J Soc Sci, 43(2): 143-151 (2015)
PRINT: ISSN 0971-8923 ONLINE: ISSN 2456-6756
DOI: 10.31901/24566756.2015/43.02.06

Constructing Homogeneous Likert-Type Summative Rating Scales According to Classical Measurement Theory

Marianne Viljoen

School of Higher Education Studies, Faculty of Education University of the Free State, Bloemfontein, South Africa 9301
Telephone: +27 848007877, E-mail: viljoenmc@ufs.ac.za

KEYWORDS Instrument. Measuring. Questionnaire. Reliability. Validity. Analyse

ABSTRACT The study of human behaviour requires the use of valid and reliable instruments in order to quantify, analyse and interpret such behaviour. It is an accepted fact in research design that the reliability and validity of quantitative research depends on the reliability and validity of the instruments applied. Many researchers are hampered in their efforts to design quantitative research due to a lack of measuring instruments. Knowing the requirements for constructing Likert-type summative rating scales will greatly increase the reliability and internal validity of research projects. This paper offers the guideline for scale construction according to the model of Classical Measurement Theory as presented by Nunnally in the 1970s. This method of scale construction requires homogeneity of scale dimensions and applies item analysis as a measure of standardisation and not factor analysis as is the case with heterogeneous scales. In executing quantitative research, the statistical properties of scales should be scrutinised for their applicability. Researchers should take care when selecting scales in quantitative research.and need experience in the methodological guidelines of scale construction. The writing of an item pool, the pros and cons of different numbers of scale steps and types of anchors are discussed with relevance to current literature. Determining the reliability and validity of a rating scale is clearly outlined. All applicable references regarding the elements of scale construction as posited by literature from the 1970s to the present (researching a time span of 40 years) have been consulted recounting the status of the Nunnally principles as they stand today.

INTRODUCTION

This paper will argue the established validity and reliability of constructing a Likert-type scale by way of the 1970's Classical Measurement Theory as presented in the Nunnally book on research methods, of that time. This method, although relatively simple, has been researched with regards to scale properties and statistical requirements and has, even in the computer age, been found to be sound and user friendly.

All researchers strive to increase the reliability of rating scales. Measuring instruments are fundamental to the implementation of solid quantitative research in education and the social sciences. As scientists study human behaviour they increasingly need valid and reliable instruments in an attempt to quantify, analyse and interpret such behaviour. Renowned authors, Kerlinger (1986) and Maas (1998) still quoted today, unambiguously state that it is a clearly established fact that the reliability and validity of quantitative research is tied in closely with the reliability and validity of the instruments applied. This fact presents a special problem to education researchers and supervisors who need to design quantitative research projects by finding valid and reliable measuring instruments which measure human attributes of concern to planned studies. In addition it is necessary that the construction characteristics of existing scales be examined for the appropriateness of use in intended research and to ascertain their applicability. The basic premise in quantitative research design is that a scale should only be used for measurement in a sample of which the hypothesised population is similar to the population for which the original scale was designed; otherwise it should be adapted or a more applicable one written. Therefore it becomes essential that researchers using quantitative designs scrutinise all measuring instruments before they are applied to test groups and in order to do so researchers need at least minimal proficiency in scale construction, which is provided by this paper.

The range of standardised instruments available for measuring human behaviour in the study field of education and social sciences can be broadly categorised into three main classes, namely the measurement of achievement and aptitude, the measurement of personality traits and the measurement of sentiments (Nunnally 1970). Included under *sentiments* is the measure-

ment of attitudes. All of the above measurements, except the measurements of attitudes, require intensive expertise and the knowledge of highly specialised statistical procedures. Consequently they are designed over long periods of time requiring several pilot studies. Factor analysis requires that the scale be administered to large samples whereas attitude scales can be standardised with a smaller sample.

Attitudinal rating scales are easy to design and are widely applicable provided that the principles of homogeneity of questions are adhered to. According to Kerlinger (1986: 494) "A rating scale is a measuring instrument that requires the rater or observer to assign the rated object to categories or continua that have numerals assigned to them." Arnold et al. (2007) agree that Likert-type attitude scales are "One of the most popular methods of measuring attitudes..."

