Comparing Interface Influence on Users with Varying Expertise

Joel S. Elson(✉), Gina S. Ligon, and Doug C. Derrick

University of Nebraska, Omaha, NE 68182, USA
jselson@unomaha.edu

Abstract. The design of a system interface can impact user judgements among expert and novice users alike. With information systems, fundamental design choices can either augment or distract individuals in identifying patterns of converging data points. The goal of this effort was to compare the influence of Likert and categorical type rating scales in a system used to guide analysts through a content analysis process. While these scales have been examined in the context of psychological assessment literature, little has been said about their impact on decision makers from a human computer interaction perspective. We conducted a laboratory experiment to explore the effect of using Likert and categorical scales in an intelligence assessment task using unstructured data. The dependent variables included (1) Likert versus categorical type scales and (2) analyst experience (novice versus expert). Results indicated that expert and novices both had greater confidence and more creative, accurate responses in the interface utilizing Likert decision scaling.

Keywords: Interface design · Likert and categorical scaling
Novice and expert decision makers

1 Introduction

The design of decision support system interfaces has been studied in a number of different contexts and historically has shown mixed results in regards to their overall effectiveness [1, 2]. When designing these interfaces, it is important to realize that fundamental design choices can either augment or distract individuals in identifying patterns of converging data points. One such design consideration is the use of Likert and categorical type rating scales. While these scales have been examined in the context of psychological assessment literature, little has been said about their impact on decision makers when implemented in decision support systems from a human computer interaction perspective.

Leadership and organizational psychology literature suggests that novices versus experts consume and think about information differently from unstructured data about leader decision-making [3]. Moreover, when examining unstructured data in general, novices attend to ancillary information such as contextual dates, locations, and time frames, while experts tend to extract principles and concepts [4, 5]. This may be because experts, having more robust mental models based on experiences, have more working memory freed up when processing information that may be considered foreign

F. F.-H. Nah and B. S. Xiao (Eds.): HCIBGO 2018, LNCS 10923, pp. 275–284, 2018.
https://doi.org/10.1007/978-3-319-91716-0_21

or dense to novices [6]. It is important that decision aids be designed in a way that allows individuals with differing levels of experience and decision-making styles (e.g., tolerance for ambiguity, need for cognition) to quickly and accurately recognize patterns across unstructured data sources [7]. Decision support systems can do this by scaffolding the way individuals look and analyze data, drawing attention to information that experts consider important in an analysis task.

The goal of this effort was to compare the influence of an interface design used to guide an analyst through a content analysis process. To do this, we compared two types of common interfaces in psychological scaling: (1) Likert and (2) categorical. For the present effort, the following research question (RQ) was assessed via an experimental design.

RQ: What type of interface (e.g., categorical versus Likert scale) is most effective to assess the cognitive lens of leaders from a distance?

We designed a controlled laboratory experiment, using neurophysiological instrumentation, to assess the degree to which two interfaces impact decision-makers in assessing and interpreting leader intent from unstructured data. The experimental data resulted in recommendations about potential design considerations that could be offered to reduce the cognitive load of individuals interpreting intelligence indicators and recognizing patterns in unstructured data about an individual's likely interpretation of deterrence messaging.

1.1 Likert and Categorical Scales

Likert-based scales were first introduced by Likert [8], and can vary from 3, 5, or more options. When items are constructed to form a scale (e.g., confidence scale), they manifest an interval scale of measurement. One benefit of this is that it allows for multivariate analyses of analyst assessments. From a decision-making standpoint, Likert-based aids generally result in greater satisfaction and perceived ease of use in a host of populations when making judgments about unstructured data [9–11].

Conversely, categorical scales, usually represented by a bivariate response option (e.g., present versus absent) can result in faster decision-making [12], but incorporating fewer attributes may also reduce confidence and validity of responses in general [13]. In addition, the type of scales utilized by an interface has not been investigated in a population of intelligence professionals charged with making assessments from unstructured data.

