RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES

Social Science Journal

Likert Scale Labelling and Variation in Response: A study comparing three-response labels

By

Essa J. H. Al-Harbi
Associate Professor of Measurement and Evaluation Islamic University of Madinah Saudi Arabia
Email: Alhrbi909@gmail.com

Abstract

Based on the familiarity hypothesis, respondents are affected while using response categories in answering surveys. This may affect the research findings where alternative scales such as, "strongly disagree- strongly agree", or percentages are used to obtain response. This study explores the impact of using three different Likert Scale labeling on the participants' responses to questionnaire items. The study explored whether there are statistically significant differences in the responses of the participants attributed to the Likert Scale labeling (on three option types: description - scores - percentage). It also calculated whether there is a correlation between the responses and the type of Likert scale labeling (description - scores - percentage). The questionnaire used was administered thrice using the three different Likert Scale labeling methods.. In the first, the items were answerable with the alternatives, (strongly agree into strongly disagree), in the second, the questionnaire items were answerable with scores (from 0 to 10), and the third version provided participants with items answerable in percentages (0% into 100%). The questionnaires were administered to 382 participants. Findings showed that the average scores of the respondents are more inflated in the case of description label of the Likert Scale compared to the scores and percentages labelling. Similarly, the mean scores of the respondents are more inflated in the case of using scores label of the Likert Scale compared to the third method (percentages label of the Likert Scale). Based on the mean values, the study recommends using the scores labeling of the Likert Scale as in this case the respondents are the most careful in choosing an appropriate option for their responses.

Keywords: Likert scale labeling, questionnaire, descriptive alternatives, statistically significant

1. Introduction

The Likert Scale is a commonly used tool for measuring attitudes or opinions in social science research. It consists of a series of statements or items that respondents rate on a scale from the most common naming levels as follows: i. Strongly disagree: This response option indicates a very negative attitude or opinion towards the statement being evaluated. ii. Disagree indicates a negative attitude or opinion towards the statement being evaluated. iii. Neutral response option indicates a lack of opinion or a neutral attitude towards the statement being evaluated. iv. Agree indicates a positive attitude or opinion towards the statement being evaluated. v. Strongly agree indicates a very positive attitude or opinion towards the statement being evaluated (Ponsiglione et al., 2022). Figure 1 depicts these Likert Scale options pictorially.

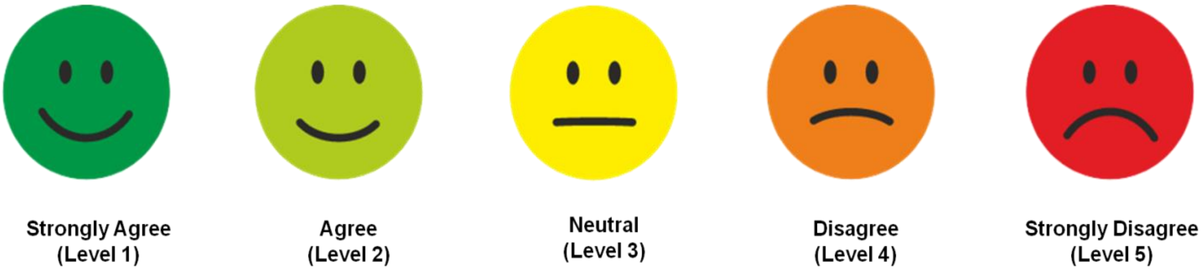

Figure 1. *Schematic representation of Likert Scale*

The responses are usually converted numerically and then, summed to create an overall score for each respondent. These five naming levels allow for a balanced and easily interpretable Likert Scale that clearly shows the distinctions between different levels of agreement or disagreement with a given statement (Moreno-Garcia et al., 2022). When it comes to naming the different levels of the Likert Scale, there are several different conventions that researchers use. Here are a few examples: (a) Five-point scale: This is the most common type of Likert Scale, and it consists of five response options: "strongly agree," "agree," "neutral," "disagree," and "strongly disagree." (b) Seven-point scale: Some researchers prefer to use a seven-point scale, which includes additional response options such as "somewhat agree" and "somewhat disagree." (c) Ten-point scale: A ten-point scale is another option, which provides even more response options for participants. This can be useful if the researcher wants to capture more nuanced range of attitudes or opinions. It's important to note that the number of response options on a Likert Scale can affect the reliability and validity of the results, so it is absolutely imperative for researchers to choose a scale that is appropriate for their needs (Heo et al., 2022).

