# Muddles in Pentatonic Likert-type scale: Accuracy Cost in Psychometric Measurements for Small Enterprise Development

Francis Okumu Omillo



**Abstract**

Likert-type scale is ordinal, hence not compatible with parametric techniques. Disregard of this fact causes flawed research outputs. Enterprises get themselves in precarious situations as ultimate consumers flawed outputs. This paper is motivated by the dearth desire by entrepreneurs to make accurate and valid decisions harvested from a dependable measurement scale. Identifying the pitfalls of Likert-type scale and remedies to address the weaknesses, form the objectives of the study. The study is anchored on the Classical Test and Generation theories. Reviewing literature and from own personal experiences in assessing students' thesis at university level in Kenya found traditional pentatonic Likert-type scale highly favored by most young researchers in enterprise development. The researchers treated the Likert scale outputs as interval data. Consequently most of them got wrong inferential techniques and findings. This study suggests transformation of ordinal data into binary data, interval or ratio before going into parametric analysis. Secondly, increase the number of points on the Likert scale, preferably to seven (7) to enhance reliability, validity, discriminating power and respondent preferences. Thirdly, adopt newest models of Likert type scale, that is; novel fuzzy Likert scale, phrase completion scale and two-stages Likert scale for measuring direction and intensity dimensions seperately. Finally, Likert

ILIRIA International Review – Vol 9, No 1 (2019)

type scale could be improved by Rasch analysis, too. The findings and suggestions of the study are relevant for researchers in both academic, clinical and enterprise development for attainment of the *Kenya Vision 2030*.

**Key words:** Likert type scale; psychometric measurements; small enterprise development;

## 1. Introduction

Micro and Small Enterprises (MSEs) are economic activities in both informal and formal sectors employing between 1 and 50 people. The focus on small enterprise development worldwide has been informed by the fact that Micro and Small Enterprises (MSEs) form the backbone and engine of both developing and developed economies. They have proved to improve livelihoods by increasing per capita output and incomes. Further, MSEs have been found to be seedbed of entrepreneurial skills and innovation and greatest creator of jobs (Alliance for Financial Inclusion, 2017). It is because of this that the Kenyan government has prioritized MSEs as drivers to its industrialization and making the country a middle-income economy as contemplated in the *Kenya Vision 2030*. Despite this recognition, MSEs are encountering a myriad of challenges. Key among them is inadequate access to information related to market, financing, and competitors among others (Muturi, 2015). Because of their limited budgets, MSEs cannot afford highly research and development personnel. Because of limited skills, they lack capacity to acquire, interpret and effectively use research information. They have to depend on open source information whose research outputs are questionable due to flawed data collection techniques. Using such outputs predisposes them to coming up with wrong managerial decision.

In advent of information age, enterprises compete on the amount of accurate information they acquire to inform their decisions. Enterprises with the newest and the most accurate information at marketplace are at an advantage point over rivals. This has made research ubiquitous in small enterprise sector. Research has therefore become an important logical and systematic process used to investigate and find solutions to problems facing enterprises. Accurate information being an output of a good research is used by entrepreneurs in industry and businesses as a competitive strength and decision making to enhance productivity and to improve the quality of products. Researchers design methods and strategies to gather

information about the problem, measure the extent of the problem, and even predict its future manifestation. Desire for accuracy of the research output requires that the researcher embraces scientific approach that determines the kind of information or data to be determined and right the scale of measurement.

Researchers in the fields of psychology, sociology, education and now entrepreneurship seek information to predict people's reaction towards a phenomenon through measuring their attitudes. They have therefore improvised and adopted psychometric tools to measure attitude. Amongst them are Thurstone scale and Likert-type scale. Thurstone is individualized, quite expensive and is unable to measure large groups of items. Its attributes have discouraged its adoption my most researchers. Likert scale has gained popularity among most behavioral scientists, of late. Why? Because is simple to construct, easily readable and yields psychometric consistency when completed by respondents (Junior & Costa, 2014). Murray (2013) discovered that the scale is highly reliable, whether using parametric or non-parametric statistical techniques to analyze its outputs. A fact hotly contested in the psychometric measurement foray.