1.2 Expertise in Decision Making

Expertise has been considered an important characteristic when studying decision making and judgment evaluation [14]. Expertise includes those skills and knowledge that are requisite in performing a specialized task and is developed through training and prior experience. Research in this area has been mixed, with expertise leading contributing to better decision outcomes as well as being a factor resulting in mistakes [14]. This may be explained by the fact that experts often rely on heuristics, or mental shortcuts, that enable experts to make decisions quickly and with less cognitive effort in comparison to domain novices [15]. Domain experts have been shown to consume information differently than novices, in addition domain experts engage in different processes or strategies when problem solving. Familiarity with a decision aid itself has also been shown to impact system users, requiring mastery of using a decision aid before functional area knowledge of an expert can be applied to a specific problem or task [16].

2 Methodology

An experiment was developed to assess the differences between scaling options in a decision support system and compared how expert and novice decision makers varied in their assessment of a foreign adversary. The experiment required participants to read background information and speech excerpts from a fictional foreign government leader then analyze his decision-making style using either a Likert type or categorical decision aid.

2.1 Experiment Development

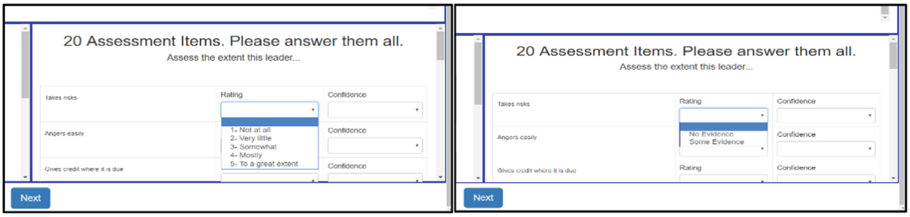
Experimental materials were all developed based on actual background information about a real foreign leader, as well as literal speaking excerpts taken at two different points in the given leader's history. One speech excerpt was taken after an interview six-months prior to the bid for a major international sporting event, which was classified as a relatively neutral period in the leader's history (i.e., unmarked by subsequent escalation). This speech was thus labeled the "Neutral Speech." Another speech was selected at the same time point (six months) preceding a military event in a neighboring country. Given then subsequent regional conflict that ensued, this speech was labeled the "Escalatory Speech." The order the participants saw these speeches was counterbalanced in order to minimize bias across participants (e.g., participants were randomly assigned to read and assess the escalatory versus the neutral speech first). No significant differences were found based on this ordering. The speeches were the exact same length.

Subject matter expert feedback was obtained for each of these speeches and the materials were refined to conceal the actual identity of the foreign leader. This was done to minimize a priori biases participants may have held about the leader on which the materials were based.

There were two methods in which we varied the visual arrays: (1) the type of unstructured data from which analysts drew conclusions about the effectiveness of deterrence, and (2) the scaling options that were provided to support the decision-making of the analyst making the assessment (Fig. 1).

2.2 Participants

To assess differences in visual array processing that might arise given level of experience, two samples representing expert and novice decision makers were recruited for this effort. Analysts were recruited from the Department of Defense (DoD) via an



20 Assessment Items. Please answer them all.
Assess the extent this leader...

Takes risks | Rating | Confidence
1- Not at all
2- Very little
3- Somewhat
4- Mostly
5- To a great extent
Angers easily | Confidence
Gives credit where it is due | Confidence

Next

20 Assessment Items. Please answer them all.
Assess the extent this leader...

Takes risks | Rating | Confidence
Angers easily | No Evidence
Some Evidence | Confidence
Gives credit where it is due | Rating | Confidence

Next

Fig. 1. Likert versus categorical scaling visual array format

email. This effort resulted in 43 individuals with varying levels of experience (a) assessing leader decision-making, (b) using all source intelligence, and (c) years working in the Department of Defense. Of these, 23 individuals reported having 10 or more years of analyst experience. The second sample was recruited from a participant pool at the University of Nebraska Omaha College of Business and consisted of both graduate and undergraduate students. The overwhelming majority of this sample had minimal or no experience (a) assessing leader decision-making, (b) using all source intelligence, or (c) years working in DoD. Thus, this second sample was used to represent novices. We considered those individuals with 10 or more years of experience to be experts, aligning with the definition of "expert" from Ericsson and Charness [17]. The following figure illustrates years of experience differences across the two samples (Fig. 2).