One of the most used tools for gauging opinions, preferences, and attitudes is the Likert Scale. Analysis may be based on a single item or the total of many things that make up a scale (South et al., 2022). A Likert item is often distinguished from a Likert-type item because the former have bivalent and symmetrical labels centered on a middle or neutral point (Wu et al., 2022). The Likert Scale is still widely used, however, the several problems that responses so generated are obtained need deeper evaluation. First, all points can be labeled, however, sometimes it's only possible to identify the endpoints. Likert initially labeled each choice to be selected, but this may have taken away from the interval character of the options, thus only end-defined labels were included (Anjaria, 2022). Four different labellings of 5-point scales were examined, and it was discovered that although they may have different variances, their means, and reliability are not likely to change. Reliability is based on item-item correlations, which are independent of variation. Additionally, they discovered that using more absolute endpoints might lead to frequencies being concentrated in the center, and vice versa, and recommended using fewer absolute labels. Studies may elect to utilize several labels, with even a blank label being one of the choices. Second, there is a lack of consensus over the appropriate number of scale points. Most studies employ four to seven points, although others may go as high as ten or eleven. Likert Scales are commonly used in surveys and questionnaires to measure the attitudes or opinions of respondents. They consist of a series of statements or items, each of which is accompanied by a response scale across a response range. The respondent is asked to indicate their level of agreement or disagreement with each statement using the response scale (Jebb et al., 2021).

1.1 Research questions

1. Are there statistically significant differences in the response iterations or lineages of the participants attributed to the Likert Scale labeling (description - degree – percentage)?

2. Does the distribution of responses differ depending on the Likert Scale labeling (description - degree - percentage)?

2. Literature Review

The work of Jebb et al. (2021) provided psychological researchers with information on more recent psychometric developments in the design of Likert Scales. Wu and Leung (2017) examined this argument and hold the same opinion, but they use simulation to create fake data from symmetrical normal and skewed distributions when the underlying measure is known in advance. Mircioiu and Atkinson (2017) employed a methodical approach to evaluate real Likert data on responses from different professional subgroups of European pharmacists about practicing competencies. Chyung et al. (2017) examined research findings from many fields to show that there are instances in which a midpoint should be included and others in which it should not. O'Neill (2017) offered three figures and a "quick reference" table to help readers understand how IRA numbers vary and how IRA is interpreted will be greatly influenced by the statistic used. The purpose of the study was to examine the challenges and issues related to evaluating the validity and analyzing data from a Likert Scale, as well as how to construct an effective Likert Scale (Mirahmadizadeh et al., 2018). The study of Douven (2018) examined the various Likert-type scale forms and how they relate to data quality. Overall, none of the earlier studies examined the effect of response choice, if any, on the responses obtained which qualifies as a pertinent area of research for obvious reasons. This gap in the available literature has been filled by the current study.

3. Methodology

3.1Dataset

The study sample consisted of 382 people who named their responses using the Likert method on a questionnaire with ten items that was given to them in three distinct ways (description - scores - percentage). SPSS (version 23) was used to process the findings.

3.2 Label the levels of the staging of the responses

10 items were submitted for the study sample. In a questionnaire after reviewing, auditing, and arbitration, these items are answered in three different ways:

The first way

It is the description labelling, which includes the five responses in a descriptive form, ranging from (strongly disagree) to (strongly agree), and these five responses take the five numerical values from (1) to (5).

Second way

It is the grading method, which includes the five responses in the form of degrees, ranging from (0) to (10) degrees, and these five responses also take the five numerical values from (1) to (5).

