심리학석사학위논문

The Effect of the Survey Length on the Response Quality

설문지의 길이가 설문의 응답에 미치는 영향

2015년 2월

서울대학교 대학원

심리학과 계량심리 전공

유 동 주

The Effect of the Survey Length on the Response Quality

설문지의 길이가 설문의 응답에 미치는 영향

지도교수 김 청 택

이 논문을 심리학석사 학위논문으로 제출함.

2015년 1월

서울대학교 대학원

심리학과 계량심리 전공

유 동 주

유동주의 심리학석사 학위논문을 인준함.

2015년 2월

위 원 장 박 주 용 (인)

부 위 원 장 이 훈 진 (인)

위 원 김 청 택 (인)

ABSTRACT

The Effect of the Survey Length on the Response Quality

Dong Joo Yoo

Department of Psychology

The Graduate School

Seoul National University

To measure the effect that the survey length has on the response quality, several statistical indices were used. In Experiment 1, the survey length was 30-minute long, containing 192 items from various inventories. Three different types of questions were used in the survey: 5-Likert point scale items, open-ended questions, and stem-and-branch questions. Standard deviation and entropy analyses showed significant decrease in response quality when the items were located at the latter part of questionnaire. The open-ended questions also showed significant decrease in the analysis. Straight-line responses, middle responses significantly increased toward the end of the survey. In Experiment 2, the survey was lengthened to 60 minutes, containing 505 items from various inventories compiled. Standard deviation, entropy, and open-ended questions analyses showed significant decrease in response quality toward the end of survey.

Missing values analysis, straight-line responses and middle responses showed significant increase.

Keywords: item location, item position, response quality, survey length, nonsampling error

Student Number: 2009-22833

Contents

Lists of Tables

Introduction

What is a survey? According to Lessler (1992), a survey is a scientific study of a population typified by persons, institutions, or physical objects, and it attempts to quantify the characteristics of the given population. There are many types of surveys, such as face to face interviewing, mailing questionnaires, telephone and online questionnaires.

In the process of data collecting through a survey, survey error inevitably occurs. There are two types of survey error, sampling error and nonsampling error (Lessler, 1992). Sampling error can be evaluated statistically and can be controlled by designing well-developed theories or by increasing the population size. On the contrary, nonsampling error cannot be controlled nor can it be estimated statistically.

There are three general types of nonsampling error: (1) frame error, (2) nonresponse error, and (3) measurement error. Frame error occurs when the structure of the sample frame is designed poorly, and nonresponse error occurs when the survey fails to collect response from the respondents (Lessler, 1992). Measurement error is attributable to the interviewer, wording of the questions, data coding, or the questionnaire itself (Biemer, 1991).

Survey length is part of measurement error which is a type of nonsampling error. When the length of the survey becomes longer, respondents may show signs of fatigue, loss of interest, and thereby careless responses manifest (Lessler, 1992). As a result, the quality of responses may decrease in terms of accuracy and credibility. In business settings, however, to increase reliability and reduce costs, the establishments often increase the length of the survey to obtain as much information as they possibly can (Biemer, 1991). To minimize the nonsampling error which comes from increasing the length of the survey, appropriate length and time shall be explored for the best response quality. One way is through analyzing the location of the items in a survey.

Studies on the effects of the survey length on the response rate and the response quality have been done for many decades. Meta-analyses done by Heberlein and Baumgartner(1978) and Yammarino, Skinner, and Childers(1991) have shown that longer surveys result in lower response rate. Helgeson and Ursic(1994) have also found that the position of the questions, either at the beginning or at the end, of the questionnaire can affect response quality as well. Some studies have found increased rate of “don’t knows” as the length of the survey increases (Krosnick et al., 2002) and some have found “straight-line response” pattern, indicating boredom or fatigue due to the survey length (Herzog and Bachman, 1981).

Among the studies on the survey length, the responses to open-ended questions were also probed. Based on the experiment led by Johnson et al.(1974), the length of the response was shorter when it was at the end of the survey than when the open-ended question was at the beginning of the survey . On the contrary, the research done by Burchell and Marsh (1992) obtained the opposite results. Such conflicting results require more inspection into the effect of survey length on open questions as well.

