【Original Article】

A New Type of Screening Necessity We Face with Online Surveys: Evaluating the Function of Instructed Response Items for Identifying Inattentive Respondents

[1] Shibutani Hirohide, [2] Masuda Shinya, [3] Murakami Fumio, [3] Yoshimura Harumasa

[1] Aomori University
[2] Keio University
[3] Nara University

Abstract

More and more social surveys are conducted online now than ever. Therefore, we investigated a method, planting instructed response items (IRIs) in a survey, to eliminate inattentive respondents as a necessary screening. Two web surveys were conducted; there were four IRIs in study 1 (n=2,490) and three in study 2 (n=2,000). The objectives were twofold; finding an appropriate number of IRIs in a web survey and finding the differences between the two groups, the original and the screened data, categorized by IRIs. In study 1, of the respondents who passed the first three IRIs, 1,935 out of 2,490 were considered attentive; the rest (555) were eliminated from the data analysis based on the response tree analysis. The two groups were compared on the quality-of-life scale with 24 items, with all respondents and only the attentive. The difference in mean scores between the two groups was statistically significant. Still, the difference was minor because of the shared respondents, 1,935 respondents between the two groups. The item characteristic curves from the 2-parameter logistic model were compared between the attentive (n=1,935) and the inattentive (n=555) respondents. The differences were distinctively visible, and the decision to eliminate 555 respondents was supported. In study 2, the birthday was asked and used to calculate the age. Then, the calculated age was checked by comparing the provided age by the web survey company. The rates of correct responses increased monotonically with higher levels of attentiveness. We conclude that evidence indicates IRIs function well for detecting inattentive respondents. We also tentatively recommend that three IRIs in a survey work well to detect inattentive respondents. Finally, the treatment of the respondents with the gray-zone attentiveness was discussed.

Keywords; web survey, instructed response items, attentiveness, IRT.

1. Introduction

1－1. Background

We use online surveys so often that it feels as if every questionnaire survey is done online these days. A significant advantage of web surveys is that large sample sizes can be easily achieved relatively quickly. In addition, online surveys are much cheaper than traditional paper-based surveys and much more accessible since web survey companies provide state-of-the-art digital technologies covering the needs of researchers. So, there are many good reasons for conducting web surveys. However, there are some concerns about non-sampling errors specifically associated with web surveys. Survey errors are typically categorized into two broad groups; sampling errors and non-sampling errors. Sampling errors are the errors associated with not being able to include all the target population members in your data. Non-sampling errors are categorized into three distinctive categories; coverage, measurement, and operation errors (e.g., Yoshimura, 2017; McNabb, 2014). The ways to handle these errors have been studied well, and web survey companies can usually provide precautions pointed out in the previous research; however, more than these precautions may be needed for a web survey that has recently become popular.

One of the advantages of a web survey is the ability to acquire large samples quickly. The primary reason why many respondents can attend web surveys is that they can answer survey items whenever they want, typically 24 hours a day and seven days seven a week. Furthermore, they can prepare something to answer the items, such as questionnaires, pencils, and pens. They can also respond in a relaxed atmosphere, such as while watching T.V. and talking to family members, which may distract respondents from answering items attentively. A new threat to web surveys comes from the inattentiveness of respondents while answering the survey items.

Errors caused by inattentive respondents are severe threats to web surveys (Silber et al., 2022; Arias et al., 2020; Vecchio et al., 2020; Meade & Craig, 2012). Errors caused by the inattentiveness of respondents are categorized as non-sampling errors. Non-sampling errors may affect data differently from sampling errors; sampling errors affect the selection process of the respondents in the sample, so data integrity is not considered affected. However, errors caused by respondents' inattentiveness, such as not reading questionnaire items well, may distort the statistical analysis results (Arias et al., 2020; Maniaci & Rogge, 2014). For example, the data integrity may be severely compromised if respondents read only a part of the questionnaire items. In addition, non-diligent participants add noise and can significantly decrease results reliability (Arias et al., 2020; Vecchio et al., 2020).

Generally, data cleaning is performed after conducting a survey, such as checking for missing responses, illogical responses, and out-of-rage answers. However, data cleaning only solves the problems of respondents' irregular reactions in general. We still have to face the issues from the assumption that all respondents answered questionnaires attentively. Another solution to the respondents' inattentiveness is related to the rewards for responding to a web survey. Generally, paid respondents are more motivated and attentive in answering survey questions, although excellent intention differs from valid data. So, it is still necessary to evaluate the quality of respondents after the survey is done.

