A Proposal for New Subjective Evaluation Analysis Method using the Visual Analog Scale

Naruki Shirahama^{a*}, Satoshi Watanabe^b, Naofumi Nakaya^c and Yukio Mori^d

^aDepartment of Creative Engineering, Information and Systems Engineering Course
National Institute of Technology, Kitakyushu College

5-20-1, Shii, Kokuraminami-ku, Kitakyushu-shi, Fukuoka 802-0985, Japan

^bSchool of Health Sciences at Odawara, International University of Health and Welfare
1-2-25, Shiroyama, Odawara-shi, Kanagawa 250-8588, Japan

^cFaculty of Health Sciences, Tsukuba International University
6-20-1 Manabe Tsuchiura, Ibaraki 300-0051, Japan

^dDepartment of Mechanical and Electronic Engineering, Salesian Polytechnic
4-6-8, Oyamagaoka, Machida-shi, Tokyo 194-0215, Japan

*Corresponding Author: naruki@kct.ac.jp

Abstract

We propose a new subjective evaluation experiment and analysis method. By adopting VAS, it became possible to perform statistical analysis to comprehend the overall tendency even in subjective evaluation experiment of a small sample. The other feature is to visualize the experimental results by overlaying a box-whisker plot and a univariate scatter plot. We recommend that it is essential to grasp the overall trend by showing all the data when the number of samples is small. We experimented with 20 subjects on conversations with communication toy robot. We investigated the user's emotions from the conversation with a robot, and whether it relates to the smoothness of the conversation. Impressive results were obtained that the question on negative emotions showed the opposite trend from expectation.

Keywords: visual analog scale, small sample, box-and-whisker plot, univariate scatter plot, histogram, subjective evaluation.

1. Introduction

We have proposed a new analysis method for subjective evaluation experiment⁽¹⁻⁴⁾ for several years. We described how to analyze experimental results on how the conversation with the communication robot affects user's emotions or feelings⁽⁵⁾. We proposed a concept of a new experimental method, which has three features, for

subjective evaluation of human sensitivity. The first feature is to set redundant questions for questionnaires to subjects. We set two similar questions in our previous studies. For example, one is the value on the interval $[0, 1]$, the other is the value on the two pole attributes. Adopting continuous quantity allows us to determine if the object has nonlinear features. The second feature is to adopt Visual Analog Scale (VAS). Likert scale is an order scale, especially for small samples, it is not suitable for statistical processing. In addition, we solved the problem of troublesome counting of VAS by using VAS measurement application called VASpad^(4,5).

The third feature is to visualize the experimental results by overlaying two graphs, a box-whisker plot and a univariate scatter plot⁽⁵⁾. Weissgerber et al.⁽⁶⁾ recommend using univariate scatter plot instead of bar graphs or line charts to visualize data when the number of samples is small. They proposed to show the full distribution of the data for small sample size studies. From our experiences, we have found that there are many outliers in subjective evaluation experiments, and most cases are small samples. That is why there is no guarantee that subjective information will follow a normal distribution without outliers. We also pay attention to this point and propose to show the experiment results by superimposing the box-and-whisker plot and the univariate scatter chart. We examine the effectiveness of this proposed method in this paper. We conducted the same experiments as before for different subjects.

We also show a histogram showing the trend of the whole data.

2. Subjective Evaluation Experiment

Here is a brief description of the outline of VAS. VAS is a method of measuring the degree of subjectivity by himself on a horizontal straight line. We use the VAS as a psychometric response scale used in questionnaires. It is a better measurement instrument for subjective information that cannot be measured directly.

Fig. 1. Comparative image of VAS and LS

At first, we set the minimum and maximum conditions at both ends with a horizontal line of 10 cm. Subjects put a mark on the line. We measure its length from 0 with a ruler and calculates its value. Likert Scale (LS) is often used in general questionnaire. However, we adopted the VAS. We can express user's subjectivity directly and can confirm the trend from the whole distribution even when the number of samples is small. Figure 1 shows a sample of VAS and LS. The upper side shows VAS, and the lower side shows LS. We can understand intuitively that it is hard to find trends from whole distribution in small samples when LS is used.