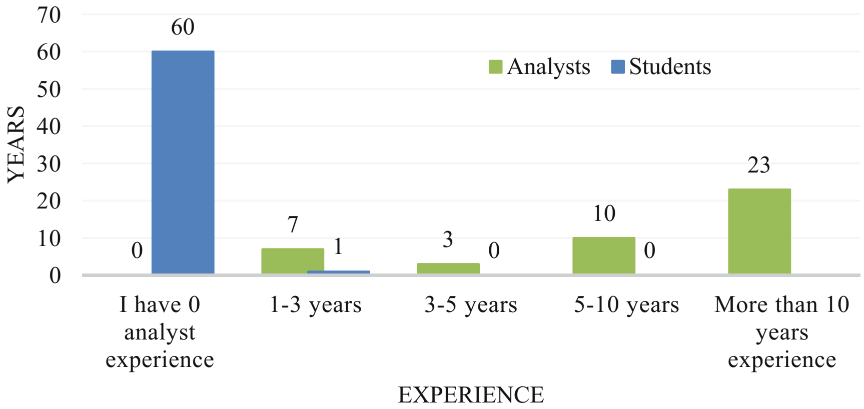


Fig. 2. Years of experience compared between analysts and students.

As previously mentioned, precautions were taken to conceal the identity of the foreign leader used in the development of the experimental task. Responses from participants indicated they were less familiar with exact backgrounds of leaders from this foreign country, and thus were unlikely to identify the characteristics and speeches of the actual identity of the leader.

2.3 Procedure

Participants first completed a battery of demographic information and the measures of problem solving style. Next, participants were assigned to one of four experimental conditions. Since the central research question of interest was to examine differences between two different interfaces among expert and novices, a 2 × 2 design was employed, varying decision-support array between conditions and type of unstructured data within condition.

To compare varying levels of expertise in assessing leader decision-making, two samples participated in an experiment at the University of Nebraska Omaha's Jack and Stephanie Koraleski Commerce and Applied Behavior Lab. Here, they participated in a series of individual difference measures (e.g., demographic survey about years of experience studying leaders from afar) and an experimental task. The process lasted anywhere from 45 min to 1.5 h, and all data was collected in the fall of 2017 on the university campus with IRB approval and was in compliance with the full HRPO human subjects' protection requirements.

In the experimental task itself, participants first read biographical information about the leader, and then answered four open-ended items about his decision-making style. After reading the biographical information, participants either first read the neutral or the escalatory speech excerpt and answered the same series of questions about each. Participants assessed the leader's decision-making style three times based on three types of unstructured data: (1) biographical data, (2) Neutral Speech, and (3) Escalatory Speech. Depending on whether they were assigned to the categorical or Likert condition, participants then completed 20-items related the attributes of the foreign leader. Specifically, participants judged, based solely on the background information available, the leader's likely cognitive lens through which he received and interpreted deterrence messages. For each of these items, participants also completed a confidence rating to provide some indication of the confidence they felt in their assessment.

2.4 Analysis

Prior to conducting analyses, several analytic steps were taken to prepare the data. First, open-ended responses to the three sets of analysis questions following the biographical data and speech excerpts were converted to quantitative scores to allow multivariate analyses. Four raters, unfamiliar with study hypotheses, were trained to assess responses on fluency, flexibility, complexity, and type of affect identified. Training lasted five hours, and raters achieved appropriate interrater reliability ($\alpha = 0.91$) across all scales.

For example, because participants were asked to provide a list of descriptors about the leader after reading through the stimulus materials, fluency was assessed by counting the total number of adjectives, while flexibility was assessed by counting the unique categories of adjectives for each participant (Table 1).

In addition, some index of accuracy of assessment was taken via the assessment of the responses for positive versus negative affect. Specifically, because the speech excerpts were selected during varying levels of escalatory activity, they manifested characteristics of positive versus negative affect. In the neutral speech, and particularly

Table 1. Fluency and flexibility scales.