Third way

It is the ratio method, which includes the five responses in the form of percentages, ranging from (0) to (100%), and these five responses also take the five numerical values from (1) to (5). The following table 1 shows the three ways of naming the levels of the Likert Scale:

Table 1 *Methods to Label the Likert scale levels and numerical scores corresponding to each response*

The corresponding numerical value for each response	**1**	**2**	**3**	**4**	**5**
Methods	Responses				
Description	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Scores	0	3	5	7	10
Percentage	0%	25%	50%	75%	100

3.3 Study tool

The researcher prepared a questionnaire consisting of 10 phrases, to be answered on a five-point scale, with three naming methods of Likert calibration as shown in Table 1. The reliability and validity coefficients of this scale were calculated after applying it to a survey sample consisting of 75 individuals who were randomly selected outside the sample of the study, where its reliability and validity were calculated in the following ways:

3.4 Reliability and validity

The Cronbach's alpha coefficient Alpha-Cronbach (by the volume of scale expressions) and every time one of the phrase scores are omitted from the overall score of the scale, we carry out the calculation of coefficients of correlation between the scores of the phrase and the overall score of the scale. The validity of the scale statements was calculated in this manner. When deleting the score of the phrase from the total score of the scale, given that the rest of the elements of the scale are a test of the veracity of the statement, the results were as shown in the following Table 2:

Table 2 *The stability and validity coefficients of the scale (n = 75)*

Items	**The correlation coefficient of the degree of the item with the total score of the scale when deleting the degree of the item from the total score of the scale (validity)**	**The correlation coefficient of the degree of the item with the total score of the scale (reliability)**	**Cronbach's alpha coefficient**
1	0.49**	0.60**	0.815
2	0.55**	0.65**	0.809
3	0.47**	0.59**	0.816
4	0.55**	0.65**	0.809
5	0.60**	0.70**	0.803
6	0.56**	0.66**	0.808
7	0.50**	0.62**	0.814
8	0.58**	0.69**	0.806
9	0.49**	0.60**	0.815
10	0.38**	0.52**	0.827
	Cronbach Alpha -: The overall year of the scale =0.828		
	The overall ability coefficient of the scale by split half method for Spearman brown=0.845		

The following are clear from Table 2:

- In the absence of any of the items being less than or equal to Cronbach’s coefficient alpha scale, all items in the existing questionnaire help in raising the overall stability coefficient for the scale.

- That all the coefficients of the expression degree correlate with the total degree for scale at a statistically significant level (0.01), which indicates the internal consistency and stability of all the questionnaire statements.
- The overall stability of the scale as a whole was established using in Alpha-Cronbach's, the half hash of Spearman-Brown which showed significant high coefficients implying total stability for the questionnaire.
- That all the coefficients of the expression degree correlation with the total degree for scale (in the case of deleting the degree of the statement from the total score of the scale) shows statistical significance at the level (0.01), which indicates the validity of all statements of the questionnaire.

From the foregoing procedures, the researcher confirmed the stability and validity of the questionnaire.

3.5 Statistical methods

A set of statistical methods were used to answer the research questions of this study. These were:

- Chi-Square Test Square
- Crone Bach Alpha
- Repeated measures analysis of variance Analyzing data with repeated-measures ANOVA and then a least-significant-differences test LSD (Least significant difference) for multiple comparisons.

4. Results

RQ1: Are there statistically significant differences in the response iterations or lineages of the participants attributed to the Likert Scale labeling (description - degree – percentage)?

Iterations or lineages of the responses were established using the chi-square test results. Chi-square was computed for the frequencies of the responses according to the three labeling methods using the Likert Scale in each of the five response options (n = 382). The results were as shown in the following Table 3 and Figures 2-4:

Table 3 *Responses for description, scores, and percentage*

		Responses					
The numerical value of the response		**1**	**2**	**3**	**4**	**5**	**Total**
Description		**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**	
Scores		0	3	5	7	10	
Percentage		0%	25%	50%	75%	100%	
Description	Response	10	127	1119	1683	881	3820
	Frequency	0.26%	3.32%	29.29%	44.06%	23.06%	100%
Scores	Response	74	850	1557	1160	179	3820
	Frequency	1.94%	22.25%	40.76%	30.37%	4.69%	100%
Percentage	Response	758	1475	1160	356	71	3820*
	Frequency	19.84%	38.61%	30.37%	9.32%	1.86%	100%
Chi-square value		1225.0**	1225.0**	1113.6**	91.5**	1026.1**	3820
		0	0	3	5	10	100%