The first rating scale was constructed by Likert in 1932 (Arnold et al. 2007) and consequently scales with scale steps are referred to as Likert-type scales. The attribute of homogeneity is a necessary prerequisite of this form of scale construction. Homogeneity, as discussed further on, implies that there should be a positive relationship between the sub dimensions of the scale and that questions should be related. Therefore a person answering positively on one dimension of the scale will be inclined to answer positively on other dimensions as is the case in attitude scales. Gliem and Gliem (2003: 82) describe the applicability of Likert-type scales: "Oftentimes information gathered in the social sciences, marketing, medicine, and business, relative to attitudes, emotions, opinions, personalities and descriptions of people's environment involves the use of Likert-type scales".

Regarding the appropriateness of Likert-type summative rating scales to measure attitudes it is necessary to state that the term attitude by no means inhibits the range of applicability of such scales in research. Reflecting on the term *attitude* as defined by various resources, it becomes apparent that the application of attitude scales to measure human behaviour are extremely varied. Coleman (2001: 63) defines an attitude as "...a more or less consistent pattern of affective, cognitive or conative, and behavioural responses (or of feeling, thinking, and behaving) towards a psychological object,...".Kerlinger (1986: 495) offers a similar definition wherein he states that an attitude "...is an organised predisposition to think, feel, perceive, and behave toward a referent or cognitive object". Applying this construct to a research paradigm, it is clear that attitude scales may be used to measure attitudes towards an array of different attitudinal objects in various fields such as the social and business sciences, and education. Apart from constructs such asacademic achievement and People Living with HIV and AIDS, some innate personal qualities such as motivation, self-efficacy and time perspective of students can also be measured with attitude rating scales.

This paper will present a method for constructing and standardising homogeneous Likert-type attitude rating scales by means of item analysis which was designed by Nunnally and which is still highly relevant and applicable today. This method, endorsed by Nunnally in 1970 and 1978 is referred to as Classical Measurement Theory (Nunnally and Bernstein 1999) and the validity of all the steps in scale construction will be argued against the recommendations of modern day researchers.

The guidelines that will be discussed are the following: Defining the attitudinal construct, writing of an item pool, types of anchors, selecting the number of scale steps, piloting of the scale as well as reliability and validity of the scale.

CONDITIONS FOR CONSTRUCTING A LIKERT-TYPE SUMMATIVE ATTITUDE RATING SCALE

Many researchers shy away from developing measuring instruments because of the sophisticated and laborious task of standardising such instruments by means of factor analysis. Nunnally, however clearly indicates that a scale need not be standardised by means of a factor analysis provided that the scale is constructed from an item pool which is homogeneous with regard to the object being measured. " In the construction of most types of psychological measures, factor analysis of an item pool should be considered only as a last resort, after efforts to hypothesise homogeneous scales have led to naught" (1970: 437).

The condition, under which a less experienced person may construct a summative rating scale for measuring attitudes, is that the scale must consist of homogeneous items. Homogeneous items are also called uni-dimensional items

and such a test is standardised by doing an item analysis (La Trobe and Acott 2000) instead of a factor analysis, which is more complicated. Gliem and Gliem (2003) describe homogeneity by stating that an underlying quantitative dimension should connect the items in a scale. If for instance you are writing a scale to measure the psycho social background of students you could have sections that measure socio economic status, social support and level of depression. Each of these topics would measure different dimensions of psycho social background but according to the theory, these aspects all form part of the dimensions of psycho social backgrounds and are related.

It is also necessary to consider that Classical Measurement Theory (CMT) differs from Item Response Theory (IRT) which underpins latent trait scales. Polit and Beck (2008: 476) state "Whereas the items on a CMT scale are designed to be similar to each other to tap the underlying construct in a comparable manner, items on a latent trait (IRT) Likert scale are carefully chosen and refined to tap different degrees of the attribute being measured." Items written with IRT as basis lie on a continuum indicating varying degrees of difficulty in choosing a rating. In lay terms this means that item number 5 in the scale should be emotionally more difficult to rate than item number 1. This method of scale construction is also referred to as the Rasch model (Polit and Beck 2008) which is applied to measure uni-dimensionality and item fit (Chachamovich et al. 2008)

PROCEDURE FOR WRITING OF A LIKERT-TYPE SUMMATIVE RATING SCALE ACCORDING TO CLASSICAL MEASUREMENT THEORY

The following summary contains a description of how to go about writing a rating scale. Each step will be described in more detail further on.

