Effect of Words, Numbers and Colours on Subjective Interpretation of Rating Scales

Niall Miranda

Cranfield University, United Kingdom

SUMMARY

Rating scales are tools that enable researchers to obtain empirical data from a theoretical construct. The quality of data obtained is influenced by subjective interpretation of the rating scale or measurement bias. A survey was conducted on eighty-two participants to determine the subjective interpretation of three distinct linear interval rating scales that employed either words to assign polarity, numbers to assign divisions or an experimental colour gradient. The data obtained in this study identified deficiencies in two commonly used questionnaires and recommended rating scale designs to attenuate measurement bias and improve the quality of data.

KEYWORDS

Measurement Bias, Questionnaire Design, Rating Scale, NASA TLX, SART

Introduction

Self-rated questionnaires are often employed to obtain empirical data on human performance. However, several studies aimed at improving questionnaire designs indicated variations in participants' interpretation of rating scales (Saris & Gallhofer, 2007). Deriving empirical evidence from questionnaires requires considerations for the quality of measured data influenced by these variations in responses (Rodgers & Herzog, 1992). Andrews (1984) attributed these variations to measurement bias developed by subjective interpretation of rating scales. Studies on cognitive load reduction strategies aimed at improving the quality of measured data concluded that it did not improve subjective interpretation of the questionnaires (Brosnan et al., 2021). However, employing colours to designate ordinal intervals in questionnaires with ordinal rating scales, such as the Cooper–Harper Handling Qualities and the Bedford Workload Rating (Roscoe & Ellis, 1990), are known to improve human interpretation of rating scales.

The study conducted in this paper involved two questionnaires with linear interval scales. The first questionnaire, the NASA Task Load Index (TLX), measured the subjective workload on a polarity scale, that is, words denoting the opposite ends of the scale (Hart & Staveland, 1988). The second questionnaire, the Situation Awareness Rating Technique (SART), measured subjective situation awareness on a numerical scale (Taylor, 2017). Besides the differences in the types of rating scales, the two questionnaires differed in the number of divisions in the rating scale. Whilst the NASA TLX questionnaire has 21 divisions in its rating scale, the SART scale has seven divisions in its rating scale. Prior studies have indicated seven divisions as ideal for rating scales (Miller, 1956; Preston & Colman, 2000), however, a study conducted by Matell & Jacoby (1972) attributed the certainty and accuracy of responses to increased number of divisions. Cox (1980) attributed variations in responses to factors such as the number of divisions in a rating scale and the ability and interests of participants. Therefore, this study investigates the effects of the types of scales and the number of divisions on subjective interpretation to attenuate measurement bias. Key differences in the rating scales used in these questionnaires are shown in Table 1.

Table 1: Key differences in the rating scales used in the NASA TLX and SART questionnaires.

	NASA TLX	SART
Type of Rating	Polarity	Numbers
Scale Divisions	21	7

Method

A human-in-the-loop experiment was conducted to determine participants' responses to aural cockpit checklists in an aircraft simulator environment whilst attempting to perform a landing (Miranda et al., 2024). During the experiment, participants provided self-rated scores for mental workload and situation awareness in the NASA TLX and SART questionnaires respectively. Each questionnaire had been modified to implement three distinct linear interval rating scales. After the experiment, a survey was conducted among the participants to obtain insights and subjective preferences for the three linear interval scales. The survey also determined the type of rating scale that was most appropriate to convey subjective scores that were required for the experiment.

Participants

The research involved a sample size of eighty-two participants with ages that ranged between 19 and 64 years (M = 27, SD = 6.75). Approval of the Cranfield University Research Ethics System was obtained before conducting the research. The data obtained from the research were made available only to the research team and stored under the Ethical Code and the Data Protection Act of the university.

Equipment

A high-fidelity cockpit simulator developed by Cranfield University, Rolls-Royce, and DCA Design, called the Future Systems Simulator (Korek et. al, 2022) was used to conduct a human-in-the-loop experiment. Subjective mental workload and situation awareness scores were obtained from NASA TLX and SART questionnaires respectively in the conventional pen-and-paper method.