	List as many adjectives as you can about the leader in this scenario…
Fluency (number of descriptors) = 8	Aggressive, dominant, powerful, commanding, commandeering, angry, experienced, bitter
Flexibility (number of different categories of descriptors) = 3	Aggressive, dominant, powerful, commanding, commandeering, angry, experienced, bitter

because the given leader was in a period of trying to influence the global community to host the Olympics in his country, the use of positive wording around decision-making far exceeded the use of negative wording. Conversely, prior to escalation activities of the invasion of Crimea, speech patterns turned more negative in affect, comparable to what has been seen in speeches of other leaders prior to escalation [18]. For this rating, a benchmark scale was developed to guide raters on assessing the primary Affect (or emotional valence) of a given response following either the neutral or the escalatory speech excerpt (Table 2).

Table 2. Affect benchmark scale.

Primary type of Affect of response. This is a 1–5 scale, use all points on the scale 1 (negative)–5 (positive)	
1	Response had all negative tone (e.g., annoyed, opportunistic, aggressive)
2	List was primarily negative in tone (e.g., hostile, angry, boasting)
3	Mix of negative and positive affect/balanced (e.g., powerful, defiant, fair, competitive)
4	List was primarily positive in tone (e.g., passionate, hard-working, achieving, angry)
5	Response had all positive tone (e.g., passionate, determined, strong)

Finally, responses were assessed for degree of complexity, or the amount of abstraction participants were able to complete based on the information available. Because decision support aids and visual arrays were meant to increase the analysts' capacity for abstraction to the cognitive lens of deterrence of a given leader, an assessment of participants capacity for abstraction was assessed using a 5-point behaviorally anchored benchmark scale. Particularly, since some leaders can perceive deterrence messaging as a direct affront to them personally [18], we assessed participant response to the item "Describe this leader's decision-making style when faced with what he perceives a personal betrayal" for complexity and capacity for abstraction to variables related to deterrence (Table 3).

3 Results

The guiding research question behind this work looked to see what type of interface scale (e.g. categorical verses Likert type) would be most effective to assess the cognitive lens of a foreign leader. We specifically looked at three facets of an analysts

Table 3. Complexity benchmark scale.

Degree of complexity/chunking of response. This is a 1–5 scale, use all points on the scale	
1–2	Response was organized by superficial groupings/characteristics; re-iterated only cues in the prompt; perceptually salient descriptors rather than abstract or complex ones *Participant Example: Angry*
3	Moderate level of Complexity; re-iterated some of the cues in the prompt, but also added new ones; ideas are moderately complex and convey multiple meanings with at least 1 word/concept *Participant Example: Leader would likely get mad and try punishing the personnel who betrayed him. Seeing as he has been in five physical altercations growing up he is quick-tempered so he may try fighting*
4	Response was organized by abstract/complex groupings/characteristics; conceptually combined 2 or more concepts; ideas are complex and convey multiple meanings with at least 2 words/concepts *Participant Example: Based on his rapid advancement in the KGB, this leader probably understands a certain level of restraint is necessary, but a clear message must be sent. He will likely take action based upon the severity of the slight against him and the organization, while calculating the perceived loyalty and value of the subordinate or person who has betrayed him. Those with potential and value will likely face less reprimand than someone who is a threat or a persistent problem to him. He does not want to make the punishment too overt or over-the-top, as this could draw negative attention and ire towards him from his superiors, whom he consistently seeks to please*
5	Response had all positive tone (e.g., passionate, determined, strong)

assessment: confidence in, complexity, and accuracy of assessments. The results of this endeavor found that interfaces utilizing Likert scaling outperformed categorical scaling interfaces on nearly every metric.

3.1 Confidence in Assessments

Decision confidence was first compared between individuals who used the Likert-type scaling (the Likert group, $N = 43$) and those who used the categorical scaling (the categorical group, $N = 49$). The two groups were comprised of individuals with both high experience (greater than 10 years of experience) and low experience (less than 10 years of experience). Confidence scores from the 72 items were summed to generate a total confidence score for each participant. A Welch's t-test was ran to compare the average confidence score between the two groups. There was a significant difference in the scores for the Likert ($M = 232.33$, $SD = 22.52$) and categorical ($M = 209.86$, $SD = 29.16$) conditions; ($F(1, 90) = 16.75$, $p < 0.001$). This means that individuals who used the Likert-type interface were more confident in their decision on average than individuals who used the categorical interface (Table 4).