- Define the attitudinal object or construct, and if applicable, it's dimensions which are to be measured. Have a specific population in mind.
- Write an item pool (begin with 40 item statementsand later reduce to 20).
- Determine the number of scale steps you will use.
- Choose the anchors.
- Decide on the length of the scale.
- Pilot the scale.
- Calculate reliability.
- Argue validity.

Writing of an Item Pool

Summative rating scales for measuring verbalised attitudes consist of a number of item statements (begin with 40 and reduce to 20) from the content domain, with which respondents may agree or disagree on a six-or-more point scale. The scores of a respondent are summed to give a total which then indicates his/her stance on the measure. Neutral or extreme statements should not be used as this creates less variance (Nunnally 1970). When scoring the negative statements, scores are inverted to give the "true" score. The use of both positive and negative statements is an attempt to avoid response style as a confounding variable.

Homogeneous attitude scales are developed from an underlying theory or hypothesis which is also called the content domain of the variable. Therefore it is essential that before writing statements for the item pool, a well-articulated definition of the variable under investigation should be given (Albaum 1997; Dawis 1987). Abeles (1987) agrees that items should be based on preconceived structures. Ambiguously understood constructs will certainly lead to invalid test scores (Polit and Beck 2008).

Collecting possible item statements can be done by employing qualitative techniques where respondents have the opportunity to voice their feelings and experiences regarding the attitudinal object. A team of professionals may work together in formulating statements. A technique which is advised by Polit and Beck (2008: 482) involve sending the item pool to a number of professionals in the field and having them rate each item on its relevance to the construct being measured. Dawis (1987) suggests that statements should be back-translated by experts in the field who know the theory but were not involved in the development of the scale. In so doing it is possible to affirm the relevance of the items to the theory. This also serves to establish coverage of the content domain which determines validity

The mother-tongue of the respondents is an important consideration when formulating the statements: When statements are written in the second language of the respondents, special

attention should be paid to the level of comprehension of the statements. Turner (1993) stresses this aspect when stating that language limitations affect the accuracy of the responses on scale items and inhibits the reflection of a true attitude. Dawis (1987) calls this aspect "readability" and emphasises that often it is better to use lay language than more sophisticated academic language. Polit and Beck (2008) recommends that statements be clear, short and should not contain jargon. Furthermore they advise against double negatives and double barrelled item statements.

Another aspect of scale construction that may confound researchers is the aspect of positive and negative scale item statements where reversing of scores are employed when summating the scores over all the items. Nunnally (1970) suggests that the scale should consist of equal numbers of moderately positive and moderately negative statements. He posits that this may minimize response bias or acquiescence as Polit and Beck (2008) call it. There is however another phenomenon to consider namely that second language speakers or inexperienced respondents could become confused when confronted with both positive and negative items in the same scale. Polit and Beck (2008) thus distinctly advise the use of only positively worded item statements and declare that there is modern day evidence to support this theory.

Number of Scale Steps

The number of scale steps that is appropriate for reliable and valid research has been a bone of contention among scale designers and researchers since 1932 when Likert wrote the first summative rating scale. The original model contained only five steps;

- Highly agree
- Agree
- Neither agree nor disagree
- Disagree
- Highly disagree

The debate surrounding the number of scale steps, centre around threeissues; how the number of scale steps affect reliability; whether parametric or non-parametric statistics may be applied to data gathered by such a scale; and whether the number of scale steps should consist of an even or uneven number.

Nunnally (1970: 425) cites Guilford (1954) who said that according to psychometric theory it has been demonstrated that "...the reliability of individual rating scales is a monotonically increasing function of the number of steps". This increase in reliability is said to level off when seven to eleven scale steps are used. When implementing a scale with too many steps the researcher has to consider the possibility that respondent fatigue may set in for a respondent having to gauge his perceptions, feelings or thoughts about an object on so many levels.