The data integrity must be evaluated in terms of the attentiveness of respondents before conducting a routine analysis, especially for online surveys, to maintain the internal validity of a study (Arias et al., 2020; Vecchio et al., 2020; Maniaci & Rogge, 2014). Although different approaches have been practiced for solving the inattentive respondents related problems, such as paying respondents and organizing reliable web survey reserve monitors as potential respondents, there is no panacea for them. It is still necessary to identify inattentive respondents in the data. Therefore, several methods to check respondents' response patterns are proposed to eliminate inattentive respondents from the data analyses (Silber et al., 2022; Meade & Craig, 2012). Response time analysis, acquiescence analysis, and monotone-response-pattern detection are typically used as response check methods; these methods can be applied after the data collection. One of the most frequently used methods to assess response quality is placing instructed response items (IRIs). This method has to be used carefully so that the flow of the main questionnaire items would not be disturbed (Kung et al., 2018). One drawback of the IRIs method is that it has to be done before the survey is conducted, a big difference from the techniques, such as response time analysis and monotone-response-pattern detection. Sometimes different labels are used for IRIs, including trap questions, red herrings, validation questions, and verification ratings (Jones et al., 2012).

1－2. Instructed Response items

One of the easily applicable methods to spot inattentive respondents and eliminate them from the analysis is planting IRIs in a survey. IRIs are survey questions that ask respondents to follow specific instructions instead of asking typical questions. An example of the IRI can be "in this question, please select "Yes, I agree" regardless of your opinion about the response options. Then, attentive respondents respond to the IRI with the designated option as the answer. The frequency of the specified response of an IRI, such as the above example, should be the same as the sample size if every respondent is attentive. An IRI is usually placed at the end of grid questions that construct a scale so that it would not affect the responses to the rest of the scale items. Since IRI would detect respondents' inattentiveness to a particular item, sometimes it is designated as a local indicator to a grid, not the indicator for the entire survey. Naturally, a question such as how many IRIs are needed to detect invalid respondents becomes a valid research question; however, more information must be provided to answer the question.

Gummer et al. (2021) enumerated the advantages of using IRIs in web surveys;

> "Since (i) easy to implement in a survey, (ii) does not need too much space in a questionnaire (i.e., one item in a grid), (iii) provides a distinct measure of failing or passing the attention check, (iv) requires no interpretation by the respondent, (v) is not cognitively demanding, and (vi)-most importantly- provides a measure of how thoroughly respondents read items of a grid."

As mentioned above, there are many advantageous characteristics of using IRIs for checking the attentiveness of respondents in web surveys.

2. Objective

The present study evaluates the function of IRIs in eliminating inattentive respondents from data analysis. The present study explores two simple research objectives related to implementing IRIs in web surveys and the consequences of removing data from inattentive respondents. Therefore, we conducted two web surveys to answer the following two questions.

1. How many instructed response items do we need to detect inattentive respondents reliably?
2. What are the differences in the results of data analysis between the data with and without the inattentive respondents?

3. Methods

3-1. Two surveys

Two web surveys were conducted, one in December 2022 and another in January 2023. The ethical aspects of the survey method utilized for conducting the two surveys were evaluated by the ethical committee of Aomori university college of Sociology and approved. A Japanese web survey company was used for performing both surveys. We used this company's web-based survey system to develop two survey questionnaires. The questionnaires are grid-based, and only a limited number of items are independent questions. Both survey questionnaires ask respondents' opinions regarding timely social issues such as changing our constitution to enable the Japanese government to equip itself with armed forces and LGBT issues, which are presently heated legal problems in Japan. This web survey company claims to have 140,000 monitor respondents in all 47 prefectures in Japan. It has an excellent reputation with many clients; therefore, we have decided to use their system for our web surveys. After the surveys is done, the system provides the client with the following information besides the data from the client's questionnaire; respondents' age, residing prefecture, sex, marital status, occupation, occupational field, family income, residing status, number of children, responding device, amount of time responding to each item, time and date of completion, and spent time to complete the survey.

We planted four IRIs in the first survey and three in the second. In addition, the second survey included an item asking the respondents' birth year and day so that we could calculate their age to compare to the provided respondents' age. The planted IRIs in both surveys are very similar; all were grid items, and respondents were asked to choose a particular option. The number of possible choices in IRIs was from 4 to 11.

3-2. Analysis

The positions of all IRIs in both surveys were evenly spaced so that attentiveness was evaluated throughout the surveys. In each survey, IRIs were treated as a parcel as if they measured a psychological trait. So, respondents who successfully complied with more IRIs were considered more attentive than respondents who successfully responded with fewer IRIs in both surveys. In the first survey, respondents were categorized based on a tree-like structure developed from the success or failure of each IRI in a way that respondents who passed all the IRIs were classified as most attentive. Several criteria separated attentive and inattentive respondents into two distinctive groups. Then, the quality of life and happiness scales, which have been used for the past 20 years with established factor structures and high reliability, was used to check whether there are significant differences between the two groups. The criteria that successfully categorized respondents into two groups with significant differences were considered helpful for identifying the attentive respondents.

The same tree-like structure was used in the second survey to categorize respondents into four groups. Then, each respondent's age was calculated to check if there was any discrepancy from the company-provided age. Finally, the frequencies of the differences were compared between the four groups in the age discrepancies.