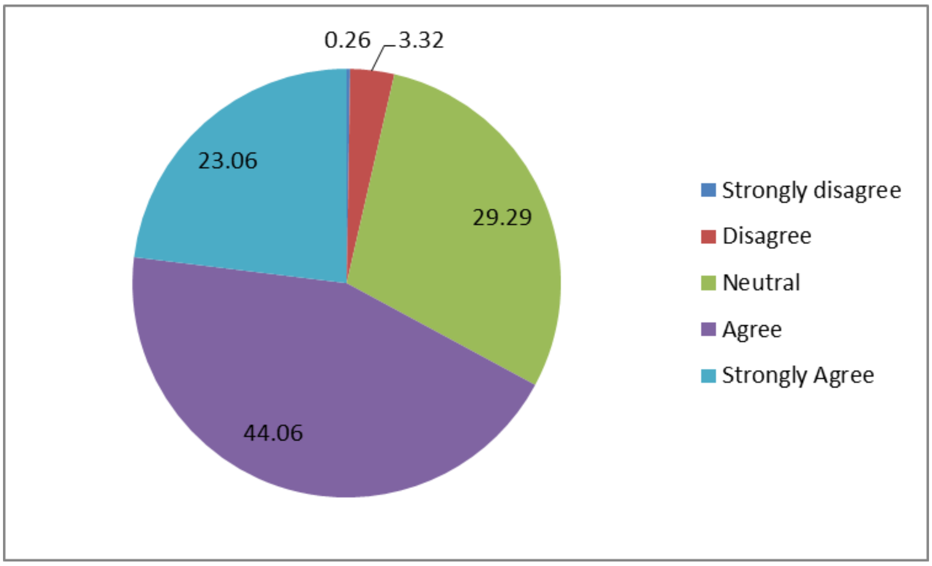

Figure 2. *Responses for description*

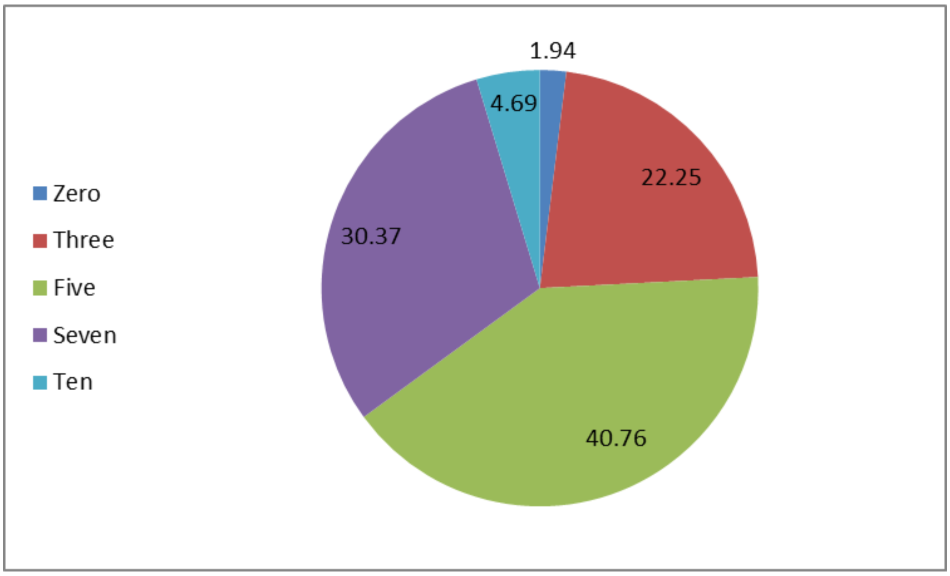

Figure 3. *Responses for scores*

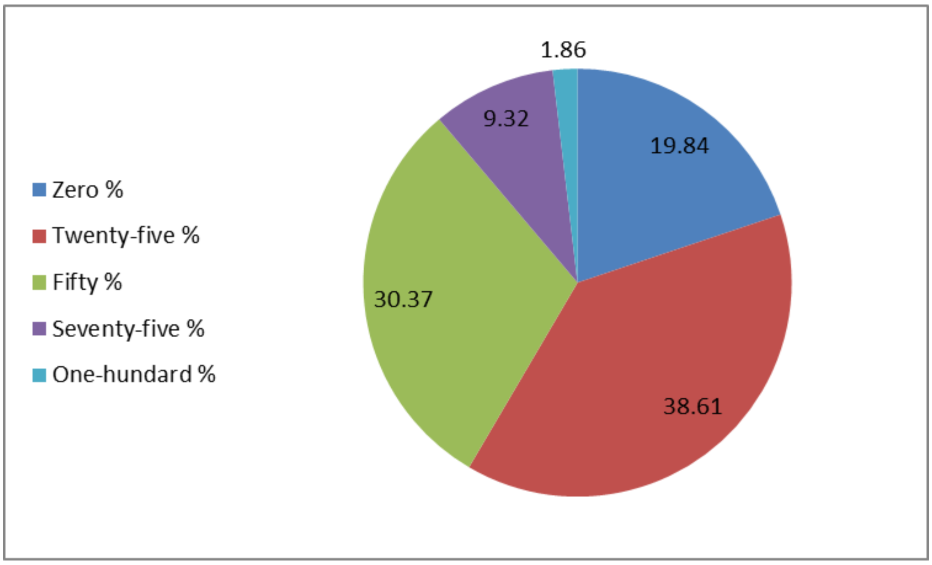

Figure 4. *Responses for percentages*

It is clear from Table 3 that:

- There are statistically significant differences (at the level of 0.01) between iterations according to the five response options (description-scores-percentages). As is evident from the chi-square test values which are equal to (1026.1, 838, 91.5, 1113.6, 1225) in the case of the five responses, respectively, statistical significance stands at (0.01).
- When examining the differences in the highest response, it was found that the differences are in favor of the response Strongly Agree which is the response that corresponds to the highest numerical value according to the first method (the descriptive method of labeling the Likert Scale levels) compared to the frequencies of individuals who chose the label corresponding to this response according to the other two methods, ie. scores and percentages.
- When examining the differences in the second highest response, the differences were found to be in favor of the response Agree which is the response that corresponds to the second highest numerical value according to the first method as well (the description method) compared to the frequencies of individuals who chose the label corresponding to this response in the other two methods (scores and percentages).
- When studying the differences in the third response, it was found that the differences are in favor of response 5 (five), which is the response that corresponds to numerical value 3 according to the second method (Likert scale labeling scores) compared to the frequencies of individuals who chose the label corresponding to this response according to the first two methods (descriptive and percentages).
- When studying the differences in the fourth response, it was found that the differences are in favor of the response (25%), which is the response that corresponds to the fourth highest degree in the classification according to the third method (Likert scale labeling in percentages) compared to the frequencies of individuals who chose the label corresponding to this response according to the first two methods (descriptive and scores).
- When studying the differences in the lowest response, it was found that the differences are in favor of the response (0%), which is the response that corresponds to the lowest numerical score according to the third method (percentages) compared to the frequencies of individuals who chose the label corresponding to this response according to the first two methods (descriptive and scores).
- It is notable that the first method (descriptive) attracted the respondents to respond mostly with (strongly agree, agree), which are the two responses that corresponded to the highest two degrees in the scores label.
- Table 3 shows that the second method (scores) attracted the respondents to respond by naming or choosing response (5), which is the response that corresponds to the middle degree or the median of degrees corresponding to the labels or responses.
- Table 3 shows that the third method (percentages) attracted the respondents to respond mostly (25%, 0%), which corresponded to the lowest two numerical values in the scores label.

RQ2: Does the distribution of responses differ depending on the Likert Scale labeling (description - degree - percentage)?