### 1.1. Statement of the Problem

Despite Likert-type scale being a widely used model in behavioural studies; it is abused with equal measures. As a result inaccurate findings, wrong interpretations, flawed conclusions and theory constructions are highly observed. Likert scale generates ordinal data. However, most researchers handle the Likert scale outputs as interval data (Henson, Hull, & Williams, 2010), instead of ordinal. When a behavioural measurement model such as Likert-type scale is flawed, it jeopardizes prediction and explanation functions of scientific research. As observed by Awang, Afthanorhan and Mamat (2012), outputs of abused technique lead to meaningless findings. Consequently, the consumers of such research output are misinformed ending up with wrong business decisions. Wrong decision due to flawed research output would fuel fatality rate among firms, especially those exposed to vulnerabilities of smallness and newness.

### 1.2. Objectives of the study

1. To identify strengths and pitfalls of Likert type scale as psychometric measurement scale
2. To find out appropriate Likert-type scale modifications that yield reliable and accurate research outputs for MSE development

## 2. Literature Review

In order to clearly show the rationale and need for new and improved psychometric measurement scale, background and pre-existing knowledge about the problem is reviewed. Literature review unveils what is already known, what is not known and what this study investigates and the reason for investigation about Likert scale. This section delves into relevant theories and prior literature to achieve the study objectives.

### 2.1. Theoretical review

Theories are highly developed models of reasoning explaining the occurrence of phenomena (Denney & Tewksbury, 2012). The theories found to explain why a problem exists in the use of Likert-type scale and why it needs to be corrected are: Classical test and Generalization theories.

#### 2.1.1. Classical test theory

Classical Test Theory (CTT) is a psychometric theory that predicts each examinee's true score by eliminating error from observed score in a behavioural set of measurement item. It aims at studying reliability of test scores, correlation of two random variables after filtering out error and true score confidence interval. According to CTT, observed score is composed of a true score and error, as shown in the model:

$$X = T + E.$$

Whereas X is observed score, T is the true score and E is the error score. True scores are invisible and the error is a normal distribution random variable. Every person examined has a true observable score obtained if errors were eliminated. The theory, therefore, prescribes maximum determination of information about an individual and minimization of measurement error (Bichi, 2016). The theory presumes that tests are fallibly imprecise tools because of errors. To improve tests, CTT theory suggests

that the researcher focuses on difficulty, discrimination and reliability when designing a measurement scale. Difficulty is the number of the examinees that got the item correctly. Discrimination is the variance between the low and high scoring examinees, that is; the ability of a test item to discriminate between highest and lowest ability examinees. Reliability is a test level statistics showing dependability. The CTT has been broadly adopted in education and psychology as a scientific framework to improve test analysis and test refinement procedures. Because enterprise development studies borrow a lot from educational and psychological techniques of research, CTT can equally be used to improve psychometric measurements in entrepreneurial phenomena of interest.

Despite the fact that CTT has been hailed for its power to assess score dependability in behavioural or psychometric studies, studies have also found out that the theory is oversimplified to address the measurement challenges of the world we currently live. For example, CTT assumption that observed score is made of a true score and error whose sources are indistinguishable is a big weakness in the theory (Prion, Gilbert, & Haeling, 2016). Other observed weaknesses include incapability to estimate numerous reliability aspects at once and inability to distinguish relative from absolute rank order decisions where psychometric measurement is applied.

**2.1.2. Generalizability theory**

In light of the aforesaid weaknesses, in 1963 Cronbach, Gleser, Nanda and Rajaratnam came up with a more robust comprehensive framework. The framework is called Generalization (G) theory. It estimated consistency of scores with more than one source of measurement error distinctively and simultaneously (Vispoel, Morris, & Kilinc, 2018). Beyond evaluating reliability of psychometric measurements, G-theory identifies sources of both systematic and unsystematic error variations using Analysis of Variance (ANOVA) methodology. The framework also differentiates the relative (rank-order) from absolute decisions. It is more powerful than CTT. It compels the researcher to see reliability not in tests but in the scores (Thompson & Crowley, 1994). Simultaneously, G-theory estimate multiple sources of errors, their variance and interactions in reference to the true score. This makes G-theory most recommended, modern and essential where CTT is insufficient.