Stem-and-branch questions were also investigated in several studies. Stem-and-branch questions are usually the questions that have the format of one yes or no question followed by various related questions, or it may contain responses that has many boxes to check. For example, on many self-reported medical surveys asking for reporting symptoms, respondents often show attenuation (Duan et al., 2006, Kessler et al. 1994, 1998, 2000, Jensen et al., 1999).

Attenuation is a form of survey conditioning where the responses are affected by previous experiences in the same or similar surveys before, thereby already “conditioned” to the format of the survey (Duan et al., 2006). The attenuation has found to cause respondents to underreport over time. In particular, once the respondents have learned the format of the items, they may answer negatively to the stem questions in order to avoid the branch questions that

follows.

Another study regarding the stem-and-branch questions were performed to investigate the order effects (Jensen et al., 1999). Jensen and his colleagues the locations of the items in the survey and found order effects (Jensen et al., 1999). In particular, when the items were presented earlier in the survey, people tended to check more boxes than when the items were presented later. However, the differences in the location of the items in the survey were not always statistically significant. So the investigation on the stem-and-branch questions is required for better understanding of the order effect as well as the survey length effect.

Indices for Response Quality

In the present study, some new methods will be proposed to analyze the quality of the responses, along with the traditional approach that has been used throughout the psychological research field.

Cronbach α: Derived from the Classical Test Theory, Cronbach-α measures the reliability of the survey (Cronbach, 1951). Reliability is a traditional method in the field of psychometrics which measures the correlation of the items in a survey or a test

(Nunnally, 1997). It is a form of estimating measurement error; which means, higher the reliability, lower the measurement error, and vice versa.

$$\text{Cronbach } \alpha = \frac{k}{k-1}(1-\frac{\sum \sigma_i^2}{\sigma_x^2}) \ ,$$

where k is the number of items, σ_x^2 is the variance of total score, and σ_i^2 is the variance of item i.

However, the length of the survey may change the reliability. Sometimes, increasing the items in the survey can increase the reliability coefficient (Cronbach-α), but the higher reliability of the survey is not necessarily a positive one. The survey may have high reliability, or high precision, but its accuracy would be low. Accuracy refers to the closeness of a measurement to an accepted value, while precision refers to the closeness of a series of measurements of one another (Tarendash, 2013). In other words, when a survey contains too many items than necessary, no matter how high its reliability is, it might not be accurate.

Entropy: According to Shannon and Weaver (1949), entropy measures the disorder of the responses of an individual in a given survey. If the diversity of an individual's responses is low, the entropy will be low, and the reverse will be true for higher diversity of an individual's responses. However, interesting phenomena about the entropy may be observed. For example, if the response of an individual happens to be 1-1-1-1-1-2-2-2-2-2 in five-multiple choice items, the entropy is the same as the series of responses 1-2-1-

2-1-2-1-2-1-2. Entropy only concerns with the variety of answers and is not order-sensitive. However, straight-line response analysis, introduced by Herzog and Bachman (1981), is more sensitive to the order.

Along with the entropy analysis, the average, the standard deviation, and the missing analyses of the individuals will be shown to compare the differences in response quality between the earlier the latter part of the survey. Unlike the entropy analysis, which is a new method for measuring response quality, the three analyses are traditional methods.

Person-fit Index: Derived from the item-response theory (IRT), person-fit index refers to the statistical assessment of IRT model-fit at the level of the individual person (Embretson and Reise, 2000). IRT is a new theory introduced into the field of psychometrics and has gained some popularity and its usefulness has been accepted by many psychologists who are interested in individual differences analysis.

First introduced by Lord and Novick (1968), IRT is a model-based measurement which measures latent trait by measuring the persons' responses and on the item properties such as item difficulty, item discrimination, and item guessing. Person-fit index, labeled L_z, attempts to measure the accuracy, or the validity, of the IRT measurement at the individual level by standardizing the log-likelihood (LogL) of an item response vector (Drasgow, Levine, and Williams, 1985).

$$L_0 = \sum \left[X_j \ln p_j + (1 - X_j) \ln (1 - p_j) \right]$$

$$L_z = \frac{L_0 - E(L_0)}{[Var(L_0)]^{\frac{1}{2}}} \sim N(0,1)$$

where X_j is the response of j'th item, p_j is the probability of X_j predicted by IRT model, $E(L_0)$ is expected value of L_0 and $Var(L_0)$ is variance of L_0.