4. Results

4-1. Study 1

The sample size of study 1 is 2,490, and the main demographic characteristics are shown in table 1 and figure 1 (age distribution). The mean age (59.3) is about ten years older than the mean age in Japan (48.6). Sixty-six percent of respondents are married, slightly higher than Japan's national average (male;60.8%, female;57.0%). On the other hand, the male ratio (66.9%) among the respondents is much higher than the national average in Japan (48.2%).

The number of response options in the four IRIs planted in study 1 was 6, 6, 4, and 7, and the instructed response was option 6, 1, 1, and 4 in each item. The above information will help us imagine how difficult it is to pass all the IRIs by chance. The suggested response options for the first three items were all extreme options. Only in the last IRI, the middle option, the fourth option out of 7, was the asked response option. The path for success and failure responses in the flow of the four IRIs is presented in figure 2; successfully responded cases are described as passed, and the opposites are as failed.

As a first step, most extreme respondents were categorized into two clear groups: the most careful

Table 1. Demographic characteristics of the respondents in study 1 (n = 2,490)

variable	mean	s.d.	Min.	Max.
age	59.31	12.88	16	92

sex: male = 1,788, female = 722; s.d.: standard deviation
marital status: married = 1,643(66.9%) not married = 847
children: with = 1,447 without = 1,043
device: computer = 1,943 smartphones or tablet = 547