**2.2. Challenges of Psychometric Measurement in Behavioural Studies**

The primary goal of measurement in behavioural science is to clarify and quantifying links between unobservable latent variables and observable variables. The researcher gets the opportunity to evaluate and relate the latent and the observable variables to discover uniformities of elements and patterns. This presupposes that the scientific model must be accurate enough to generate correctly transform qualitative – behavioural data into right scale of quantitative data. It proves trickier when constructing a scale to measure attitude (Rattanalertnusorn, Thongteeraparp, & Bodhisuwan, 2013). But it is of utmost importance that in measuring a phenomenon of interest (e.g. perceived enterprise performance, customer delight, loyalty, workers' motivation among others), the researcher has to achieve a stable consistent measure of the respondent's level on that scale, for analysis of severity, proper decision making and appropriate choice of business strategy.

Before anything else, the researcher must choose the scale to numerically measure data with so as to determine the right statistical analysis. Brown (2011) admits that there are four flavours of scales of measurement; that is nominal, ordinal, interval, or ratio. Each scale of measurement is useful in its own rights. Whereas nominal scale is best at measuring categorical data, ordinal scale is best at ranking attitude responses, interval scale orders things with equal intervals between the scale points. Ratio is good at measuring things requiring zero values and points along the scale such as temperatures. Confusing one scale of measure for the other would expose the researcher to appropriate statistical analyses technique that would lead to flawed research outputs and misinformation.

**2.3. The Likert-type Scale**

Likert-type scale is a summative multi-item gradation scale meant to gauge psychological attitudes of a population about a phenomenon. It was invented by Rensis Likert, an American civil engineer and sociologist, in 1932. The original Likert type of scale has 5-point order (pentatonic). Respondents are expected to give their extent of agreement or disagreement to a series of attitude statements or questions relating to phenomenon. Every response is assigned a point with quantitative value. All values are then summated in a rating scale.

**Table 1:** Example of original Likert-type scale

| Statement | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|
| 1. I understand the difference between Likert items and Likert scales | 1 | 2 | 3 | 4 | 5 |
| 2. I understand how to analyze Likert items | 1 | 2 | 3 | 4 | 5 |
| 3. I like using Likert items | 1 | 2 | 3 | 4 | 5 |

**Source:** Brown (2011)

According to table 1 above, the items/statements 1to 3 are ranked or ordered on a scale of one to five. Strongly disagree is assigned the value of 1, disagree 2, neutral 3, agree 4 and strongly disagree 5. Despite the order, the scale does not disclose the interval distances between the points (Brown, 2011). To misconceive the ranks as interval data at equidistance would attract serious errors in data computation and interpretation.

**2.3.1. Weakness of traditional mode of Likert five point scale**

Studies have absolved confusion, issues and challenges in using traditional pentatonic Likert type data. The worst occurs when Likert output is treated as interval values (DeWinter & Dodu, 2010). The scale is should neither be interpreted interval nor ratio scale but ordinal scale. Misinterpreting the Likert scale (as is commonly done) amounts to abuse and undermining parametric technique strength. Other observed weaknesses include participants tending to avoid extreme categories (central tendency bias). Instead of upholding honesty, they portray themselves in a socially favourable manner (social desirability bias) and that that pleases the researcher (Subedi, 2016). It has also been found out that Likert scale luck reproducibility and validity. Lucian (2016) observed that the traditional Likert type scale equally suffers from inability to measure the amount of change and the degree of favourableness by respondents. The scale only permits rational conjectures, rarely justifies the choice of n-point values and unable to analyze parametric data (Lucian, 2016). In addition it does not address cross-cultural issues (Murray, 2013).

### 2.4. Mending flaws in Traditional Pentatonic Likert type scale

The effort to come up with a more effective scale of measure must be anchored on classical measurement theory which stresses on a pool of homogeneous items. It must also adhere to the Nunally principles of scale construction according to classical measurement theory and generalization (G) theory (Viljoen, 2015). The Nunally principles include: (i) identify and measure attitudinal objects and their dimensions with a specific population in mind; (ii) start writing an item pool of 40 and reduce it to 20; (iii) choosing the best number of scale points to use; (iv) selecting the anchor; (v) deciding the length of the scale; (vi) piloting the scale; (vii) measure the reliability; and (viii) determine the validity. In order to realize more faithful results in parametric tests, one has to eliminate the observed limitations above of the five point scale Likert type scale through the following modifications.

