score reliability in web or internet-based surveys: unnumbered graphic rating scales versus Likert-type scales". *Educational and Psychological Measurement*, Vol 61, pp 697-706.

Cummins, R.A. & Gullone, E. (2000). "Why we should not use 5-point Likert scales: The case for subjective quality of life measurement". *Proceedings, Second International Conference on Quality of Life in Cities* (pp.74-93). Singapore: National University of Singapore.

Dawes, J. (2008) "Do data characteristics change according to the number of scale points used?" *International Journal of Market Research*, Vol 50, No. 1, pp 61-77.

DeVellis, R. F. (2003) *Scale development: Theory and applications* (2nd ed.), Sage Publications, Thousand Oaks.

Dillman, D. A., Smyth, J. D. & Christian, L. M. (2009) *Internet, mail and mixed-mode surveys: The tailored design method,* John Wiley & Sons Inc., Hoboken, N.J.

Fink, A. (1995) *How to ask survey questions,* Sage Publications, Thousand Oaks.

Foddy, W. (1994) *Constructing questions for interviews and questionnaires: Theory and practice in social research,* Cambridge University Press, Cambridge.

Hatton, C., Emerson, E., Rivers, M., Mason, H., Swarbrick, R., Mason, L. Kiernan, C. Reeves, D. & Alborz, A. (2001) "Factors associated with intended staff turnover and job search behaviour in services for people with intellectual disability". *Journal of Intellectual Disability Research*, Vol 45, No. 3, pp 258-270.

Hui, C.,&Triandis, H. (1989) "Effects of culture and response format on extreme response style". *Journal of Cross-cultural Psychology*, Vol 20, pp 296-309.

Huynh, Q., Howell, R. T. & Benet-Martinez, V. (2009) "Reliability of Bidimensional Acculturation Scores : A Meta-Analysis", *Journal of Cross-Cultural Psychology*, Vol 40, No 2, pp 256-274.

Krieg, E.F. (1999) "Biases Induced by Coarse Measurement Scales". *Educational and Psychological Measurement*, Vol 59, pp 749-766.

Kubovy, M. & Psotka, J. (1976) "The predominance of seven and the apparent spontaneity of numerical choices". *Journal of Experimental Psychology: Human Perception and Performance*, Vol 2, pp 291-294.

Lee, J. A. & Soutar, G. (2010) "Is Schwartz's Value Survey an Interval Scale, and Does It Really Matter?" *Journal of Cross-Cultural Psychology,*Vol 41, No 1, pp 76–86.

Lee, Y., Douglas, J. & Chewning, B. (2007) "Techniques for developing health quality of life scales for point of service use". *Social Indicators Research*, Vol 83, pp 331–350.

Likert, R. (1932) "A technique for the measurement of attitudes". *Archives of Psychology*, Vol 22, No. 140, p 55.

Likert, R., Roslow, S. & Murphy, G. (1934) "A simple and reliable method of scoring the Thurstone attitude scales". *The Journal of Social Psychology*, Vol 5, pp 228-238.

Miller, G. A. (1956) "The magical number seven, plus or minus two: Some limits in our capacity for processing information". *Psychological Review*, Vol 63, pp 81-97.

Mobley, W.H., Horner, S.O. & Hollingsworth, A.T. (1978) "An evaluation of precursors of hospital employee turnover". *Journal of Applied Psychology*, Vol 63, No. 4, pp 408-14.

Mobley, W.H., Griffeth, R.W. Hand, H.H. & Meglino, B.M. (1979) "Review and conceptual analysis of the employee turnover process". *Psychological Bulletin*, Vol 86, No. 3, pp 493 – 522.
Mobley, W.H. (1982a) *Employee turnover: causes, consequences and control*, Addison-Wesley Publishing Company, INC., Philippines.
Mobley, W.H. (1982b) "Some unanswered questions in turnover and withdrawal research". *Academy of Management Review*, Vol 7, No. 1, pp 111-116.
Molenaar, N. J. (1982) "Response effects of 'formal' characteristics of questions". In W. Dijkstra and J. Van der Zouwen (eds), *Response behaviour and the survey interview*, Academic Press, New York.
Oppenheim, A. N. (1966) *Questionnaire design and attitude measurement*, Heinemann, New York.
Pedhazur, E. J., & Schmelkin, L. P. (1991) *Measurement, design, and analysis: An integrated Approach,* Lawrence Erlbaum, Hillsdale, NJ.
Pejtersen, J.H., Bjorner, J.B. & Hasle, P. (2010) "Determining minimally important score differences in scales of the Copenhagen Psychosocial Questionnaire". *Scandinavian Journal of Public Health*, Vol 38, No 3 (Suppl), pp 33-41.
Rattray, J. & Jones, M.C. (2007) "Essential elements of questionnaire design and development". *Journal of Clinical Nursing*, Vol 16, pp 234–243.
Remenyi, D., Onofrei, G. & English, J. (2010) *An introduction to statistics using Micorsoft Excel*, Academic Publishing Limited, Reading.
Riggs, C. J. and Rantz, M. J. (2001) "A Model of staff support to improve retention in Long-Term Care". *Nursing Administration Quarterly*, Vol 19, No. 4, pp 43 - 54.
Roberts, J. S., Laughlin, J. E. & Wedell, D. H. (1999) "Validity Issues in the Likert and Thurstone Approaches to Attitude Measurement". *Emotional and Psychological Measurement*, Vol 59, No. 2, pp 211-233.
Rodebaugh, T. L., Woods, C.M. & Heimberg, R.G. (2007) "The reverse of social anxiety is not always the opposite: The reverse –scored items of the Social Interaction Anxiety Scale do not belong". *Behavior Therapy*, Vol 38, No. 2, pp 192-206.
Rosch, E. (1975) "Cognitive reference points". *Cognitive Psychology*, Vol 7, pp 532-547.
Schuman, H. & Presser, S. (1996) *Questions & answers in attitude surveys: Experiments on question form, wording, and context*, Sage Publications, Thousand Oaks.
Smithson, M. (2006) "Scale construction from a decisional viewpoint". *Minds & Machines*, Vol 16, pp 339–364.
Symonds, P. M. (1924) "On the loss of reliability in ratings due to coarseness of the scale". *Journal of Experimental Psychology*, Vol 7, pp 456-461.
Terre Blanche, M. & Durrheim, K. (1999) *Research in Practice: Applied Methods for the Social Sciences,* University of Cape Town Press, Cape Town.
Tourangeau, R. Rips, L. J. & Rasinski, K. (2000) *The Psychology of Survey Response,* Cambridge University Press, Cambridge.
Wong N., Rindfleisch, A. & Burroughs, J. E. (2003) "Do Reverse-Worded Items Confound Measures in Cross-Cultural Consumer Research? The Case of the Material Values Scale." J*ournal of Consumer Research*, Vol. 30, No. 1, pp 72-91.