The goal of L_z is to identify the pattern of response that individuals produce very unlikely, thereby locating the possibility of error. The higher the L_z is, the higher the possibility of aberrant test behaviors is, such as cheating, fumbling, or carelessness (Birenbaum, 1986).

Middle response (3 on Likert scale) frequency: In this research, we are concerned with how often the respondents choose the answer in-the-middle, "3". The respondents would choose "3" when they become tired as the survey progresses due to the length of the survey, and they may be less attentive about the contents of the items. We are also concerned with how often the respondents selected "2-3-4" response pattern. By detailed analysis, we expect to find that some individuals may try to be "smart" and select "2" or "4" to make it seem that they were paying attention although they were just as careless about the content of the item and choose answers randomly.

Experiment 1

Experiment 1 was designed to investigate the effects of questionnaire length upon individuals' response qualities. An assumption in this study is that the locations of items in a questionnaire reflects questionnaire lengths. In other words, the n'th item shows the behaviors of questionnaires with length n. The questionnaire used in Experiment 1 includes 6 different batteries with 5-Likert scale as means of measurement, 8 open-ended questionnaires, and 6 stem-and-branch questions. The total items in the questionnaire was 192, with 148 5-Likert scale questions, 8 open-ended questions, and 6 stem questions with 30 branch questions. The batteries were counterbalanced to form two types of surveys, Type A and Type B. The estimated time for completing the *short survey* was 30 minutes.

Experimental Design

Participants: Participants were collected by two means: paid and non-paid. The non-paid group was recruited from Seoul National University, located in Seoul, South Korea. Only the undergraduate students who were taking psychology courses from SNU could participate. One credit was granted for 30-minute participation, and only the gender information was collected from this group. The analyses were based on whether they were paid or nonpaid, and type A or B.

The paid group was more complicated to collect because it involved recruiting from the universities located in Korea outside of Seoul National University. To reach this population, SNS was used as a mode of communication and the recruitment information regarding the research was posted on various university community web pages.

There was no limitation for age or gender to participate, but since the questionnaire was in Korean, foreigners could not participate in the study. Total of 387 undergraduate students participated with 123 males, 264 females and 206 in type A and 181 in type B (see *Table 1*).

Table 1. Sample size of Experiment 1 by gender and type (ratio)

	Type A	Type B	Total
Males	62(0.30)	61(0.33)	123(0.32)
Females	144(0.70)	120(0.67)	264(0.68)
Total	206(0.53)	181(0.47)	387

Instrument: The content of the survey were tailored to the undergraduate students in that the students take many exams during the semester so the 2 Test Anxiety Scales were chosen, developed by Spielberg (1980) and Hwang (1997), and placed in front and at the end of the questionnaire. Likewise, the 2 Self Consciousness scales were chosen, developed by Kim (1991) and Kim (1993), and placed in front and at the end of the questionnaire. These scales were carefully selected to compare the responses in the beginning of the questionnaire and at the end.

Two more scales were included in the survey: the Self Concept Scale (Lee, 1997) and the Academic Motivation Scale (Kim, 2002). They were placed in the middle of the survey. Also, 8 open-ended questions and 6 stem questions were divided into two sets, Set 1 and Set 2. Each set contained 4 open questions and 3 stem questions, and they were placed before and after the Self Concept Scale and the Academic Motivation Scale to observe the effect of the survey length. The order for type A and type B survey is shown in <*Table 2*>.

Table 2. Order of the questionnaires in the short survey (Number of questions)

Type A	Type B
TA 1 (20)	TA 2 (20)
SC 1 (17)	SC 2 (17)
Set1: Open (4)	Set2: Open (4)
Stem (3)	Stem (3)
M(44)	SN (30)
SN (30)	M(44)
Set2: Open (4)	Set1: Open (4)
Stem (3)	Stem (3)
SC 2 (17)	SC 1 (17)
TA 2 (20)	TA 1 (20)

Note. Acronyms for scales: TA stands for Test Anxiety Scale, SC stands for Self Consciousness Scale, M stands for Academic Motivation Scale, SN stands for Self Concept Scale.

Masking: The purpose of the survey was to observe the effect of survey length by the location of items in the survey. However, if the participants knew about the purpose beforehand, they might respond differently, i.e., try to be more attentive even at the end of the survey. The response should be more natural and thus the real purpose and the real title of the survey were masked for the best scientific result.