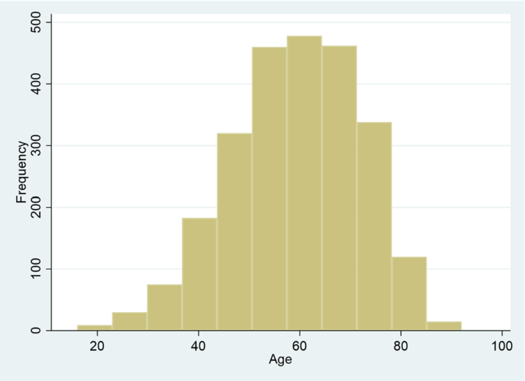

Figure 1. Distribution of age in study 1

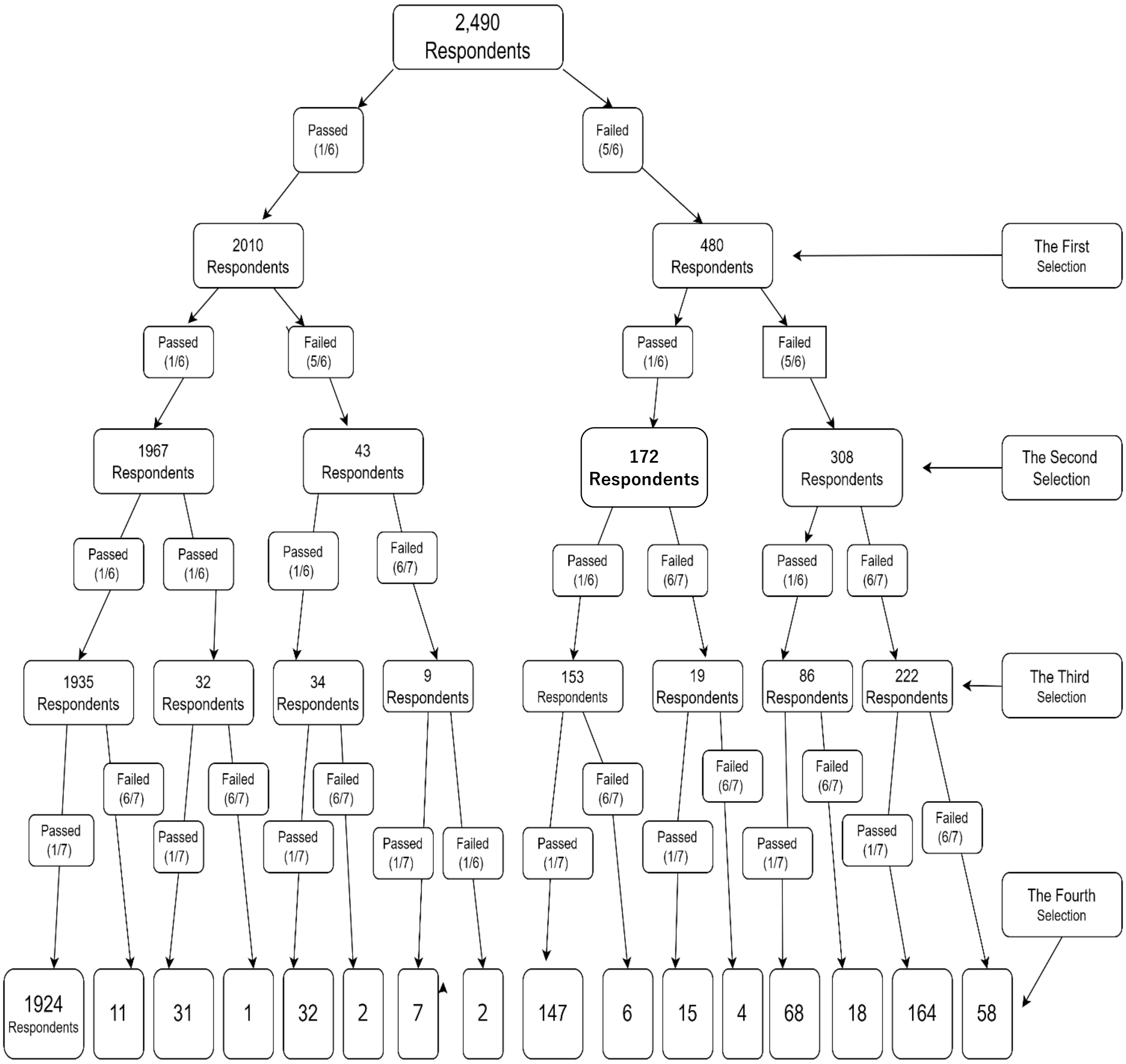

Figure 2. Response tree analysis of the four instructed response items

group passed all four IRIs, and the least careful group failed all four. Then the rest were classified into several groups with in-between attentiveness levels based on the tree. For example, the first IRI separated 480 (19.3%) respondents as inattentive, while the rest, 2,010 respondents (80.7%), as attentive (figure 2). The passing rates of the IRIs increased from the first IRI to the last IRI constantly, starting from 80.7% to 97.9, 98.4, and 99.4 at the fourth IRI shown in the leftmost path in figure 2. The passing paths finally came up with the most attentive group of 1,924 respondents (77.3%).

On the other hand, the failing path, shown in the rightmost path in figure 2, indicated an increasing tendency up until the third IRI, starting from 19.3% to 64.2 and 72.1, although the fourth IRI failing rate dropped to 26.1%. The failing paths finally came up with the least attentive group of 58 respondents. (2.3%) in a reasonably consistent manner. Placing four IRIs in one survey may become too much disturbance to the respondents. A possible reason for the failure rate dropping the fourth IRI may be related to the position of the instructed option, the only middle choice among all four IRIs. Eleven respondents (.6%)

Table 2. Cross tabulation of the last IRI with perfect success in the first 3 IRIs

Variable	Outcome	Fourth IRI	
		Passed	Failed
First 3	Perfect	924(99.4%)	11(.6%)
IRIs	Not perfect	464(83.6%)	91 (16.4%)

failed in the fourth IRI out of 1,935 who succeeded in all three previous IRIs. Ninety-one respondents (16.4%) failed the fourth IRI among those who failed some of the first three IRIs.

The difference in the failure rates on the fourth IRI is heavily conditional on the first three IRIs, which is another clear evidence that IRIs works well. However, 464 respondents (83.6%) passed the fourth IRI despite failing somewhere in the first three IRIs. This rate of 83.6% is too high for inattentive respondents, who might have gotten lucky since there are seven response options in the fourth IRI. The frequencies regarding the fourth IRI tabulated by the success in the first three IRIs are shown in table 2. The fourth IRI is the last item in the grid-15 question, with 13 items with seven response options.

Response option four, the instructed option in the fourth IRI, was the highest choice in all grid-15 questions without exception. All respondents tend to respond to the middle options in the grid-15 questions. It is likely the reason for the disproportionately high success rate of 86.3% of 464 respondents in the last IRI, not because they complied with the fourth IRI. It suggests we use the options with the lowest expected frequency as the instructed option for IRIs. Finally, a response pattern analysis was summarized in table 3; five groups were formed based on the 4-point scoring indicated in score* in table 3.

Table 3. Response pattern analysis with explanations of respondents' perspective

Response Pattern (score*)	Logical Category Score	Frequency of the last IRI	Possible explanations on respondents' perspective*
1111 (4)	Most Attentive (7)	1924/1935	I read all the questions in the survey
1110 (3)	Attentive (6)	11/1935	I made an error with carelessness or fatigue in the last item.
0111 (3)	Very attentive (6)	147/153	I was trying to figure out what to do with the first item, and then I realized the rules about IRIs.
1101 (3)	Attentive (5)	31/32	I missed the third item, most likely by a careless error.
1011 (3)	Attentive (5)	32/34	I missed the second item, most likely by a careless error.
1100 (2)	Maybe attentive (3)	1/32	I knew the check items existed, but I missed two questions by a careless error.
1010 (2)	Maybe attentive (3)	2/34	I knew the check items existed, but I missed two questions by a careless error.
0011 (2)	Maybe attentive (4)	68/86	It took two items to realize how the check items work.
0110 (2)	Maybe attentive (3)	6/153	I should have noticed the last item by a careless error.
1001 (2)	Maybe attentive (2)	7/9	I missed two items; at least the last item got right, not by a chance error.
1000 (1)	Not attentive (1)	2/9	I missed the last three items in a row because I needed to read them, so the first item was a chance.
0100 (1)	Not attentive (1)	4/19	I got one item right by chance.
0010 (1)	Not attentive (1)	18/86	I got one item right by chance.
0001 (1)	Not attentive (1)	164/222	I got one item right by chance.
0000 (0)	Not attentive at all (0)	58/222	I got no item right because I did not read the questions.