3. Results and discussion

We experimented with 20 subjects on conversations with communication toy robot. Because this robot is inexpensive and toy level performance, the conversation often does not succeed well. Focusing on this point, we investigated the user's emotions from the conversation with a robot, and whether it relates to the smoothness of the conversation. The communication robot we used for the experiment is OHaNAS of TAKARA TOMY. Subjects talks to OHaNAS for about 5 to 10 minutes. Figure 2 shows ten questions on the impression from the conversation. Q1, Q4, Q5, Q6 have categorized the performance evaluation of OHaNAS. Q2 and Q3 are two pole emotional attributes such as "Happy-Sad."

- Q1. Do you think that the flow of conversation was smooth? (Yes - No)
- Q2. What kind of emotion has been occurred? (Happy - Sadness)
- Q3. What kind of emotion has been occurred? (Joy - Anger)
- Q4. Do you think that OHaNAS could understand Japanese well? (Agree - Disagree)
- Q5. Do you think OHaNAS could speak Japanese well? (Agree - Disagree)
- Q6. Which do you think better OHaNAS's performance? (Listening / Speaking)
- Q7. How much did you become happy after the conversation? [0% - 100%]
- Q8. How much did you become sad after the conversation? [0% - 100%]
- Q9. How much did you become pleased after the conversation? [0% - 100%]
- Q10. How much did you become anger after the conversation? [0% - 100%]

Fig. 2. Ten questions about conversation with OHaNAS

Table. 1. Measured values (case 1)

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
S1	0.46	0.64	0.49	0.52	0.44	0.71	0.56	0.00	0.58	0.00
S2	0.80	0.52	0.41	0.81	1.00	1.00	0.43	0.00	0.47	0.73
S3	0.51	0.87	0.51	0.51	0.53	1.00	0.75	0.01	0.50	0.00
S4	0.64	0.67	0.71	0.39	0.62	0.77	0.71	0.17	0.68	0.00
S5	0.70	0.53	0.39	0.93	0.83	0.41	0.34	0.00	0.15	0.47
S6	0.79	0.49	0.77	0.77	0.81	0.94	0.73	0.03	0.80	0.02
S7	0.83	0.66	0.63	0.81	0.75	0.26	0.68	0.00	0.73	0.00
S8	0.51	0.53	0.50	0.45	0.71	0.74	0.04	0.17	0.02	0.01
S9	0.60	0.48	0.19	0.44	0.97	1.00	0.70	0.28	0.47	0.76
S10	0.96	0.52	0.64	0.82	0.90	0.93	0.58	0.01	0.19	0.01
S11	1.00	0.73	0.81	0.82	0.99	1.00	0.19	0.00	0.42	0.00
S12	0.76	0.63	0.58	0.91	0.93	0.61	0.64	0.16	0.76	0.03
S13	1.00	0.37	0.01	1.00	0.35	0.52	0.27	0.69	0.27	1.00
S14	0.79	0.89	0.80	0.70	0.94	0.78	0.68	0.05	0.62	0.06
S15	1.00	0.77	0.78	1.00	1.00	1.00	0.50	0.00	0.47	0.00
S16	0.92	0.21	0.27	0.90	1.00	0.99	0.33	0.53	0.54	0.03
S17	0.82	0.87	0.92	0.35	0.91	0.73	0.69	0.00	0.44	0.00
S18	0.17	0.31	0.25	0.36	0.82	0.14	0.31	0.73	0.41	0.82
S19	0.90	0.72	0.61	0.85	0.91	0.33	0.82	0.45	0.68	0.20
S20	0.74	0.84	0.82	0.49	0.75	0.85	0.73	0.08	0.81	0.00

The four questions Q7 to Q10 turn the evaluation value on the interval [0, 1] for the two pole attribute of Q2 and Q3.

We analyze from the viewpoint of differences in questions. Table 1 shows the measured values obtained from smooth conversations and table 2 shows the measured values obtained from the non-smooth conversation.

Table 2. Measured values (case 2)