With a higher number of scale steps, equidistance of scale steps can be assumed. And this translates to higher reliability since it offers the respondents the opportunity to reflect true feelings, cognitions and perceptions in their choice between a larger number of alternatives.

Another consideration when deciding on the number of scale steps to use in an attitude scale is that a total score on a five point scale is considered to be an ordinal measurement which is analysed by non-parametric statistical analysis (Maas 1998). Six and more scale steps represent an interval scale which may be analysed with parametric statistics. According to Turner (1993), parametric statistics are more powerful and robust than non-parametric statistics. Polit and Beck (2008) posit that variability is enhanced when opting for "numerous response options". Greater variability would indicate greater variance which has statistical value in that the distribution of test scores would then approximate the normal curve and facilitate the use of parametric techniques.

Bernstein (1996: 1) who describes himself as having much experience in the practical use of Likert-scales stresses the issue of the heuristic value of using Likert scales as interval measures. With an interval scale one is able to "...quantify degrees of affect..." of test respondents.

In opposition to these statements comes the slant of modern day researchers who, according to Heiser (2007) are adamant that equidistance of adjacent points on a Likert scale cannot be claimed. According to them a Likert scale with any number of scale points is merely an ordinal scale and should thus be analysed with non-parametric statistics. However the debate continues. Heiser (2007) declares that the issue is not whether intervals are equal but whether the steps are close enough to be treated as equal which implies the use of parametric statistics in

analyses and drawing the right conclusions from hypotheses. To this question the author argues that one ought to do both parametric and non-parametric statistical analyses on such data.

The third issue, whether to use an uneven or even number of scale steps, has also been debated much. An odd number of steps make it possible for respondents to fall into the response style of choosing the middle or neutral value. Regarding this Nunnally (1970) states that there is a danger of uneven numbers of scale steps eliciting a response style when respondents have the option to choose a "neither agree nor disagree" scale step. Often this is an easy way when respondents are tired, disinterested or not willing to commit to an opinion. On the other hand he states that some researchers believe that a "neutral option" makes subjects more "comfortable" in making ratings. Modern researchers concur with this point of view (Polit and Beck 2008) regarding the applicability of scale steps with uneven numbers as more advantages.

In a research study done by Birkett (1986) 300 randomly selected respondents completed a Health Locus of Control scale with either two, six or fourteen response categories. According to the results reliability was highest when the questionnaire had six response categories. A study done by Preston and Colman (2000) tested scales that only differed in the number of response categories. The indices of reliability, validity and discriminating power were significantly higher for scales with more response options.

Notwithstanding these arguments, several modern authors advocate the use of only a five-point scale (Huysaman 2001; Delport 2005; McMillan and Schumacher 2006). Nevertheless Nunnally's (1970) arguments should be heeded.

Types of Anchors

Rating scales usually have numeric as well as semantic anchors. There are a number of different word pairs that may be used to rate attitudes. These are the scale steps, that is, 'Completely agree', 'Mostly agree' and so on. In actual fact the scale steps are numbers which are represented by word pairs. Siegle (n.d.) gives a very useful summary of the most used word pairs:

Agreement: *Strongly disagree, Mostly disagree Slightly disagree, Slightly agree, Mostly agree, Strongly agree.*

Frequency: *Always, Usually, Seldom, Never.*

Importance: *Very important, Important, Moderately important, of little importance, Unimportant.*

Quality: *Good, Acceptable, Poor*

Likelihood: *Definitely, Probably, Possibly, Very probably not.*

Other options for anchoring that is supplied by Nunnally (1970) are percentages and adjective pairs. Adjective pairs such as 'Ineffective-Effective', 'Foolish-Wise' and 'Weak-Strong' form a different type of summative rating scale known as a semantic-differential scale which in essence has all the characteristics of a Likert-type summative rating scale and is easy to construct.

Regarding the types of word pairs that form anchors, care should be taken with the precise meaning respondents attach to some words, especially if the scale is not written in their first language. The exact difference in meaning between 'Occasionally' and 'Rarely' and 'Somewhat' and 'Little' is not always clear to respondents. It is the author's opinion that the anchor pairs 'Agree/Disagree' and the range of scale steps that accompany them are more precise and easier to use because they cannot be misunderstood. Likewise the adjective pairs of the semantic differential scale are concise and easily understood.