When the survey was administered, the title of the survey was introduced as “College Life and Culture Survey”. After the survey was over, the real purpose of the survey was debriefed. If the participant did not want to participate after the true purpose was revealed, they may choose not to use his or her data.

Table 3. Demographics of Experiment 1 by type and payment (ratio)

	Type A	Type B	Total
Nonpaid	59(0.29)	51(0.28)	110(0.28)
Paid	147(0.71)	130(0.72)	277(0.72)
Total	206(0.53)	181(0.47)	387

However, no one wished to withdraw.

Results

Demographics: The demographics are shown in <*Table 3*>. The nonpaid group, which was collected from SNU, was 110, 28% of the sample population. The paid group, collected from various institutes in Korea, was 277, 72% of the sample population. It was not intended, but the paid group was more than twice the nonpaid group. When designing the experiment, the motivation behind paid and nonpaid group seemed interesting to analyze, but due to the unbalanced size of the sample, this analysis was not done. Type A participants were 206, 53% of the population, and Type B participants were 181, 47% of the population. The administration of the survey was random for the type of survey.

Table 4. Reliability coefficient (Cronbach α) for scales used in Experiment 1.

	First Half	Second Half	Δ	*p*
TA1	0.933	0.946	.013	.134
TA2	0.901	0.935	.035	.003*
SC1	0.782	0.786	.003	.913
SC2	0.799	0.847	.048	.057+

* p<.05

Reliability factor – Cronbach α: In <*Table 4*>, the reliability of the tests are shown. The tests for comparison, the 2 Test Anxiety scales and the 2 Self Consciousness scales, did not decrease toward the end of the survey. Test Anxiety scales were the first scales to be measured, both in Type A and Type B, and they were farthest apart, yet they did not show decrease. Rather, the reliability coefficient, Cronbach α, increased. Same was true for Self Consciousness scales. The higher reliability rate was even statistically significant for TA2 (p=0.012). The predicted reliability change was not observed. The reason for the unpredicted high reliability toward the end of the survey will be discussed in the general discussion section.

Person Fit Index: The IRT-based person fit index for Experiment 1 did not show any statistically significant result..

Table 5. Differences in the lengths of the responses for the open-ended questions.

	First Half	Second Half	Δ	*p*	
O1	64.19(60.46)	44.58(44.20)	-.37	.000	**
O2	52.81(39.24)	44.07(39.27)	-.22	.029	*
O3	67.76(47.34)	52.71(34.17)	-.36	.000	**
O4	74.05(51.25)	61.57(43.09)	-.26	.010	**
O5	53.46(32.39)	46.20(29.67)	-.23	.023	*
O6	56.11(46.60)	47.06(33.59)	-.23	.031	*
O7	50.61(32.93)	42.81(32.61)	-.24	.020	*
O8	70.28(58.03)	60.10(40.64)	-.21	.049	*

* p<.05 , ** p<.01 .

Analysis of open-ended questions by individuals: The results for the analysis of the lengths of the open-ended questions by individuals was as predicted. All 8 of the open questions showed differences in the lengths that were significant by the t-test analysis. The responses to open questions placed at the later part of the survey showed shorter response length compared to the responses to questions placed at the earlier part of the survey. Looking at closely, it was found that O1 and O3 had stronger significance (p<.01) than the other open questions (p<.05). The results are shown in <*Table 5*>.

Table 6. Yes/No ratio of stem questions.

	First Half (Yes/No)	Second Half (Yes/No)	χ^2	*P*
B1	87 / 13	86 / 14	0.006	.938
B2	70 / 30	79 / 21	3.586	.196
B3	87 / 13	84 / 16	0.478	.587
B4	75 / 25	66 / 34	2.742	.196
B5	83 / 17	78 / 22	1.674	.294
B6	90 / 10	84 / 16	2.951	.196

Analysis of stem-and-branch questions: This study predicted that the negative responses to stem questions will increase as the survey progresses because the participants, having understood the scheme of stem-and-branch questions, would want to avoid the contingent problems that comes after answering affirmatively. Since the total number of Type A participants and Type B participants were not equivalent, 206 and 181 respectively, the ratio of yes and no responses were calculated and χ^2-test was conducted. All items, except for B2, tend to increase in the number of "no" responses toward the end of the survey. However, the differences were not statistically significant. The results can be seen in <*Table 6*>.

Table 7. Average of scale of individuals by location.