Score*: Empirical score that success was expressed as one and failure as zero.

The most attentive group, with a score of 4, shown in the first column in table 3, passed all four IRIs. On the other hand, the least attentive group scored 0, missing all four IRIs. The rest of the respondents were categorized in-between groups of three. Next, the logical category scores were developed for the secondary missing all four IRIs. The rest of the respondents were categorized in-between groups of three. Next, the logical category scores were developed for the secondary category criteria in case the first category did not produce enough supportive empirical pieces of evidence for the function of IRIs. The logical category scores were developed from the possible explanations from the respondents' perspectives and the frequency analysis in the response tree (figure 2). Another reason for developing the logical category scores is that even the same 4-point total score can be categorized into different groups based on the response patterns shown in the response tree (figure 2). For example, four response patterns produce a 4-point score of 3; however, those patterns are scored 5 or 6 in the logical category score. This type of response pattern scoring helps interpret the outcome of the tree model family of item response theory, although it is beyond the scope of this paper.

Finally, we have the base criteria from a 4-point score scale for dividing respondents into attentive and inattentive. The frequencies corresponding to all four IRIs are shown in table 4. The frequencies of respondents shown in table 4 indicate an essential clue for answering how many IRIs are needed to identify inattentive respondents. For example, the first IRI identified 480 (19.28%) respondents as inattentive. By using all four IRIs, 566 respondents (22.73%) were categorized as inattentive; three additional IRIs helped gain 86 respondents (19.28 to 22.73%). On the other hand, using only one IRI may be problematic since it can be missed by a moment of inattention. So, two or three IRIs are enough for identifying inattentive respondents based on the results of study 1. However, as explained in the other part of this section, the fourth IRI could have been more efficient for identifying inattentive respondents since it gained only 11 respondents, from 555 to 566. So, the best number of IRIs in the case of study 1 was 3; therefore, 1,935 respondents were categorized as attentive, and 555 were not.

Table 4. Instructed response items and accumulated response frequencies.

IRIs	Pass (%)	Fail (%)	Total
1st IRI	2,010 (80.72%)	480(19.28%)	2,490
Previous IRIs	1,967 (79.00%)	523 (21.00%)	2,490
Previous IRIs	1,935 (77.71%)	555 (22.29%)	2,490
Previous IRIs	1,924(77.27%)	566 (22.73%)	2,490

Table 5. QOL scale descriptive statistics.

Variable	group	Mean s.d.	Max. Min.	Sample
QOL total score	All group	50.8 10.7	22 84	2,490
	IRI group	50.4 10.5	22 84	1,935

S.D.: Standard Deviation
Min. Max.: Minimum, Maximum

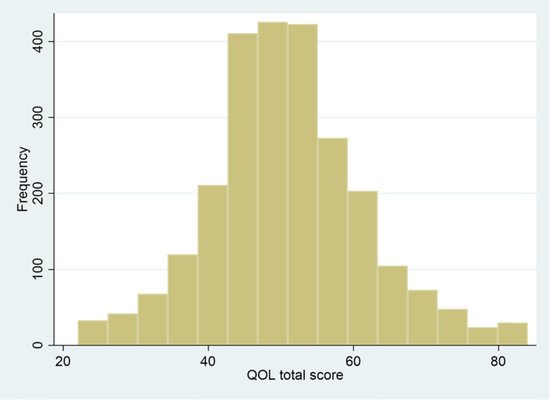

Figure3-1. Score distribution in All group

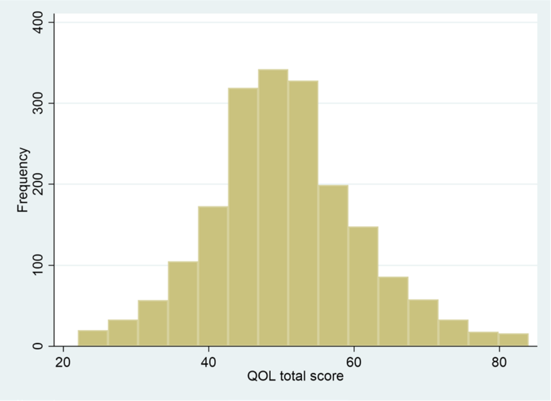

Figure3-2. Score distribution in the IRI group

Figure 3. Distribution of total QOL score

We compared the results of the analysis on the quality-of-life scale (QOL) between the two groups, a group with all the respondents (All group; n=2,490) and another group without respondents who failed in the first three IRIs (IRI group; n=1,935).The QOL scale comprises 24 items with eight sub-scales with three items each, housing, income, family, friends, relationships, free time, health, and happiness. The reliability for the QOL scale was .92 for the All group and .92 for the IRI group. The QOL scale score was chosen as the comparison variable because of the high scale reliability, a clearly defined factor structure, and the content of the items ranging from reasonably concrete to somewhat abstract. The descriptive statistics for the QOL total score and the distribution in both groups are shown in table 5 and figure 3, respectively.