Repeated measures analysis of variance was used followed by the least significant difference test LSD (Least significant difference) for multiple comparisons between the three labelling methods in the total average of the ten items. The results were as shown in the following Tables 4 and 5:

Table 4 *Outcomes of LSD*

Source of Variance	Sum of squares	Degrees of freedom (df)	Mean of squares	Value (q) F	Significance level
Between methods	439.19	2	219.60	2683.73	0.01
Inside (error) methods	62.35	762	0.82		

The results of a repeated-measures analysis of variance when examining the differences between the total means show that:

- ➢ There are statistically significant differences (at the level of 0.01) between the total means of the grades on all items in the questionnaire attributable to the Likert scale labels (description - scores - percentages). This is evident from the value of (q), which is equal to (2683.73), and it is statistically significant at the level of (0.01).
- ➢ Less difference test results or LSD for multiple comparisons is used to determine the direction of the statistically significant differences when studying the differences between the total means of the grades of individuals in the study sample on all items attributable to the Likert Scale labels (description - scores - percentages).

Table 5 *Outcomes of Standard deviation and Mean (1st, 2nd, and 3rd methods)*

Methods of Labeling the Likert scale levels	Average Mean	Standard deviation (SD)	Description	Scores	Percentages
Description	3.86	0.30	-		
Scores	3.13	0.39	0.73**	-	
Percentages	2.35	0.41	1.51**	0.78**	-

It is clear from Table 5 that there is a statistically significant difference (at the level of 0.01) between the averages of the total scores for individuals in the sample of the study on all items for Likert scale responses in Description and Scores, in favor of the former. That is, the average scores of the study sample according to the first method for Likert scale designation (Description) are statistically significantly higher than the respondents' choices of the second method of the Likert scale (Scores). This indicates that the scores of the respondents are more inflated in the case of the first method, which is the descriptive method of Likert label scale, compared to the second method, which is the scoring method, as displayed in Figure 5.

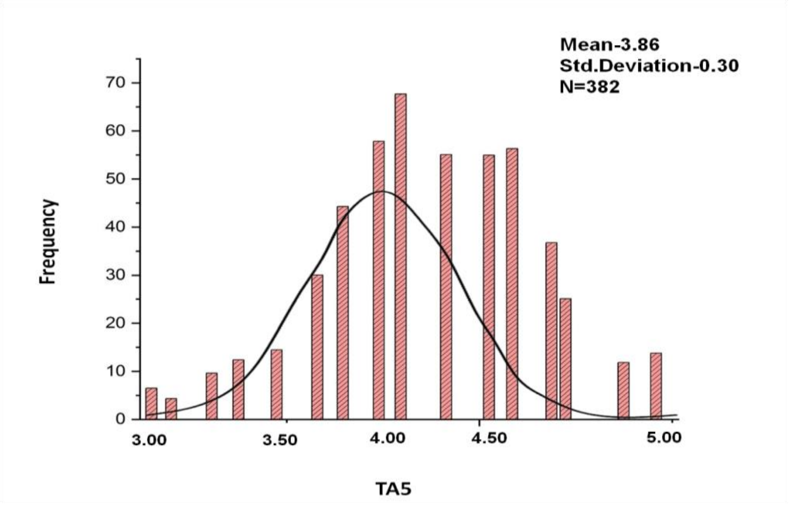

Figure 5. *Histogram representation of Standard Deviation and Mean of description labeling*

Table 5 indicates that there is a statistically significant difference (at the level of 0.01) between the averages of the total scores for the sample in the study on all items using Likert Scale Description and Percentages in favor of the former. That is, the average scores of the study sample according to the first method for Likert scale description labeling are higher in statistical significance than their counterpart according to the third method of Likert Scale labeling (percentages), and this confirms that the scores of the respondents are more inflated in the case of the first method, which is the description of Likert Scale labelling compared to the third method, which is percentages in Likert Scale labeling as seen in Figure 6.