Procedure

Participants were presented with three scenarios that varied in automation assistance and weather. After each scenario, participants provided their perceived workload and situation awareness scores on the NASA TLX and SART questionnaires respectively. Block randomisation of the rating scale was ensured during the allocation of questionnaires in each of the three scenarios.

Questionnaire Design

The original questions and dimensions of the rating scales were retained; however, for this research, the rating scale was altered to contain either polarity, numbers, or a colour gradient. The rating scales in the two questionnaires were modified to implement numerical and polarity rating scales in both questionnaires. In addition to numerical and polarity rating scales, an experimental colour gradient rating scale was also implemented in both questionnaires. Appendices 1-3 show scaled-down NASA TLX and SART questionnaires with polarity, numerical and colour gradient rating scales respectively.

Findings

The survey conducted on the participants' preference for the types of scale did not indicate an absolute majority. A simple majority of 29 participants preferred the numerical scale whilst 24 participants preferred the colour gradient, and 21 participants preferred the polarity rating scale. Figure 1 graphically illustrates the results from the survey. Eight participants, however, did not perceive differences in the rating scales.

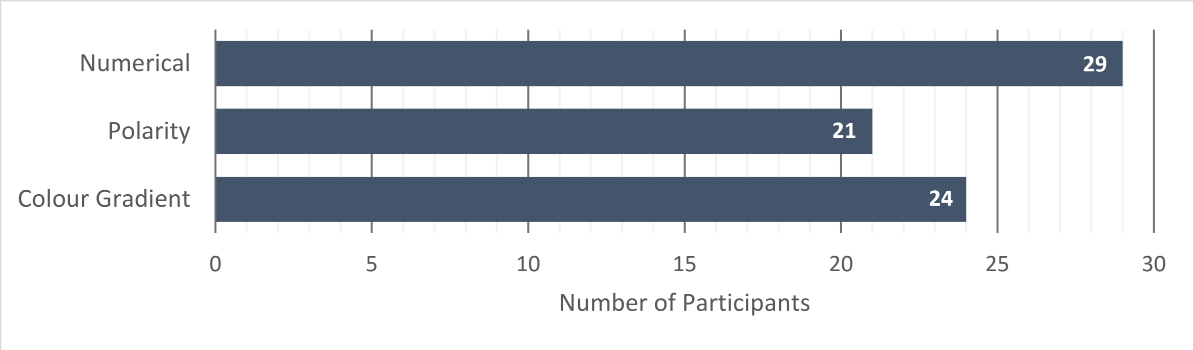

Figure 1: Distribution of participants' preference to numerical, polarity, and colour gradient rating scales

Numerical Rating Scale

Participants who preferred the numerical rating scale cited their ability to easily enumerate the twenty-one divisions of the NASA TLX questionnaire with the numerical rating scale in contrast to polarity and colour gradient rating scales. However, participants did not cite similar ease or difficulty in gauging the number of divisions in the SART questionnaire. This finding is consistent with Miller's (1956) and Preston & Colman's (2000) suggestions that people cannot provide more information than seven divisions on a scale. The numerical rating scale also directed participants to select divisions denoted by numbers in contrast to selecting intermediate regions of divisions in polarity and colour gradient rating scales as shown in Figure 2 below.