Length of the Scale

Nunnally (1970) advises that ascale include approximately 20 items. Polit and Beck (2008) state that variability of the scale, is enhanced by including a larger number of items, but they do not specify an exact number. The concern should be to strike a balance between the numbers of steps that would ensure optimal variance as opposed to a scale that creates response fatigue.

Pilot Test the Scale

The piloting of the scale is part of determining the reliability of the scale. Although the final scale should have approximately 20 items, begin with 40 items:

- Administer the scale to a group of subjects of which the size is at least five times the number of items in the item pool. The more subjects, the better.
- Mark the test and reverse the scores if negative items are present.

- Sum each item for all respondents to obtain its total scores.
- Correlate each item's total with the total score of the group using a Coefficient Alpha.
- The item with the highest correlation coefficient becomes the gauge against which the other item totals are measured. It is called the determining item.
- Calculate the correlation of each item's total score with that of the determining item's total score.
- Select the test items in the test with the highest correlation with that of the determining item.
- Pilot test the shortened version of the test again and calculate the Cronbach's Alpha statistic. It should ideally be above 0.70 (Embi 2007 in Pallant 2002).

VALIDITY

Being able to report on the reliability and validity of a scale is intrinsic to the whole process of scale construction. Content validity is assumed when scale items are grounded in theory, provided that scale items were taken from "...the universe of items related to the construct being measured" (Maas 1998: 11). This is possible when a scale is constructed by professionals who know the content domain of the intended test. When researchers agree that the construct being measured is grounded in theory, they may assume construct validity.

The Cronbach's Alpha of the final scale as described above is also the reliability coefficient and should be in the order of 0.7 and above. There is an option to pilot the scale directly and analyse the Cronbach's Alpha with deleted variables. This table indicates the effect of each item on the total alpha coefficient by deleting it and reporting the alpha value without that item in the scale. The effect of the item on the total reliability is then indicated. Items, which when deleted cause the alpha value to rise, indicate that they have a negative effect on the scale and should be deleted. In this manner only scale items which are highly reliable are kept in the scale.

CONFOUNDING VARIABLES OF LIKERT-TYPE RATING SCALES

The following confounding variables need to be controlled by the researcher;

Response Styles

The most obvious of the measurement confounders in summative scales are response styles. There are three typical response styles that may be biased. Firstly respondents who continually choose the neutral option may be indicating fatigue, lack of motivation or lack of cooperation. Conversely there may truly be respondents with no opinion regarding the object being rated. To counter this bias, Nunnally (1978) suggests using an even number of scale items.

The second response style bias may come in the form of respondents being overly positive regarding the attitudinal object or the test administrator. This is similar to the confounding variable known as the halo effect. Such may be the case when respondents favour the researcher and wish to impress or gain favour from the researcher (Kerlinger 1986; Turner 1993). One of the ways in which to counter this reaction is to keep the inventory anonymous. Of course anonymity prevents different scores of the same respondent being recorded which in some cases defy the objective of the research.

Thirdly, the response style bias of being overly negative may be found among respondents who feel negative towards the test situation or who are overly discontented in mood. In this case they will mark the statements more negatively then they really feel towards the attitudinal object.

Test-retest Reliability and Time Interval

In assessing the test-retest reliability of a summative rating scale, the decision of the time interval between first and second testing is a challenging aspect to consider. Huysamen (1976) warns against the effects of test sensitization. It takes a certain time to elapse before subjects forget the questions and answers they gave to the test. If they remember the questionnaire and give the same answers, a false test-retest reliability is achieved.

In a study done by Inandi et al. (2001) a scale was developed to determine perceptions about physicians. The retest was done with a 7-10 day interval on the same group of 220 respondents Clearly this time period is too short and allows for test sensitisation to confound the Pearson correlation of 0.61 ($p<0.001$) that was calculated. Challenging this view one may argue that if the time interval between test and re-test is too long,

attitudes might have actually changed and one would then obtain a non-significant correlation when the test is actually reliable.

Direction and Intensity

In Likert-type scales respondents have to access two emotions or perceptions in themselves simultaneously, namely *direction* (feeling positive or negative towards an object) and *intensity* of the feeling (choosing the matching scale step). One may assume that respondents may experience confusion regarding these two aspects.