The difference in means of QOL scores between the two groups is surprisingly small, although it was statistically significant. As in this survey, with a large sample size, a practically meaningless difference can be statistically significant based on a small standard error. However, the difference in the two distributions shown in figure 3 may cause more severe problems. The distribution in the All group is more heavy-tailed than the IRI group, especially in the very high score region. So, the mean scores among different age groups were compared (table 6). The differences in QOL means are surprisingly small again in all age groups, the largest discrepancy was .7; the age group from 16 to 19 was not included in the comparison since there were only four observations. One factor that must be considered here is the overlap of samples between the two groups, which is 1,935 respondents out of 2,490; we are not comparing attentive respondents (1,935) to inattentive (555). The mean score of the QOL among 555 respondents who were dropped from the analysis was significantly different (t=2.88, p=.004) from the mean of the remaining 1,935 respondents, 51.9 (11.4).

Table 6. Factor analysis of the QOL scale

Age Group	All Group		IRI Group	
	Mean	n	Mean	n
16-19	45,7 (8.1)	3	37.0 (n.a.)	1
20-29	52.1 (9.0)	36	52.1 (9.0)	22
30-39	55.5 (12.0)	145	54.8 (12.4)	106
40-49	53.9 (11.6)	374	53.9 (11.1)	279
50-59	53.2 (10.8)	649	53.3 (10.7)	504
60-69	49.6 (9.8)	679	49.3 (9.5)	543
70-79	46.7 (9.1)	495	46.0 (8.5)	391
Over 80	45.2 (8.1)	109	44.7 (7.8)	89
Total	50.8 (10.7)	2,490	50.4 (10.5)	1,935

Mean: mean (standard deviation)
n: sample size

Previous studies reported that we must check the attentiveness of the respondents, especially in a web survey (Arias et al., 2020; Vecchio et al., 2020; Maniaci & Rogge, 2014; Mead & Craig, 2012). Therefore, we chose to drop 555 respondents instead of 566 because a valid reason could be drawn only from the first three IRIs. However, an issue of importance remains in this topic; we still have to provide information on what to do with the respondents categorized in the in-between groups.

The last comparison was conducted using classical testing theory (CTT) and item response theory (IRT). To fit the 2-parameter logistic model, one of the most popular binary models of the IRT family (Shibutani & Watanabe, 2010; Shibutani, 2007), the four response options in each QOL item were collapsed into either positive or negative responses. Therefore, the QOL total score range differs from the 4-point scoring system shown in table 5 and figure 3. First, it must be specified that the comparison was performed between the IRI group (n=1,935) and the dropped group (n=555). The comparisons are made by overlaying the estimated item characteristic curve for each item and the summed scores based on classical test theory over the range of IRT-based estimation of the latent scores for the QOL. An example of the results is shown in figure 4. Comparisons of this type are conducted in every item on a scale, so comparisons were performed 24 times in this study; only one typical example is shown here. The item characteristic curve (ICC), a line in figure 4, is an estimated probability of responding positively to the item, in this case in item 7, along with the estimated QOL scale score similar to standard scores, shown from -4.0 to 4.0. The dots in figure 4 are the QOL total scores based on the binary responses plotted along with the ICC for item 7. The higher the probability of responding positively to the item, the higher the CTT total score. The locations of the dots are exactly the same in the two figures since they were based on CTT, not IRT; the differences are in the location and slope of ICC. Therefore, in IRT analysis, one of the best methods to evaluate the fit of the utilized model to the given data is

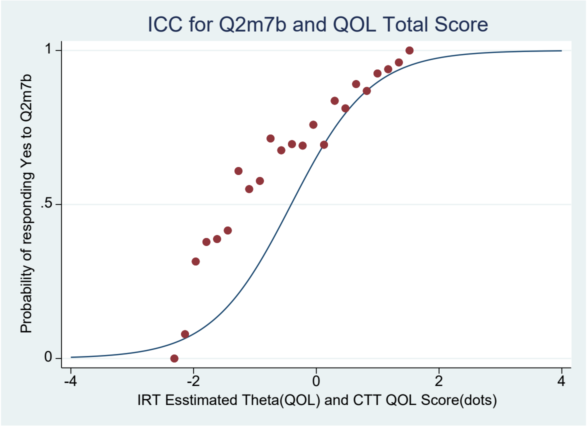

Figure 4-1. Overplot for the dropped group

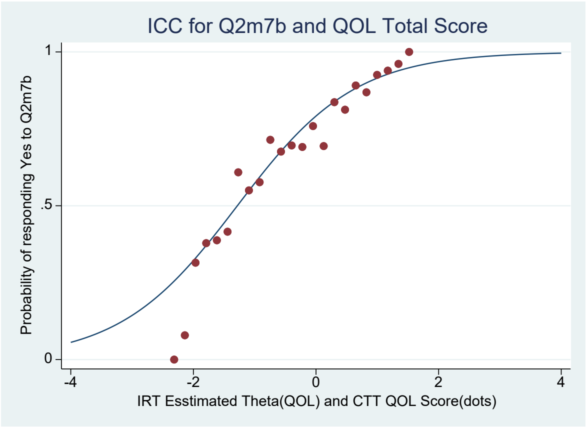

Figure 4-2. Overplot for the IRI group

Figure 4. Overplot of Item characteristic curve and QOL total scores.