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
S1	0.36	0.39	0.28	0.21	0.53	0.74	0.12	0.20	0.23	0.36
S2	0.34	0.52	0.18	0.14	0.93	1.00	0.09	0.63	0.11	0.72
S3	0.21	0.49	0.22	0.00	0.80	1.00	0.52	0.00	0.19	0.53
S4	0.18	0.32	0.51	0.14	0.79	0.86	0.79	0.15	0.81	0.00
S5	0.64	0.51	0.39	0.89	0.58	0.45	0.27	0.04	0.18	0.36
S6	0.05	0.49	0.53	0.27	0.77	0.86	0.73	0.04	0.66	0.04
S7	0.01	0.30	0.30	0.60	0.45	0.11	0.33	0.19	0.14	0.01
S8	0.28	0.41	0.39	0.38	0.47	0.58	0.01	0.10	0.01	0.01
S9	0.20	0.48	0.15	0.10	0.84	0.68	0.17	0.25	0.00	0.83
S10	0.06	0.50	0.19	0.09	0.07	0.92	0.08	0.02	0.00	0.68
S11	0.00	0.50	0.34	0.17	0.20	1.00	0.00	0.00	0.23	0.19
S12	0.30	0.61	0.30	0.27	0.83	0.93	0.51	0.11	0.51	0.04
S13	0.20	0.15	0.51	0.24	0.81	0.79	0.40	0.30	0.30	0.75
S14	0.54	0.56	0.76	0.38	0.71	0.85	0.83	0.14	0.75	0.22
S15	0.21	0.39	0.40	0.00	1.00	1.00	0.08	0.00	0.21	0.13
S16	0.56	0.14	0.29	0.50	0.78	0.97	0.11	0.15	0.72	0.05
S17	0.26	0.40	0.49	0.31	0.44	0.66	0.32	0.27	0.27	0.19
S18	0.16	0.23	0.16	0.00	0.74	0.26	0.00	0.88	0.10	0.86
S19	0.18	0.30	0.11	0.33	0.84	0.92	0.23	0.92	0.20	0.81
S20	0.36	0.35	0.37	0.31	0.38	0.32	0.30	0.40	0.52	0.72

Table 3. Fundamental Statistics (case 1)

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
max	1.00	0.89	0.92	1.00	1.00	1.00	0.82	0.73	0.81	1.00
min	0.17	0.21	0.01	0.35	0.35	0.14	0.04	0.00	0.02	0.00
mean	0.75	0.61	0.55	0.69	0.81	0.74	0.53	0.17	0.50	0.21
median	0.79	0.64	0.60	0.79	0.87	0.78	0.61	0.04	0.49	0.02
var	0.05	0.04	0.06	0.05	0.04	0.07	0.05	0.06	0.05	0.12
sd	0.21	0.19	0.24	0.23	0.19	0.27	0.22	0.24	0.22	0.34

Table 4. Fundamental Statistics (case 2)

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
max	0.64	0.61	0.76	0.89	1.00	1.00	0.83	0.92	0.81	0.86
min	0.00	0.14	0.11	0.00	0.07	0.11	0.00	0.00	0.00	0.00
mean	0.26	0.40	0.34	0.27	0.65	0.75	0.29	0.24	0.31	0.38
median	0.21	0.41	0.32	0.26	0.76	0.86	0.25	0.15	0.22	0.29
var	0.03	0.02	0.03	0.05	0.06	0.07	0.07	0.07	0.07	0.11
sd	0.18	0.13	0.16	0.22	0.25	0.27	0.26	0.27	0.26	0.33

For the sake of simplicity, we refer to a smooth conversation as “case 1” and a non-smooth conversation as “case 2”. The subjects are the same 20 people in all cases. “Q1” to “Q10” in the first row of the table represents the question number, and “S1” to “S20” in the first column of the table represent subjects, respectively.

All values are within the interval of 0 to 1, rounded to two decimal places. Table 3 and table 4 show the fundamental statistics of table 1 and table 2, respectively. It turns out that the average value shows different trends for each question from the tables. Both standard deviation of Q6, Q8, Q10 is slightly large.

Fig. 3. Box whisker and univariate scatter plot (all)

Fig. 4. Box whisker and univariate scatter plot (case 1)

Fig. 5. Box whisker and univariate scatter plot (case 2)

Next, in order to grasp the tendency of the distribution as a whole, overlay of box-and-whisker plot and univariate scatter plot is shown in figure 3. The data of case 1 is drawn in light blue, and the data of case 2 is drawn in bright red. The thick line inside the box is the median, the upper side of the box is the upper quartile, and the lower side of the box is the lower quartile. In order to make it easier to see the feature, figure 4 shows only the data of case 1, and figure 5 shows only the data of case 2.

Fig. 6. Histogram for each question (case 1)

Fig. 7. Histogram for each question (case 2)

The upper and lower lines outside of the box are connected by the dotted line from the position of the upper (lower) quartile to the maxima (minima), respectively. Small circles represent outliers.

The histogram is also a graph that grasps the tendency of the entire distribution. Figure 6 shows the histogram of case 1, and figure 7 shows the histogram of case 2.