Confidence in the decision task was also assessed between high experience ($N = 19$) and low experience ($N = 73$) raters, each group was comprised of individuals who utilized both the Likert-type and categorical type interfaces. A Welch's t-test was on the summed confidence scores for each individual. There was not a significant difference in the scores for the high experience group ($M = 218.16$, $SD = 30.04$) and

Table 4. Confidence by Likert vs categorical.

Type of question	Analyst mean (SD)	Student mean (SD)
Categorical	209.44 (21.92)	203.80 (32.75)
Likert	227.00 (28.83)	235.92 (14.38)

the low experience group (M = 220.93, SD = 28.22); (F (1, 90) = 0.14, p = 0.707). This means that there was no significant difference in decision confidence scores between individuals with high compared to low experience.

Confidence scores for users of the Likert-type interface were examined at the macro level to compare differences between users with high experience and low experience. There was not a significant difference in the scores for the high experience group (M = 226.92, SD = 25.76) and the low experience group (M = 234.42, SD = 21.23); (F (1, 41) = 0.96, p = 0.333). This means that for users of the Likert-type interface only; there was no significant difference in decision confidence scores between individuals with high experience compared to low experience.

Confidence scores for users of the categorical type interface were examined at the macro level to compare differences between users with high experience and low experience. There was not a significant difference in the scores for the high experience group (M = 203.14, SD = 32.76) and the low experience group (M = 210.98, SD = 28.80); (F (1, 47) = 0.43, p = 0.516). This means that for users of the categorical interface only, there was no significant difference in decision confidence scores between individuals with high experience compared to low experience.

Confidence scores for individuals with high previous experience were examined between users of the Likert-type interface and the categorical interface. There was not a significant difference in the scores for the Likert-type condition (M = 226.92, SD = 25.76) and the categorical type condition (M = 203.14, SD = 32.76); (F (1, 18) = 3.09, p = 0.097). This score is approaching significance and it is expected that with an increased sample size, users would have reported being more confident in the Likert-type condition.

Confidence scores for individuals with low previous experience were examined between users of the Likert-type interface and the categorical interface. There was a significant difference in the scores for the Likert-type condition (M = 234.42, SD = 21.23) and the categorical type condition (M = 210.98, SD = 28.80); (F (1, 71) = 14.64, p < 0.001). This means that individuals with low previous experience, those who used the Likert-type interface were more confident in their decision on average than individuals who used the categorical interface.

3.2 Complexity of Assessments

Complexity of assessments, as manifested by the nuanced interpretation of speech excerpts, varied between analysts and students, and also across speech type (escalatory versus neutral). However, the type of interface did not result in varied complexity of assessments. Analysts on average produced more complex responses when compared to students, but the type of decision-making aid provided did not impact the complexity of either student or analyst assessments (Table 5).

Table 5. Complexity by categorical vs. Likert (escalatory speech).

Type of question	Analyst mean (SD)	Student mean (SD)
Categorical	2.50 (1.41)	2.20 (1.13)
Likert	3.00 (1.36)	2.21 (1.01)

3.3 Accuracy of Assessments

While there are many metrics that could speak to the overall "accuracy" of the assessment, given the nature of the research question (what type of decision-making scale impacts accuracy of assessments), for the present effort we selected the degree to which participants could discern the affect manifested by a particular speech excerpt. Moreover, since one reliable indicator of a leader escalating aggression is his increased use of negative language (e.g., verbs, references to past grievances), participants' assessments in this study were assessed by the inference they were able to make about the effect, or the valence of the emotional imagery conveyed, identified in either the escalatory or the neutral speech. Through using this process, both the analyst and the student participants were able to accurately assess the affect in a given speech. Moreover, descriptors of the leader following the neutral speech—a time when the leader discussed engagement in the global community and hope for growth—were far more positive in affect (analyst M = 3.09, SD = 1.49 versus M = 4.83, SD = 0.5) in the categorical condition; (M = 3.18, SD = 1.66 versus M = 4.95, SD = 1.16) in the Likert condition.