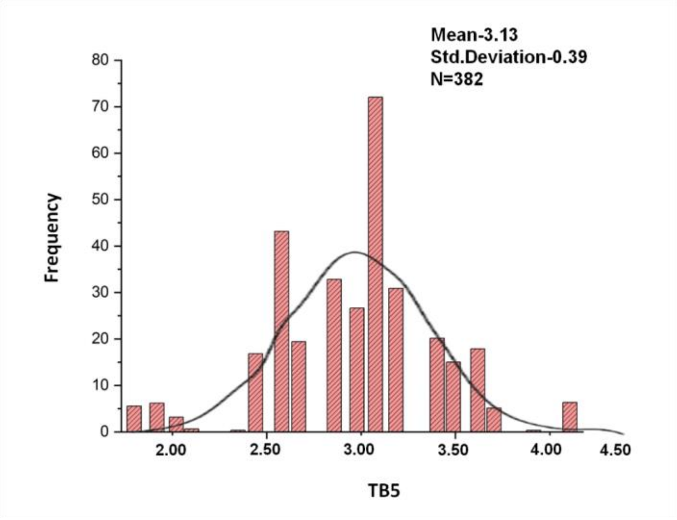

Figure 6. *Histogram representation of Standard deviation and Mean of scores labeling*

Table 6 also shows that there is a statistically significant difference (at the level of 0.01) between the averages of the total scores in a sample of the study on all Likert scale items using designation (scores) and Likert scale designation (percentages), in favor of the former (scores). That is, the average scores of the study sample according to the second method of Likert Scale labeling (scores) are statistically significantly higher than its counterpart according to the third method of Likert Scale labeling (percentages), and this indicates that the scores of the respondents are more inflated in the case of the second method, which is the scores method for naming in Likert Scales, compared to the third method, which is the percentages in Likert Scales labeling as seen in Figure 7.

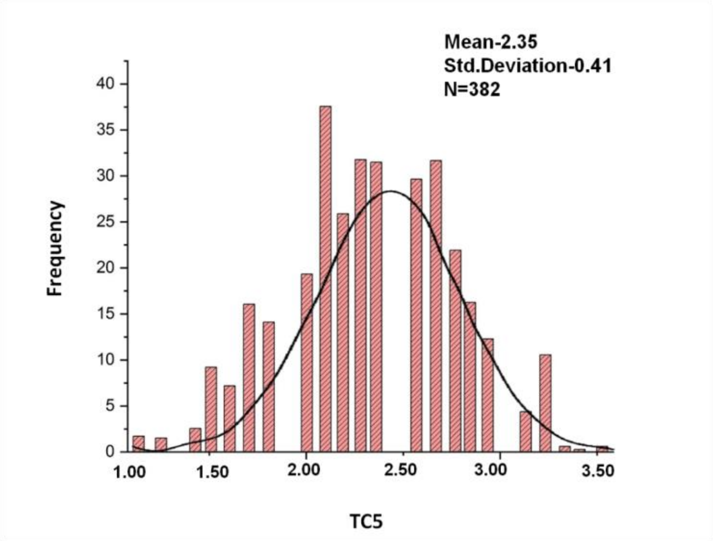

Figure 7: *Histogram representation of Standard deviation and Mean of percentages labeling*

Ranking the three methods in order of higher averages or high means or inflated grades in numerical terms indicates that the first method (the description labelling of Likert Scales) ranked first, followed by the second method (scores), and finally, the third method (percentages).

5. Conclusion

Findings of the study reported that the description label of the Likert Scale attracted the respondents to respond mostly with the two labels or the two responses (strongly agree and agree), which corresponded to the highest two degrees in the scores label. This finding agrees with Chyung et al. (2017) who examined research findings from many fields to show that there are instances in which a midpoint should be included and others in which it should not. O'Neill (2017) offered three figures and a "quick reference" table to help readers understand how IRA numbers vary and how IRA is interpreted will be greatly influenced by the statistic used. The purpose of the study was to examine the challenges and issues related to evaluating the validity and analyzing data from a Likert scale, as well as how to construct a Likert scale effectively (Mirahmadizadeh et al., 2018).

Findings also indicated that scores label of Likert Scales attracted the respondents to respond by selecting response 5, which corresponds to the middle degree or the median of degrees in the possible responses. Further, that the percentages label of Likert Scale attracted respondents to respond mostly with two responses (disagree and strongly disagree), which corresponded to the lowest two degrees in the scores label. This finding is in line with Lietz (2010) which found that participants preferred to choose the option of ‘don't know’ or a middle alternative.