#### 2.4.1. Adding more scale points to the five-scale type

Original tool was a pentatonic scale but with observed weaknesses, it has been improved overtime. For example, to optimize reliability, the scale has been modified to a 7- point scale (Croasmun & Ostrom, 2011). Recent parametric structural equation modelling found out that a decatonic (10 point) of Likert type scale was more efficient than the original pentatonic one (Awang, Afthanorhan, & Mamat, 2012). According to Junior and Costa, 2014, Likert type scale works less reliably when items are measured using pentatonic scale and below and more reliably when items are measured by more than 7 points. In a study of 149 respondents from store and restaurant setups, 2-point, 3-point and 4-point scale performed dismally as compared to scales with 10-poin, 7-point and 9-point. The later scales demonstrated significantly higher reliability, validity and discriminating power (Preston & Colman, 2000).

The most optimal no of points on modified Likert type scale for reliability, validity, discriminating power and respondent preferences is seven(7). This has been strongly echored by a study on likert items and scales (Johns, 2010). Modification by increasing more scale points should be cautiously done. It is likely to make respondents lazy and increase primary effects in Likert scales. The primary effects include respond-order effect where respondents tend to choose first response available on answer scale. Second primary effect is the donkey vote effect; respondents select same responses for all question. Finally is the central tendency effect where

neutral responses tend to be the choice for all questions. When these primary effects prevail, measurement errors are high (Li, 2013).

**2.4.2. Phrase completion scale**

In 2003, Hodge and Gillespie developed a standard 11 point range scale from 0 to 10 to address weaknesses in the traditional pentatonic Likert type scale (Hodge & Gillespie, 2007). In phrase completion scale, the integers are sequentially arranged relating to the intensity of respondents feelings. It starts with zero (0) represents missing attributes. The phrase completion points increases measurement reliability and validity in surveys by increasing value points to 10 and inserting intensity in the Likert scale as shown in the table below.

**Table 2:** Phrase completion scale sample table

| My level of satisfaction with the service was: | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Too small | | | Moderate | | | | | Too big | | |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

**Source:** Author (2019)

Responders are expected to show their satisfaction degree by finishing the sentence within the 11-point range above. After comparing measurement and verification scale of Likert-type and Phrase completion scales through a study of 229 responses, Junior and Costa (2014) found out that phrase completion yielded better reliability and functional consistency.

**2.4.3. Novel fuzzy Likert scale**

A fuzzy Likert scale is a new construction of fuzzy rating score on a traditional pentatonic Likert scale using survey questions based on fuzzy set theory. On a new form, each respondent is expected to identify a level of agreement which will be matched with membership degree, as shown in the table 4 below. A recorder will design a decision tree to figure out the probability of the degree of membership of the rest. It is therefore an ideal tool for transforming ordinal data into interval format (Li, 2013).

**Table 3:** The summarized result of the fuzzy rating scores

| Statement item | Agreement level | | | | | Traditional rating | Fuzzy rating |
|---|---|---|---|---|---|---|---|
| Do you agree with Thai government's policy to prohibited using Liquefied Petroleum Gas (LPG) for a new car? | SD (1) | D (2) | NN (3) | A (4) | SA | | |
| Respondent 1 (Mr. A) | | | | 0 | 1.0 | 5 | 5 |
| Respondent 2 (Mr. B) | | | | 0.7 | 0.3 | 5 | 4.3 |
| Respondent 3 (Mr. C) | | | | 0.2 | 0.8 | 5 | 4.8 |
| Average | | | | | | 5 | 4.70 |

Adopted from ICEAS (2013)

In addition, it can capture lost and distorted information. According to Rattanalertnusorn, Thongteeraparp and Bodhisuwan (2013), the new fuzzy likert scale precisier than the 5-point likert scale.

**2.4.4. Two-stage Likert scale**

In 1997, Albaum modified the Likert scale in two stages to address the central tendency effect and equally capture more extreme options in responses. The scale measures the *direction* and *intensity* dimensions in responses by splitting attitude questions in two stages. Stage one is concerned with measuring the direction dimension of attitude; whether the respondent agreed or disagreed to the item. Stage two measures the intensity (strength) dimension of the respondent's attitude; that is the degree of agreement, whether strong or weak. Measurement of interaction effect can then follow (Albaum, 1997).