	# of questions in between	First Half (s.d.)	Second Half (s.d.)	Effect Size	*p*	FDR- *p*
TA1	L:128, O:8, S:6	2.79(0.72)	2.57(0.76)	-.29	.005	.021*
TA2	L:128, O:8, S:6	3.00(0.68)	3.09(0.75)	.13	.217	.290
SC1	L:91, O:8, S:6	3.38(0.48)	3.39(0.46)	.02	.808	.808
SC2	L:91, O:8, S:6	3.56(0.47)	3.49(0.48)	-.16	.119	.237

* p<.05 .

Average of scale by individuals: <*Table 7*> shows the average of the Test Anxiety scales and the Self Consciousness scales by individual level analysis depending on their location in the survey. Between the pair of the Test Anxiety scales, TA1 and TA2, located were 128 Likert-scale items, 8 open questions, and 6 stem questions with 30 optional branch questions. Between the pair of Self Consciousness scales, SC1 and SC2, there were 91 Likert-scale items, 8 open questions, and 6 stem questions with 30 optional branch questions.

For the average analysis, false-discovery rate of p-values were reported. False-discovery rate (FDR) controls for the increased type I errors due to multiple comparison. (Benjamini & Hochberg, 1995). In the following average analyses for standard deviation, entropy, straight-line responses, false-discovery rate will be reported.

Table 8. Standard deviation of scales of individuals by location.

	# of questions in between	First Half (s.d.)	Second Half (s.d.)	Effect Size	*p*	FDR- *p*
TA1	L:128, O:8, S:6	0.97(0.24)	0.87(0.31)	-.34	.001	.002**
TA2	L:128, O:8, S:6	1.04(0.27)	0.87(0.33)	-.56	.000	.000**
SC1	L:91, O:8, S:6	0.98(0.27)	0.96(0.28)	-.05	.643	.643
SC2	L:91, O:8, S:6	1.04(0.28)	0.95(0.33)	-.29	.005	.006**

* p<.05 , ** p<.01 .

Although Test Anxiety scales were the farthest apart in the survey, only TA1 showed decrease in average. The Self Consciousness scales also showed conflict in the differences when comparing the test placed at the beginning and the end. The changes were not consistent, nor was it significant.

Standard deviation of scale by individuals: As can be seen in <*Table 8*>, the average of the standard deviation of the individuals generally decreased for all scales depending on their placement. Except for SC1, all the other scales showed statistical significance. Unlike the average of the individuals which showed no noticeable pattern, standard deviation of the individuals decreased as the items were located later part of survey.

Entropy of scales by individuals: <*Table 9*> shows the differences in the average of the entropy of individuals depending on where the scale was placed. The difference was bigger as the scales compared are farther apart. For example, TA1 and TA2 both showed effect size of -0.60 (p=.000),

Table 9. Entropy of scales of individuals by location.

	# of questions in between	First Half (s.d.)	Second Half (s.d.)	Effect Size	*p*	FDR- *p*
TA1	L:128, O:8, S:6	1.16(0.24)	0.98(0.35)	-.60	.000	.000**
TA2	L:128, O:8, S:6	1.21(0.25)	1.03(0.34)	-.60	.000	.000**
SC1	L:91, O:8, S:6	1.15(0.24)	1.11(0.28)	-.17	.109	.109
SC2	L:91, O:8, S:6	1.16(0.24)	1.06(0.32)	-.35	.001	.001**

* p<.05 , ** p<.01 .

while SC1 and SC2 showed effect size of -0.17 and -0.35, respectively (p=.109 and p=.001, respectively). However, the effect size of SC1 did not come out to be statistically significant, same as the result of standard deviation of scale of individuals analyzed in the above section.

Average, standard deviation, and entropy of scales by items: The analyses for average, standard deviation, and entropy of scales by item level did not show any statistically significant result.

Average of missing values by individuals: Since missing values do not follow normal distribution, zero-inflated count model was used to analyze the missing values (Lambert, 1992; Mullahy, 1986). The rate of missing values were not consistent. TA1 and SC1, placed at the start of Type A survey, showed effect size of 1.14 (p=.008) and 0.69 (p=.086), while TA2 and SC2, placed at the start of Type B survey, showed effect size of -0.11 (p=.521) and -0.31 (p=.231).

Table 10. Average of missing values of individuals by location.