Albaum (1997) addresses this issue when he contends that a Likert-type scale confounds direction of attitude with the intensity dimension of the attitude. In order to deal with this problem he devised a scale that measured direction in the first stage and intensity in a second stage. The problem lies within the phenomenon of central tendency and the reluctance of respondents to give extreme scores. A second stage questionnaire circumvents this response and respondents then feel free to open up to their responses "Perhaps when faced with the standard one-stage format respondents are reluctant to express an extreme position even though they have it. The two stage test would give them the flexibility to express their "true opinions" (Albaum 1997: 331). The questionnaires used in this research used a 4-point Likert scale with one alternative being "no opinion". It is the authors contention that if a six or eight point scale is used the direction of respondents as well as intensity of feelingwill be elicited. Babbie and Mouton (2001) agree that by adding scale steps for a respondent to choose from, you not only give opportunity to measure direction but intensity of response as well.

Interpretation of a Likert-type Summative Rating Scale

It should be noted that a Likert scale is a self-report inventory, measuring only that which the respondent is willing to disclose regarding his/her feelings, thoughts and perceptions about an object, person or construct. Making scales anonymous is one way to counteract this response bias. However accuracy in research remains the responsibility of the researcher and this implies that scales should not be used randomly but be examined concerning their appropriateness for use with applicable research samples.

ADVANTAGES OF A LIKERT-TYPE SUMMATIVE RATING SCALE

Likert-type scales are easy to construct and when one knows the theory in which the content domain is imbedded, it is relatively easy to construct a scale with high internal consistency. Barcley and Weaver (1962) conducted a study indicating that the Likert-type scale as opposed to a Thurston scale is relatively easier to construct, takes less time, is more reliable and there is no need to find judges since any professional in the field may be used as item judge.

CONCLUSION

This paper illustrates why the steps in the design and construction of a homogenous summative rating scale based on item analysis which is captured under Classical Measurement Theory are still relevant. The establishing of reliability and validity, the possible confounding variables in application of Likert-type scales and the advantages of using the Likert-type scale have been discussed.

This form of scale construction is limited to scales of which the items are homogeneous or dependent. This limitation however does not detract from the applicability of these types of scales and many examples of such scales for use in the field of education and social sciences are offered. The recommendation that the reliability is done with item analysis instead of factor analysis is to the advantage of the novice researcher since factor analyses is a laborious task requiring several pilot studies, larger samples which are not always accessible.

This method of scale construction is especially valuable to academics who are not experts in the use of computer data analyses software because reliability can be calculated by means of an excellspreadsheet if need be. Alternatively the cost of having a Cronbach's Alpha calculated by a data-analyst using computer software which is necessary for this type of scale, would be low.

The information supplied in this paper will increase research capacity. It offers quantitative researchers the opportunity to design and con-

struct their own scales when standardised scales cannot be found in literature.

Furthermore, for non-quantitative researchers who need to be consumers of quantitative research, the knowledge gained from this paper could be applied in assessing the reliability and validity of existing scales. Data gathered by means of scales and variables thus quantified, could be betterevaluated. It is thus another method to be used by the quantitative researcher to increase the reliability of self-constructed measuring instruments.

This paper indicates that researchers should ensure that the measuring instruments or scales that they use in research, comply with the specifications for scale construction as outlined by Nunnally in 1970. Current research papers and books do not describe scale construction according to Classical Measurement Theory, because the theory is regarded as subject specific common knowledge.

RECOMMENDATIONS

Many researchers are in need of scales for doing quantitative research. Reliable scales in certain fields are not available on databases. Keeping in mind that this method is designed for homogenous scale construction, researchers would gain much by following these principles. This would especially increase the number of quantitative research in the fields of the humanities.

REFERENCES

Abeles HF 1973. A facet-factorial approach to the construction of rating scales to measure complex behaviours. *Journal of Educational Measurement*, 10(2): 145–151.

Albaum G 1997. The Likert scale revisited: An alternate version (product preference testing). *Journal of the Market Research Society*, 39(i2): 331-343.