Three studies carried out in three different countries on university students, Albaum(1997) found out that a two stage type of Likert scale demonstrated a more powerful predictive capacity than the old five point Likert type scale that confounded the dimensions into one-stage

measurement. However, it is still doubted if the scale could measure large amount of data as the traditional pentatonic scale.

**2.4.5. Transforming ordinal data into Binary Data**

This measurement scale is based on the Item Response Theory (IRT). It is a binary scale where respondents are required to choose between "Yes" and "No," that is; positive and negative attitudes respectively. This kind of scale eliminates much observed limitations in the traditional pentatonic Likert scale. It yields more faithful results and accommodates both parametric and non-parametric calculations. Lucian (2016) observed that dichotomous measurement scales were not only efficient, but also more effective, reliable, directional and respondent-friendlier.

This approach entails collapsing the Likert type scale into two, which is; 0 and 1. On one hand scores below 3 on the scale can be assigned zero (0) meaning "No" (negative) answer to the attitude question. On the other hand values above three could be rounded up to 1 meaning "Yes" (positive). Index values can also be transformed in similar manner. Indices above 0.5 can be rounded up to Yes value of 1 and below 0.5 to No value of 0. This transforms the Likert-type scale from ordinal to dichotomous scale and ratio scale values, respectively. After which it can accommodate highly inferential techniques of analysis. For example Logit regression is a quite powerful econometric statistical technique that can only accommodate data outputs of binary nature. In a study where Likert-type scale is used, the researcher has to transform the scale outputs into dichotomous or binary data. Brown (2011) confirmed that actually Likert scale is collapsible into bimodal data.

**2.4.6. Transforming ordinal data into indexes**

Another strategy of making ordinal data of Likert type scale be analyzed through parametric techniques is by transforming its outputs into ratio scale data. For example, a variable is measured by ordinal data on a Likert scale of 7 point by 9 attitude questions (items). The index is derived as a result of each respondent's highest score divided by the maximum expected score. Each respondents total score is divided by the maximum possible score 63 (9 questions X 7 points) to get the indices. The index will fall between zero (0) and one (1), forming ratio values that can be analyzed by parametric techniques.

Further these values can be transformed into binary data. The ratio scale can be collapsed into two, which is; 0 and 1 which is the preferred model for Logit regression. All values below 0.5 are considered to be 0 and all values above 0.5 are considered 1. Responses that score above 0.5(1) shall account for "Yes" or positive attitudes. Respondents that score below 0.5(0) shall account for "No" or negative attitude.

### 2.4.7. Rasch Rating Scale Model (RSM)

Another strategy of transforming Likert-type scale ordinal data output into interval is by Rasch analysis. This is a single statistical parameter that measures relationship between response patterns, item difficulty and the expected patterns using a linear interval scale. Whether by new or revised scale, Rasch RSM is hailed for assessing unidimensionality, differential functional outcomes, item fit, validity and reliability in psychometric studies (McCreary, Conrad, Scott, Funk, & Dennis, 2013). In a logical manner Rasch RSM calculates intervals between items and uniqueness among subjects tested in a scale (Lewis & Horn, 2017). The intervals data found could be used for parametric analysis. To generate precise calculations, the model optimizes the number of points or log its and categories of items (Bartholomeu, da Silva, & Montiel, 2016). This model does not only improve the Likert scale, but has also been found to be an effective measurement model in brain injury populations, psychometric analysis of adult women in family set-up and children's social skills(McCreary, Conrad, Scott, Funk, & Dennis, 2013; Lewis & Horn, 2017; Bartholomeu, da Silva, & Montiel, 2016).

## 3. Conclusion

Small enterprises being the engine of economies are in dire need of accurate research outputs that would address their challenges currently fuelled by globalization and competitiveness. Small enterprise developers must focus on measurement techniques that are accurate and dependable for useful information generation and acquisition. This study found that the pentatonic Likert type scale, which is a favourite of many researchers, is obsolete and overtly abused. Consequently research outputs are flawed and highly misleading for policy makers and business executives. The study recommends modifications of the scale by observing CTT and G-theory techniques. Specifically the study suggests increase in measurement

scale points to seven. Secondly, the scale's ordinal data is transformed to binary, interval or ratio flavours. Binary and ratio data could be achieved by collapsing the scale and computing the outputs, respectively. Other modifications to enhance original Likert-type scale include phrase completion scale, novel fuzzy Likert scale, two-stage Likert scale and Rasch analysis.