Furthermore, that the average scores of the respondents are more inflated in the case of description label of the Likert Scale, compared to the other two methods: the scores label and the percentages label. Similarly, the mean scores of the respondents are more inflated in the case of the second method (the scores label of the Likert Scales) compared to the third method (percentages label of the Likert Scales). The order of the three methods in terms of higher means or inflated scores indicates that the description label of Likert Scales ranked first, followed by the second method (the scores label of Likert Scales) and finally, the third method (the percentages label of Likert Scales). These findings are confirmed by Weijters et al. (2013) who reported that labeling category 'strongly disagree' affected the responses due to their familiarity of these choices. They confirmed the familiarity hypothesis.

The study recommends using the second method when labeling a Likert scale, as labeling the scale with grades makes respondents more careful in choosing most accurate grades for their responses.

Recommendations

Based on the findings of the study, the following are recommended for future research using Likert Scales in surveys.

- When designing a Likert scale, it is recommended to label it with grades instead of descriptions or percentages, as respondents are more careful in choosing the most accurate grades for their responses.

- When constructing a Likert scale, it is important to consider whether a midpoint should be included or not, depending on the research question and context.
- When interpreting Likert scale data, it is important to be aware of the potential for inflated scores, particularly when using a description label or the scores label.

Limitations

The study's findings may not be generalizable to all populations or contexts, as the participants and research questions were specific to this study. Further, study only examined Likert scale responses and did not consider other types of survey responses or data collection methods. Lastly, study did not provide information on how to determine the appropriate number of response options for a Likert scale. Thus, further research could be conducted on these.

References

Anjaria, K. (2022). Knowledge derivation from the Likert scale using Z-numbers. Information Sciences, 590, 234-252.

Chyung, S.Y., Roberts, K., Swanson, I. & Hankinson, A. (2017). Evidence‐based survey design: The use of a midpoint on the Likert scale. Performance Improvement, 56(10),15-23.

Douven, I. (2018). A Bayesian perspective on Likert scales and central tendency. Psychonomic Bulletin & review, 25,1203-1211.

Heo, C.Y., Kim, B., Park, K. & Back, R.M. (2022). A comparison of Best-Worst Scaling and Likert Scale methods on peer-to-peer accommodation attributes. Journal of Business Research, 148, 368-377.

Jebb, A.T., Ng, V. & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. Frontiers in Psychology, 12, p.637547.Mirahmadizadeh, A., Delam, H., Seif, M. & Bahrami, R. (2018). Designing, constructing, and analyzing Likert scale data. Journal of Education and Community Health, 5(3),63-72.

Lietz, P. (2010). Research into questionnaire design: A summary of the literature. International Journal of Market Research, 52(2), 249-272.

Mircioiu, C. & Atkinson, J. (2017). A comparison of parametric and non-parametric methods applied to a Likert scale. Pharmacy, 5(2), p.26.

Moreno-Garcia, J., Yáñez-Araque, B., Hernández-Perlines, F. and Rodriguez-Benitez, L.(2022). An Aggregation Metric Based on Partitioning and Consensus for Asymmetric Distributions in Likert Scale Responses. Mathematics, 10(21), p.4115.

O'Neill, T.A. (2017). An overview of interrater agreement on Likert scales for researchers and practitioners. Frontiers in Psychology, 8, p.777.

Ponsiglione, A.M., Amato, F., Cozzolino, S., Russo, G., Romano, M. & Improta, G. (2022). A hybrid analytic hierarchy process and Likert scale approach for the quality assessment of medical education programs. Mathematics, 10(9), p.1426. https://doi.org/10.3390/math10091426

South, L., Saffo, D., Vitek, O., Dunne, C. & Borkin, M.A. (2022). Effective use of Likert scales in visualization evaluations: a systematic review. Computer Graphics Forum, 41(3), 43-55.

Weijters, B., Geuens, M., & Baumgartner, H. (2013). The effect of familiarity with the response category labels on item response to Likert scales. Journal of Consumer Research, 40(2), 368-381.

Wu, H. & Leung, S.O. (2017). Can Likert scales be treated as interval scales?—A Simulation study. Journal of Social Service Research, 43(4), pp.527-532.

Wu, W., Gu, F. & Fukui, S. (2022). Combining proration and full information maximum likelihood in handling missing data in Likert scale items: A hybrid approach. Behavior Research Methods, 54(2), 922-940.