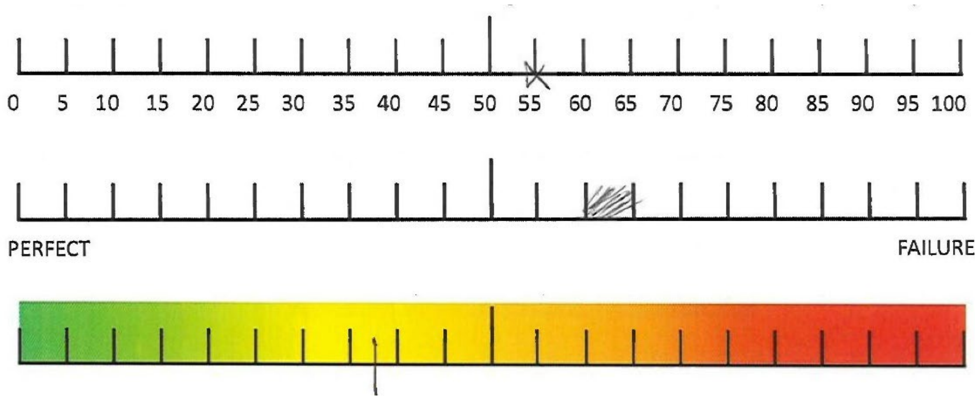

Figure 2: Types of response obtained on a NASA TLX rating scale for numerical (top), polarity (centre) and colour gradient (bottom) rating scales.

Polarity Rating Scale

Participants who preferred the polarity rating scale cited their inability to comprehend the polarity of numerical and colour gradient scales as the reason for their preference. The colour red in the colour gradient rating scales had ambiguous representations for an extreme polarity, either high or low, but neither was evident unless explicitly stated. Similarly, participants cited ambiguity in the numbers 1 and 7 on the SART scale. Unless explicitly stated, the number 1 could mean high in the ranking perspective, whilst the number 7 could also mean high by virtue of its value. Due to the absence of polarity in the numerical and colour gradient questionnaires, participants resorted to implementing their polarity on the rating scales as shown in Figure 3 below.

Figure 3: Examples of colour gradient and numerical rating scales where participants implemented their polarity

Colour Gradient Scale

Despite having the same number of divisions as the numerical and polarity rating scales, the colour gradient scale provided participants with a perception of infinite divisions. Participants who preferred the colour gradient rating scale cited that they were able to report their perceived mental workload and situation awareness scores with more certainty than the other two scales. This finding is consistent with the findings of Matell & Jacoby (1972) where the certainty in responses was linked to the greater number of divisions on a rating scale.

Conclusions

The three types of rating scales examined in two questionnaire designs provided varied insights on the design and subjective interpretation of rating scales. The feedback obtained on the original rating scale used in the NASA TLX questionnaire is consistent with previous studies suggesting that more than seven divisions are not ideal for a rating scale. However, the seven divisions used in the SART questionnaire encountered problems with the design of the original rating scale. The ambiguity with the subjective interpretation of numbers 1 and 7 prompted participants to adopt their polarity. The use of an experimental colour gradient on the rating scale enabled participants to assume infinite divisions and provide more certain responses. However, this was found to be problematic for the researchers since acquiring accurate empirical results from the rating scale requires precise selection of a division present on the rating scale. This problem can be addressed by digitising the analysis of the rating scale or the overall usage of the questionnaire.

In conclusion, the use of numbers and colours in a rating scale is superficial due to subjective interpretation of the numbers and colours. An ideal rating scale used in a pen-and-paper questionnaire would consist of numerical divisions in combination with an assignment of polarity on the ends of the rating scale.

References

Andrews, F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly,* 48.

Brosnan, K., Grün, B., & Dolnicar, S. (2021). Cognitive load reduction strategies in questionnaire design. *International Journal of Market Research*, 63(2), 125–133.

Cox, E. P. (1980). The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research*, 17(4), 407.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*, 52, 139–183. North-Holland.

Korek, W. T., Li, W.-C., Lu, L., & Lone, M. (2022). Investigating Pilots' Operational Behaviours While Interacting with Different Types of Inceptors. In D. Harris & W.-C. Li (Eds.), *Engineering Psychology and Cognitive Ergonomics*: Vol. LNAI 13307 (314–325). HCI International 2022.

Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology*, 56(6), 506–509.

Miller, G. A. (1956). The Magical Number Seven Plus or Minus Two - Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63, 81–97.

Miranda, N., Rukasha, T., & Korek, W. T. (2024). *Response to Acoustic Sounds and Synthesized Speech in an Automated Cockpit Checklist*. Manuscript submitted for publication.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1), 1–15.

Rodgers, W. L., & Herzog, A. R. (1992). Quality of Survey Measures: A Structural Modelling Approach. *Journal of Official Statistics*, 8(3), 251-275.