**List of References**

Albaum, G. (1997). The Likert scale revisted: An alernative version. *Journal of Marketing Research Society*, 331-348.

Alliance for Financial Inclusion. (2017). *SME working group survey report: Defining micro, small and medium enterprises(MSMEs) in the AFI Network.* Kuala Lumpur, Malaysia: Alliance for Financial Inclusion(AFI).

Awang, Z., Afthanorhan, A., & Mamat, M. (2012). Likert scale analysis using parametric based stractural equation modelling(SEM). *Computational Methods in Social Sciences*, 13-21.

Bartholomeu, D., da Silva, M. C., & Montiel, J. M. (2016). Improving the Likert scale on the children's social skills test by means of Rasch model. *Psychology*, 820-826.

Bichi, A. A. (2016). Classical test theory: An introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies*, 27-33.

Brown, J. D. (2011). Likert items and scales of measurement. *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, 10-14.

Croasmun, J. T., & Ostrom, L. (2011). Using Likert-type scale. *The Social Sciences*, 19-22.

Denney, A. S., & Tewksbury, R. (2012). How to write a literature review. *Journal of Criminal Justice Education*, 1-17.

DeWinter, J., & Dodu, D. (2010). Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research and Evaluation*.

Henson, R., Hull, D. M., & Williams, C. S. (2010). Methodology in our education research culture. *Educational Researcher*, 229-240.

Hodge, D., & Gillespie, D. (2007). Phrase completion scales: A better measurement approach than Likert scales? *Journal of Social Science Research*, 1-12.

Johns, R. (2010). Likert items and scales. *Survey Question Bank: Methods Fact Sheet*, 1-7.

Junior, S. D., & Costa, F. J. (2014). Measurement and verification scales: A comperative analysis between the Likert and Phrase completion scale. *Brazilian Journal of Marketing, Opinion and Media Research*, 1-15.

Lewis, F. D., & Horn, G. J. (2017). Rasch analysis and functional measurement in post-hopital brain injury rehabilitation. *International Journal of Statistics and Probability*, 50-59.

Li, Q. (2013). A novel Likert scale based on fuzzy sets theory. *Expert Systems with Applications*, 1609-1618.

Lucian, R. (2016). Rethinking the use of Likert scale: traditional or technical choice? *Brazilian Journal of Marketing, Opinion and Media Research*, 11-26.

McCreary, L. L., Conrad, K. M., Scott, C. K., Funk, R. R., & Dennis, M. (2013). Using the Rasch measurement model in psychometric analysis in the family effectiveness measurement. *Nursing Research*, 149-159.

Murray, J. (2013). Likert Data: What to use, Parametric or non-parametric? *International Journal of Business and Social Science*, 258-264.

Muturi, P. M. (2015). The role of micro and small enterprises(MSEs) in achieving Kenya Vision 2030. *International Journal of Economics, Commerce and Management*, 1337-1352.

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

Prion, S. K., Gilbert, G. E., & Haeling, K. (2016). Generalizability theory: An introduction with application to simulation evaluation. *Clinical Simulation in Nursing*, 546-554.

Rattanalertnusorn, A., Thongteeraparp, A., & Bodhisuwan, W. (2013). Fuzzy rating score on the Likert scale. *International Conference on Engineering and Applied Sciences(ICEAS, 2013)* (pp. 291-325). Osaka, Japan: ICEAS.

Subedi, B. P. (2016). Using Likert type data in social science research: Confusion, issues and challenges. *International Journal of Contemporary Applied Sciences*, 36-49.

Thompson, B., & Crowley, S. (1994). When classical measurement theory is insufficient and generalizability theory is essential. *the Annual Meeting of the Western Psychological Association* (pp. 1-18). Kailu-Kona, Hawaii: Western Psychological Association.

Viljoen, M. (2015). Constructing homogeneous Likert-type summative rating scales according to classical measurement theory. *Journal of Social Sciences*, 143-151.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of genralizability theory and their relations to classical test theory and structural equation modelling. *Psychological Methods*, 1-26.