to assess the closeness of the estimated ICC and overlayed CTT-based measurement, the dots in figure 4. The line and the dots are aligned more closely in the IRI group (n=1,935) than in the dropped group (n=555). The discrepancies between the dots and the line in the dropped group, figure 4-1, are pronounced in the region from -2.0 to 0.0, which corresponds to the lower side of the QOL. In addition, two dots are not close to the ICC in the IRI group, figure 4-2. These dots are in a region lower than the scale score of -2.0, corresponding to less than 30.0 in the T-score; most respondents with a typical score range are reasonably staying along with the ICC (figure 4-2). So, this can be considered one of the empirical reasons for the dropped group should not be included in the data for analysis.

Table 7. Number of respondents with an error in their age with different levels of attendance scores

Score*	Number of Respondents		Total
	Error	No Error	
0	26(20.47%)	101(79.5%)	127(100.0%)
1	19(14.18%)	115(85.8%)	134(100.0%)
2	26(12.15%)	188(87.9%)	214(100.0%)
3	70(4.59%)	1,455(95.4%)	1,525(100.0%)
Total	141(7.05%)	1,859(93.0%)	2,000(100.0%)

Pearson chi (3) = 67.883 P=.000
Score*: attentiveness score

4-2. Study 2

The second study has an advantage compared to the first; it asked about days and years of birth, so we can check whether the respondents answered correctly or not. In addition, there are three IRIs in the second study. Respondents are grouped into four categories depending on the success and failure of the IRIs; a score of 0 means no success and a score of 3 means success in all three. The error rates in the four groups are shown in table 7. Again, we can observe a consistent monotone decreasing tendency of errors corresponding to higher levels of success in IRIs.

The result shows 1) three IRIs work well in separating respondents for their attentiveness, 2) scores correspond well with the error ratio, especially the drop from the score 2 to 3 corresponds from 12.15% to 4.59%. Therefore, we suggest the appropriate number of IRIs is three rather than 2.

5. Conclusions

This study investigates two research objectives; finding an appropriate number of functional IRIs in a web survey and the differences between the results in data analysis with and without the inattentive respondents in data.

We have demonstrated that IRIs identify inattentive respondents effectively using the scale-like properties from a set of IRIs by showing a response tree analysis (figure 1 and tables 2 to 4). We also suggested that the options with the lowest expected frequency should be used as the instructed option for IRIs; in this way, one can eliminate the respondents responding without reading the question effectively. As explained in the previous section, the fourth IRI in the first survey could not separate inattentive respondents effectively because most respondents chose the instructed option, most likely including respondents who did not read the item. The instructed response option for the IRI was the middle option, the fourth option out of seven response options. Masuda et al. (2017) reported that respondents who do not read the questionnaire items carefully tend to choose the middle options. As a result, all respondents tend to respond to the middle options in the grid-15 questions. It is likely the reason for the disproportionately high success rate of 86.3% of 464 respondents in the fourth IRI, not because they complied with the fourth IRI. Therefore, researchers should use the options with the lowest expected frequency for IRIs.

Regarding the number of appropriate IRIs in a web survey, we recommended 3 IRIs tentatively since the fourth item behaved somewhat erratically. It deserves a mention that the fourth IRI could have been more efficient for identifying inattentive respondents since it gained only 11 respondents, from 555 to 566. So, the best number of IRIs in the case of study 1 was 3; therefore, 1,935 respondents were categorized as attentive, and 555 were not.

As it is clear by now, a web survey is a primary method in the 21st century. First, however, we must be aware that data screening and cleaning are two necessities for researchers. Experienced survey researchers possess skills to clean up data; however, screening out inattentive respondents from a sur-vey becomes a problem in online surveys.

Therefore, survey researchers must equip themselves with the skills to screen inattentive respondents. One of the most critical issues in screening is how to treat the gray-zone respondents; we dropped all gray-zone respondents. Unfortunately, only 222 out of 555 dropped respondents failed in all three IRIs, so 333 mation. Therefore, the treatment of the gray-zone respondents can be an issue for screening and data cleaning. We can choose to drop some of the gray-zone respondents from the survey altogether or only a part of the responses. There must be systematic reasons for whatever we do regarding the treatment of the gray-zone respondents. Further investigations into this topic merit the future of web surveys.

This study was supported by Grants-in-Aid for Scientific Research (22H00921). Japan's Ministry of Education, Culture, Sports, Science, and Technology provides Grants-in-Aid for Scientific Research. We appreciate the funding since this research could not be completed without the funding.

References

Arias, V. B., L. E. Garrido, C. Jenaro, A. Martinez-Molina A., and B. Arias, (2020). A little garbage in, lots of garbage out. Assessing the impact of careless responding in personality survey data, *Behavior Research Methods*, 52(2), 489-505.