	First Half	Second Half	*p*	First Half	Second Half	Effect Size	*p*
	Ratio of completed surveys	Ratio of completed surveys		Avg. of missing data	Avg. of missing data		
TA1	0.93	0.91	.439	1.36	2.50	1.14	.008**
TA2	0.89	0.92	.996	1.30	1.19	-0.11	.521
SC1	0.96	0.86	.911	1.12	1.81	0.69	.086+
SC2	0.93	0.88	.034	1.67	1.36	-0.31	.231

+ p<.10 , * p<.05 , ** p<.01 .

Straight-line response (Herzog and Bachman, 1981): There are three ways of dealing with non-responses. One way is to eliminate the individual's sample completely out of the analysis and only keep the samples with completed responses. Another way is to exclude non-response items when counting straight-line response. The third way is to treat non-response as one category of response. For example, if an individual responded to 5 items but left 10 items in one inventory, the longest straight-line response for this participant is 10. In this study, data were analyzed using the third method. TA1 and TA2 showed highly significant differences by the placement in the survey; 0.48 (p=.000) and 0.39 (p=.000) respectively. On the other hand, SC1 and SC2 did not show significant differences; 0.17 (p=.128) and 0.10 (.308) respectively. The results are shown in <*Table 11*>.

Table 11. Straight-line response by location.

	# of questions in between	First Half (s.d.)	Second Half (s.d.)	Effect Size	*p*	FDR- *p*
TA1	L:128, O:8, S:6	4.45(2.63)	6.04(3.94)	.48	.000	.000**
TA2	L:128, O:8, S:6	3.76(2.36)	5.05(3.99)	.39	.000	.000**
SC1	L:91, O:8, S:6	3.36(1.85)	3.72(2.37)	.17	.096	.128
SC2	L:91, O:8, S:6	3.69(2.20)	3.97(3.09)	.10	.308	.308

* p<.05 , ** p<.01 .

Middle response (3 on 5-point Likert scale) frequency: Unlike the straight-line response analysis proposed by Herzog and Bachman (1981), middle response frequency analysis involves the character of Likert scale, the middle scale, in particular. If the Likert scale is even numbered, for example, 6-Likert scale, the scale is regarded as measuring two big concepts, 1 to 3 responses as negative (or positive), and 4 to 6 responses as positive (or negative). If the Likert scale is odd numbered, for example, 5-Likert scale, as in the case of the survey used in this research, the scale is regarded as measuring three big concepts, 1 and 2 responses as negative (or positive), 3 as in-the-middle response (“agree”, “likely”, or whatever is appropriate for the subject being probed), and 4 and 5 responses as positive (or negative).

Table 12. Number of individuals who answered response 3.

	All "3"		Response 3 (>10)		*p*
	First Half	Second Half	First Half	Second Half	
TA1	0	2	9	12	.000**
TA2	1	4	5	30	.091+
SC1	1	1	11	8	.580
SC2	1	6	5	20	.215

+ p<.10 , * p<.05 , ** p<.01 .

As can be seen in <*Table 12*>, the results were not clear-cut as the number of individuals who answered all "3" on Likert scale did not necessarily increase at the end of the survey. However, since this is not about straight-line response analysis, the order of response did not matter, so the analysis was done again based on the number of participants who answered "3" more than 10 times for each inventory. As a result, there was definite increase in the number of participants who answered "3" later in the survey. TA1 increased by 3 participants (p=.000), TA2 increased by 6 times (p=.091), SC2 increased by 4 times (p=.215). However, SC1 decreased by 3 participants (p=.580).

Middle response (2-3-4 on Likert scale) frequency: This time, the middle response frequency analysis was done after coding response 2 and 4 as response 3. The reason for doing this analysis was to catch out the participants who selected response 2 or 4 as a way to deceive the researcher that they were

Table 13. Number of individuals who answered responses 2-3-4.

	All 2-3-4		*p*
	First Half	Second Half	
TA1	23	35	.200
TA2	14	42	.011*
SC1	44	40	.377
SC2	14	45	.033*

* p<.05 .

attentive to the contents of the items but perhaps the real motive for such behavior is not clear. One thing was clear that they indeed avoided choosing the extreme responses 1 or 5.

Including these individuals in the analysis, TA2 and SC2 showed significant increase (p=.011, p=.033, respectively), TA1 showed increase but not significant, and SC1 showed decrease which was not significant. The results are shown in <*Table 13*>.