Arnold WE, McCroskey JC, Prichard SVO 2007. The Likert–Type Scale. From <http: //www. jamescmccroskey. com/publications/25.htm.> (Retrieved on 17 May 2010).

Babbie ED, Mouton J 2001. *The Practice of Social Research*. Oxford: University Press.

Barkley JE, Weaver HB 1962. Comparative reliabilities and ease of construction of Thurstone and Likert Attitude Scales.*The Journal of Social Psychology*, 58: 109-120.

Bernstein IH 1996. Likert Scale Analysis. From <http: //www.math.yorka.ca/Who/Faculty/Monette/Ed-stat/0361.html> (Retrieved on 7 June 2010).

Birkett NJ 1986. Selecting the Number of Response Categories for a Likert-Type Scale. From <www. amstat.org/sections/srms/Proceedings/papers/1986 - 091.pdf-Similar pages> (Retrieved on 12 June 2010).

Chachamovich E, Fleck MP, Trentini CM, Laidlaw K, Power MJ 2008. Development and validation of the Brazilian version of the Attitudes to Aging Questionnaire (AAQ): An example of merging classical psychometric theory and the Rasch measurement model. *Health and Quality of Life Outcome*, 6(5): 1-10.

Coleman AM 2001. *Dictionary of Psychology*. Oxford: University Press.

Dawis RV 1987. Scale construction. *Journal of Counselling Psychology*, 34(4): 481-489.

Delport CSL 2005. Quantitative data collection methods. In: AS De Vos, H Strydom, CB Fouche, CSL Delport (Eds.): *Research at Grass Roots: For the Social Sciences and Human Services Professions*. Pretoria: Van Schaik, pp. 165-196.

Embi R 2007. *Computer Anxiety and Computer Self-efficacy Among Accounting Educators at Universiti Teknologi Mara (UITM)*. PhD Thesis, Unpublished. Malaysia: Malaysia State University.

Gliem JA, Gliem RR 2003. Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert–Type Scales. *Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, 82-88.

Heiser DA 2007. Ordinal, Nominal and Likert Scale Variables. From <http://www.daheiser.info/excel/notes/NOTE%20J.pdf> (Retrieved on 18 March 2013).

Huysamen GK 1976. *Inferensiële Statistiek en Navorsingsontwerp*. Pretoria: Academica.

Huysamen GK 2001. *Methodology for the Social and Behavioural Sciences*. Oxford: University Press.

Inandi TI, Sahun N, Guraksin A 2002. A preliminary study for the development of a scale to assess perceptions about physicians. *Journal of Evaluation in Clinical Practice*, 8(1): 71-75.

Kerlinger FN 1986. *Foundations of Behavioural Research*. 3rd Edition. Orlando: Holt, Rinehart and Winston, INC.

La Trobe HL, Acott TG 2000. A modified NEP/DSP Environmental Attitude Scale. *The Journal of Environmental Education*, 32(1): 12-20.

Maas F 1998. *Research Methods and Statistics in Occupational Therapy*. Department of Occupational Therapy. Queensland: The University of Queensland.

McMillan JH, Schumacher S 2006.*Researching Education: Evidence Based Inquiry*. 6th Edition. New York: Pearson.

Nunnally J C 1970. *Introduction to Psychological Measurement*. New York: McGraw Hill.

Nunnally JC 1978. *Psychometric Theory*. New York: McGraw Hill.

Nunnally JC, Bernstein IH 1999. Psychometric Theory. *Journal of Psycho Educational Assessment*, 17: 275-280.

Obermiller C, Spangenberg ER 1998. Development of a scale to measure consumer scepticism toward advertising. *Journal of Consumer Psychology*, 7(2): 159–186.

Polit DF, Beck CT 2008. *Nursing Research- Generating and Assessing Evidence for Nursing Practice.* 8th Edition. New York: Wolters Kluwer.

Preston CC, Colman AM 2000. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica,* 104(1): 1-15.

Siegle D Beginning Steps in Developing an Attitude Instruments. a. From <http: //www. gifted.uconn.edu/ Siegle/research/Instrument% 20 Reliability % 20and%Va> (Retrieved on 20 March 2012).

Turner J 1993. Using Likert scales in L2 research. *TESOL Quarterly,* 27(4): 736–739.