Gummer, T., Roßmann, J., and Silber, H., (2018). Using instructed response items as attention checks in web surveys: properties and implementation, Sociological methods and research, 50(1), 238-264.

Jones, S. L., House, L. A., & Gao, Z. (2012). Respondent screening and revealed preference axioms: Testing guaranteeing methods for enhanced data quality in web panel surveys, *public opinion quarterly*, 79(3), 687–709.

Kung, F. Y. H., Kwok, N. and Brown, J. D. (2018). Are attention check questions a threat to scale validity? Applied psychology: An international review, 67(2), 264–283.

Maniasi, M. R., and R.D. Roger. (2014). Carling about carelessness: Participant inattention and it's effects on research. *Journal of Market Research,* 60, 32-49.

McNabb, D. (2014). *Non-sampling Error in social surveys*, Sage publication, California, USA.

Meade, A. W., and S.B. Craig, (2012). Identifying careless responses in survey data, *psychological methods*, 17(3), 437–455.

Masuda, S., Sakagami, T., Kawabata, H., Kijima, N., & Hoshino, T. (2017). Respondents with low motivation tend to choose middle category: Survey questions on happiness in Japan, *Behaviormetrika*. 44, 593-605.

Shibutani, H., (2007). Fundamentals of a new generation of scale analysis: From classical test theory to item response theory, 2007, *Regional Study*, 15, 31-118.

Silber, H., Roßmann, J., Gummer, T. (2022). The issues of noncompliance in attention check questions: False positives in instructional response items, *Field Methods*, 34(4), 346-360.

Shibutani, H. & Watanabe, S. (2009). Risky-choice framing effect and risk-seeking propensity; An application of IRT for analyzing a scale with a minimal number of items, *Journal of Aomori University and Aomori Junior college*, Vol. 32, No. 2, pp.65-80

Shibutani, H.　& Watanabe S. (2010). An application of classical test theory, item response theory, and partially ordered scalogram analysis for evaluating the scalability of the risk-seeking propensity, *Journal of Aomori University and Aomori Junior college*, 33(2).

Vecchio, R., Caso, G., Cembalo, L., Borrello, M. (2020). Is responents' inattention in online surveys a major issue for research? *An International Journal on Agricultural and Food Systems*, 22(1), 7, 1-18.

Watanabe, S.　& Shibutani H., (2010).　"Aging and decision making: Differences in susceptibility to the risky-choice framing effect between older and younger adults in Japan," *Japanese Psychological Research*, 52(3), 163-174.

Yoshimura, H. (2017). *Non-sampling error in the social survey,* Tosindo inc. In Japanese; 吉村治正（2017）．『社会調査における非標本誤 差』，東信堂

オンライン調査における新たなタイプのスクリーニングの必要性：回答誘導項目の無効回答者選別機能の評価

A New Type of Screening Necessity We Face with Online Surveys: Evaluating the Function of Instructed Response Items for Identifying Inattentive Respondents

澁谷泰秀[1]、増田真也[2]、村上史郎[3]、吉村治正[3]

[1]青森大学、[2]慶応義塾大学、[3]奈良大学

要　旨

社会調査がこれほど頻繁にウェブで行われていることはこれまでになかった。そこで、回答誘導項目（IRI）を用いてデータスクリーニングの一環として不注意な回答者を選別に関する研究を実施した。この研究では2,490名をサンプルとした調査と2,000名をサンプルとした2回のウェブ調査が行われた。研究の目的は二項で、第一の目的は1回の調査で不注意回答者を選別するのに適切な回答誘導項目数についてで、第二の目的は選別されたグループ間における分析値の相違についてである。第1の研究では、回答誘導項目を用いた回答木分析を行い、1,935名が注意深い回答者と分類され、残りの555名が不注意な回答者としてデータから削除された。サンプル全員で構成されるグループと注意深い回答者のみで構成される3つのグループが24項目で構成される生活の質尺度（QOL）の分析で比較された。QOL得点の平均値は統計学的に有意であったが、1,935名の注意深い回答者は両グループに共通であるため、平均値の差は大きくはなかった。項目反応理論（2母数ロジスティッ

クモデル）に基づいて求めた項目反応曲線を用いた比較では、注意深いグループと不注意なグループに明確な相違がみられたことから、不注意グループに分類された 555 名をデータ分析から除外すべきという判断は支持された。第 2 の研究では、質問に誕生年月日を加えて、各回答者の年齢を算出した。この算出された年齢はウェブ調査会社に登録された年齢と比較され、誘導回答項目を用いて分類されたグループ間で正確性が比較された。注意深さ得点が高いグループでは一貫して正確度が高くなることが確認された。我々は、これらの結果に基づき、誘導解答項目は不注意な回答者選別に機能性が高いと結論した。また、1 回のウェブ調査で使用すべき回答誘導項目数はおよそ 3 項目程度を推奨するとした。更に、注意深さの程度が明確に分類できなかった回答者の取り扱いについて考察が行われた。

キーワード：ウェブ調査、誘導回答項目、有効回答者、チェック項目