We predicted the tendency of answers for each question in advance before the experiment. Figure 8 shows how the ten questions shown in figure 2 are classified. We anticipated that the measurements show opposite trends when the conversation is smooth and not smooth.

The trends of Q1, Q2, Q3, Q4, Q5, Q7 and Q9 supported the prediction. However, Q8 and Q10 did not apply to expectations. Q8 is a question about sadness, Q10 is a question about anger. Impressive results were obtained that the question on these negative emotions showed the opposite trend from expectation. Furthermore, from these results, it is shown that those who feel sadness are less than those who feel angry even if the conversation goes wrong.

Q6 is also interesting. Regardless of the smoothness of the conversation, many subjects feel that this robot has better speaking than listening. This can be seen from the difference in tendency between Q4 and Q5.

I, the histogram is also valid for grasping the overall trend. It turns out that Q8 and Q10 show an opposite tendency to the others only at a glance.

For further reference, we also show results of cluster analysis for each question in the same way as in previous research⁽⁵⁾. Figure 9 and 10 show cluster dendrogram of case 1 and case 2 by the Ward method. Q6, Q8, and Q10 have the same classification as the prediction in case 1. It is also very interesting that the remaining question groups are divided into those related to conversation performance and those related to feelings received.

4. Conclusions

In order to verify the effectiveness of the proposed method, we conducted a subjective evaluation experiment on the conversation with the communication robot for different subjects by the same method as the previous research. The difference from the experimental result of last year is Q8 in case 2. When the conversation was not smooth, many subjects felt sadness the previous year; however this time it was less. Regarding the result of clustering, we confirmed that case 1 is almost the same as the previous result, but in case 2 there is a difference. As a general trend, the analysis

result is similar to the previous one, but we could confirm such a difference. To establish a subjective evaluation experiment method in a small sample, we continue our research in the future.

Q1, Q2, Q3, Q4, Q5, Q7 and Q9 show high values when the conversation is smooth and low values in the non-smooth conversation.

1. Q8 and Q10 show high values when the conversation is non-smooth and show low value when the conversation is smooth.
2. Q6 shows various values.

Fig. 8. Trends of measurement values assumed before experiment

Fig. 9. Cluster dendrogram (case 1)

Fig. 10. Cluster dendrogram (case 2)

Acknowledgment

This work is partially supported by JSPS Grants-in-Aid for Scientific Research, No.25350950 and No.16K01890.

References

- (1) Watanabe, S., Takaue, R., Fenghui Yao, Matsumoto, Y., Tsukamoto, H., Shirahama, N., Nakaya, N. and Mori, Y., "Effects of 1/f Fluctuation Music Listening on Autonomic Nervous System Activity", Journal of the Institute of Industrial Applications Engineers Vol. 6 No.2, pp.86-91, 2018.
- (2) Watanabe, S., Shirahama, N., Tsukamoto, H., Matsumoto, Y., Nakagawa, M., Miyamoto, K., Nakaya, N., Tomita, M. and Mori, Y.: "Trend of the Subjective Evaluation Based on Visual Analog Scale and Likert Scale", ICIC Express Letters, Part B: Applications (ICIC-ELB), Volume 7, No.1, pp.37-42, 2016.
- (3) Watanabe, S., Shirahama, N., Matsumoto, Y., Tsukamoto, H., Nakagawa, M., Miyamoto, K., Nakaya, N., Tomita, M. and Mori, Y.: "Distribution Trend of the Information Skills for Students Based on Visual Analog Scale and Likert Scale", International Journal of Biomedical Soft Computing and Human Sciences, vol. 20, no.2, pp.1-6, 2016.
- (4) Shirahama, N., Watanabe, S., Ikegami, F. and Mori Y.: "Development of VAS app to Improve Youth Mental Health Research Environment", ICIC Express Letters, Part B: Applications (ICIC-ELB), Volume 7, No.9, pp.1939-1944, 2016.
- (5) Shirahama, N., Watanabe, S., Ikegami, F., Nakaya N. and Mori, Y. "A Study on the Impression Received from the Response of a Communication Robot", Proceedings of the 5th IIAE International Conference on Intelligent Systems and Image Processing 2017 (ICISIP2017) pp.335-338, 2017.
- (6) Weissgerber TL, Milic NM, Winham SJ and Garovic VD, "Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm", PLoS Biol Vol. 13, No.4: e1002128. doi:10.1371/journal.pbio.1002128, 2015.