4 Conclusion

Individuals who used Likert-style interface were more confident in their assessments. As user adoption of a new requirement (e.g., intent assessment) is an important element of training motivation, using this type of interface is recommended going forward for both experts and novices. Having the gradient-type response options may allow for great comfort when making decisions based on incomplete, unstructured raw data (e.g., speech excerpts). In addition, because assessing elements of a cognitive lens requires some comfort with ambiguity as well as a tenacious problem-solving style, using a priori categories of variables likely related to message interpretation can aid individuals in making connections between seemingly disparate data points.

When training both novices and experts, use of Likert-scaling interfaces to support analyst confidence, tolerance for ambiguity, and accuracy in assessments. Given that both analysts and students had greater confidence and more creative, accurate responses in the Likert interface conditions, it is not recommended that different scaling options be provided depending on the characteristics of the individual using them.

While the type of scaling appears to have less impact on judgments, it may be due to the nature of the speeches selected. Moreover, in pilot tests, speeches with the clearest differences were selected for the experimental materials. By doing this, we may have inadvertently made this component of the task too "clear" and less structured, possibly limiting the utility of the decision-making aids. However, more tests need to be run to vary the nature of the ambiguity in the speech presented in order to see if this is indeed a covariate.

References

1. Sharda, R., Barr, S.H., MCDonnell, J.C.: Decision support system effectiveness: a review and an empirical test. Manag. Sci. **34**, 139–159 (1988)
2. Benbasat, I., Nault, B.R.: An evaluation of empirical research in managerial support systems. Decis. Support Syst. **6**, 203–226 (1990)
3. Ligon, G.S., Douglas, D.C., Elson, J.S., Mazgaj, M., d'Amato, A., O'Malley, D., Robinson, S.: Intelligence support to deterrence operations, Omaha, NE (2017)
4. Chase, W.G.: Visual Information Processing. Academic Press, New York (1973)
5. Larkin, J., McDermott, J., Simon, D.P., Simon, H.A.: Expert and novice performance in solving physics problems. Science **208**, 1335–1342 (1980)
6. Ericsson, K.A., Kirk, E.P.: The search for fixed generalizable limits of "pure STM" capacity: problems with theoretical proposals based on independent chunks. Behav. Brain Sci. **24**, 120–121 (2001)
7. Hermann, M.G.: Content analysis. In: Qualitative Methods in International Relations, pp. 151–167. Palgrave Macmillan, London (2008)
8. Likert, R.: A technique for the measurement of attitudes. Arch. Psychol. **22**(140), 55 (1932)
9. Boone, H.N., Boone, D.A.: Analyzing Likert data. J. Ext. **50**, 1–5 (2012)
10. van Laerhoven, H., van der Zaag-Loonen, H., Bhf, D.: A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. Acta Paediatr. **93**, 830–835 (2004)
11. Lee, J.W., Jones, P.S., Mineyama, Y., Zhang, X.E.: Cultural differences in responses to a Likert scale. Res. Nurs. Health **25**, 295–306 (2002)
12. Redelmeier, D.A., Heller, D.N.: Time preference in medical decision making and cost-effectiveness analysis. Med. Decis. Mak. **13**, 212–217 (1993)
13. de Bekker-Grob, E.W., Ryan, M., Gerard, K.: Discrete choice experiments in health economics: a review of the literature. Health Econ. **21**, 145–172 (2012)
14. Shanteau, J., Stewart, T.R.: Why study expert decision making? Some historical perspectives and comments. Organ. Behav. Hum. Decis. Process. **53**, 95 (1992)
15. Logan, G.D.: Skill and automaticity: relations, implications, and future directions. Can. J. Psychol. Can. Psychol. **39**, 367 (1985)
16. Mackay, J.M., Elam, J.J.: A comparative study of how experts and novices use a decision aid to solve problems in complex knowledge domains. Inf. Syst. Res. **3**, 150–172 (1992)
17. Ericsson, K.A., Charness, N.: Expert performance: its structure and acquisition. Am. Psychol. **49**, 725 (1994)
18. Hermann, M.G.: Assessing leadership style: a trait analysis. Psychol. Assess. Polit. Lead. 178–212 (2005)