Roscoe, A. H., & Ellis, G. A. (1990). *A Subjective Rating Scale for Assessing Pilot Workload in Flight: A decade of Practical Use*. Royal Aerospace Establishment.

Saris, W. E., & Gallhofer, I. N. (2007). Design, Evaluation, and Analysis of Questionnaires for Survey Research. Wiley.

Taylor, R. M. (2017). Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. *Situational Awareness*, 111–128. Routledge.

Trujillo, A. (2009). Paper to Electronic Questionnaires: Effects on Structured Questionnaire Forms In: Jacko, J.A. (eds). *Lecture Notes in Computer Science*, vol 5610. Springer, Berlin, Heidelberg. HCI International 2009

Appendix 1: Scaled-down NASA TLX (left) and SART (right) questionnaires with numerical rating scales

NASA Task Load Index

MENTAL DEMAND: *How mentally demanding was the task?*

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

PHYSICAL DEMAND: *How physically demanding was the task?*

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

TEMPORAL DEMAND: *How hurried or rushed was the pace of the task?*

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

PERFORMANCE: *How successful were you in accomplishing what you were asked to do?*

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

EFFORT: *How hard did you have to work to accomplish your level of performance?*

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

FRUSTRATION: *How insecure, discouraged, irritated, stressed, and annoyed were you?*

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

SITUATION AWARENESS RATING TECHNIQUE (SART)

Instability of Situation: How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?

1 2 3 4 5 6 7

Complexity of Situation: How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?

1 2 3 4 5 6 7

Variability of Situation: How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)?

1 2 3 4 5 6 7

Arousal: How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?

1 2 3 4 5 6 7

Concentration of Attention: How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

1 2 3 4 5 6 7

Division of Attention: How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

1 2 3 4 5 6 7

Spare Mental Capacity: How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?

1 2 3 4 5 6 7

Information Quantity: How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?

1 2 3 4 5 6 7

Familiarity with Situation: How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?

1 2 3 4 5 6 7

Appendix 2: Scaled-down NASA TLX (left) and SART (right) questionnaires with polarity rating scales

NASA Task Load Index

MENTAL DEMAND: *How mentally demanding was the task?*

VERY LOW — VERY HIGH

PHYSICAL DEMAND: *How physically demanding was the task?*

VERY LOW — VERY HIGH

TEMPORAL DEMAND: *How hurried or rushed was the pace of the task?*

VERY LOW — VERY HIGH

PERFORMANCE: *How successful were you in accomplishing what you were asked to do?*

PERFECT — FAILURE

EFFORT: *How hard did you have to work to accomplish your level of performance?*

VERY LOW — VERY HIGH

FRUSTRATION: *How insecure, discouraged, irritated, stressed, and annoyed were you?*

VERY LOW — VERY HIGH

SITUATION AWARENESS RATING TECHNIQUE (SART)

Instability of Situation: How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?

Low — High

Complexity of Situation: How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?

Low — High

Variability of Situation: How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)?

Low — High

Arousal: How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?

Low — High

Concentration of Attention: How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

Low — High

Division of Attention: How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

Low — High

Spare Mental Capacity: How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?

Low — High

Information Quantity: How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?

Low — High

Familiarity with Situation: How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?

Low — High

Appendix 3: Scaled-down NASA TLX (left) and SART (right) questionnaires with colour gradient rating scales

NASA Task Load Index

MENTAL DEMAND: *How mentally demanding was the task?*

PHYSICAL DEMAND: *How physically demanding was the task?*

TEMPORAL DEMAND: *How hurried or rushed was the pace of the task?*

PERFORMANCE: *How successful were you in accomplishing what you were asked to do?*

EFFORT: *How hard did you have to work to accomplish your level of performance?*

FRUSTRATION: *How insecure, discouraged, irritated, stressed, and annoyed were you?*

SITUATION AWARENESS RATING TECHNIQUE (SART)

Instability of Situation: How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?

Complexity of Situation: How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?

Variability of Situation: How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)?

Arousal: How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?

Concentration of Attention: How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

Division of Attention: How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

Spare Mental Capacity: How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?

Information Quantity: How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?

Familiarity with Situation: How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?