Discussion

The results can be summarized into three aspects: consistency, accuracy, and diligence. For consistency, the reliability and person fix index analyses did not show any meaningful differences between the fore part and the latter part of the survey. It can be concluded that the quality of responses in Experiment 1 did not change throughout the survey in terms of consistency.

The results for accuracy were more satisfactory: the individual level

analyses of standard deviation and entropy of each inventory in the survey showed significant differences between the fore part and the latter part of the survey. The middle responses and straight-line responses analyses showed significant differences between the two locations in the survey. However, the analyses by item level did not show any meaningful result.

Diligence was studied by the analysis of missing values using the zero-inflated count model, the analysis of open questions using t-distribution, and the analysis of stem questions using χ^2-distribution. The missing values increased toward the end of the survey, the length of responses for open-ended questions decreased, and the number of negative responses to stem questions generally decreased. Although significant results were found only for open-ended questions, the pattern was consistent that responses become less diligent as the survey progresses.

The results for Experiment 1 still leave unanswered questions, such as why the stem questions did not yield more negative responses, and why the differences in the missing values was not significant. Although the survey contained 192 items to complete in 30 minutes, the time pressure did not seem to stress the participants enough. The researcher wanted to test if the participants would answer differently if the survey was twice as long.

Experiment 2

The result of Experiment 1 prompted the reason for a new experiment because the short survey was only 30-minute long and perhaps it was not long enough to tire the participants out so terribly. The purpose of the study is to observe when participants become fatigued and answer carelessly toward the end of a given survey. Therefore, the researchers extended the questionnaire from Experiment 1 to increase the length of the survey.

Two self-conscious scales were removed from questionnaire used in experiment 1 because they seemed redundant with self-concept scale and 9 new scales were added. All items were standardized to 5-Likert scales, and 12 more open-ended questions were added, with some branch questions that are not optional, unlike the previously used 6 stem questions that are optional if they answer negatively to the stem question.

In sum, the total items in the questionnaire was 505 questions, with 439 5-Likert scale questions from 13 different scales, 8 open-ended questions, 12 additional open-ended questions with 36 branch questions (not optional), and 6 branch questions with 30 branch questions (optional). The batteries were counterbalanced to Type A and Type B. The estimated time for completing the

long survey was 60 minutes.

The research hypothesis is that toward the end of the survey, many deviant and abnormal patterns will be observed: such as lower reliability, shorter responses to open questions, increase in negative answers to stem questions.

Experimental Design

Participants: The non-paid group was recruited from the psychology courses at Seoul National University. The students were rewarded 2-credit for one hour participation. Only their gender information was collected.

The paid group was recruited from various university community websites, such as SNS webpages. Some demographic information such as the university they are attending and their gender were collected Total of 223 undergraduate students participated with 108 males, 115 females and 119 in type A and 104 in type B (see *Table 15*).

Table 14. Sample size of Experiment 2 by gender and type (ratio)

	Type A	Type B	Total
Males	62(0.52)	46(0.44)	108(0.48)
Females	57(0.48)	58(0.56)	115(0.52)
Total	119(0.53)	104(0.47)	223

Instrument: The contents of the survey were extended from the survey used in Experiment 1, but 2 Self Conscious scales were removed. To elongate the survey, following scales were included for comparison: the 2 Self Identity scales, developed by Bennion and Adams (1986) and Park (1996); the 2 Hopelessness scales, developed by Beck (1974) and Lee (1993). The scales of each pair were placed in front and at the end of the questionnaire.

Table 15. Order of the questionnaires in the long survey (Number of questions)

Type A	Type B
TA1(20)	TA2(20)
SI1(48)	SI2(48)
HS1(20)	HS2(20)
Set1: Open (10)	Set2: Open (10)
Stem (3)	Stem (3)
M-SE(26)	SD(48)
M-FT(18)	N(12)
CG(25)	CC(30)
PO(30)	C(48)
C(48)	PO(30)
CC(30)	CG(25)
N(12)	M-SE(26)
SD(48)	M-FT(18)
Set2: Open (10)	Set1: Open (10)
Stem (3)	Stem (3)
HS2(20)	HS1(20)
SI2(48)	SI1(48)
TA2(20)	TA1(20)

Note1. Acronyms for scales: TA stands for Test Anxiety Scale, SI stands for Self Identity Scale, HS stands for Hopelessness Scale, SE stands for Self Efficacy subscale, FT stands for Failing Tolerance subscale, CG stands for Cognitive Satiation Scale, PO stands for Social Problem Solving Scale, C stands for Conscientiousness subscale of NEO-PI-R, CC stands for Self Concept Scale, N stands for Fear of Negative Comment – Short Form, and SD stands for Acceptance Desire Scale.

Note2. Academic Motivation scale(M) from Experiment 1 was divided into two different sub-scales to analyze; Self Efficacy(M-SE) sub-scale and Failing Tolerance(M-FT) sub-scale.

The scales placed in the middle to lengthen the survey are as follows: in addition to the Self concept scale (labeled CC in Table 15) (Lee, 1997) and Academic Motivation scale (labeled M in Table 15) (Kim, 2002) used for the short survey, Cognitive satiation scale (Kim, 1994), Acceptance desire scale (Keum, 1984), Social Problem-Solving Scale (D'zurilla and Nezu, 1990, Kim, 1995), Fear of Negative Comment – Short Form (Choi & Lee, 1994), and the Conscientious scale from the NEO-PI-R (Jeon, 2013) were included in the long survey. The open questions and stem questions were divided into two sets and placed before and after this middle set of questionnaires. The purpose of the open-stem set questions was to divert the attention of the participants of the Likert-scale questions for a while. At the same time, the length of the open responses and the number of negative responses to stem questions depending on placement will be probed. The order for type A and type B survey is shown in <*Table 15*>.

Masking: Similar to the short survey in Experiment 1, the purpose of the long survey was masked for natural observation and the best scientific result. The title of the survey was introduced as "College Life and Culture Survey" and when the survey was over, the real purpose of the survey was explained in detail. If the participant did not want to participate after the true purpose was revealed, they may choose to do so. However, no one wished to withdraw.

Results

Demographics: In <*Table 16*>, the demographics of the Experiment 2 is shown. The nonpaid group was 56, 25% of the total sample size, and the paid group, was 277, 75% of the sample population. Type A participants were 119, 53% of the population, and Type B participants were 104, 47% of the population. It is a coincidence, but the ratio of the type of survey was the same as the sample population for Experiment 1. The administration of the survey was random for the type of survey distributed.

Table 16. Sample size of Experiment 2 by type and payment (ratio)

	Type A	Type B	Total
Nonpaid	29(0.24)	27(0.26)	56(0.25)
Paid	90(0.76)	77(0.74)	167(0.75)
Total	119(0.53)	104(0.47)	223

Reliability factor – Cronbach α: The reliability coefficient, Cronbach α, of the tests used in Experiment 2 are shown in <*Table 17*>. Six tests were compared, the 2 Test Anxiety scales, the 2 Self Identity scales, and the 2 Hopelessness scales. Except for HS2, the reliabilities of tests increase as the test located at the later part of survey. TA1, SI1, and HS1 showed significant

Table 17. Reliability coefficient (Cronbach α) for scales used in Experiment 2.

	First Half	Second Half	Δ	*P*
TA1	0.922	0.952	.030	.001**
TA2	0.903	0.925	.021	.084+
SI1	0.710	0.791	.081	.023*
SI2	0.941	0.950	.009	.246
HS1	0.904	0.936	.031	.007**
HS2	0.958	0.957	-.001	.850
SE	0.856	0.792	-.064	.012*
FT	0.815	0.840	.025	.311
CG	0.920	0.904	-.016	.200
PO	0.941	0.924	-.016	.096+
C	0.908	0.900	-.007	.600
CC	0.914	0.914	.000	.001
N	0.877	0.903	.025	.108
SD	0.882	0.900	.017	.268

+ p<.10 , * p<.05 .

increase, .001, .023, and .007, respectively, and they were all positioned earlier in Type A survey. TA2, SI2, and HS2 were located earlier in Type B survey, and they did not show significant differences. Although measuring the same test anxiety, and with same distance apart, TA1 showed significant increase in reliability, while TA2 did not showed significant increase. Similar patterns were observed for Self Identity scales and Hopelessness scales. The prediction for the decrease in the reliability is not observed. Other tests included in the

Table 18. Differences in the lengths of the responses for the open-ended questions.

	First Half	Second Half	Δ	*p*	
O1	71.71(69.16)	51.29(80.93)	-.27	.046	*
O2	58.16(51.49)	39.13(33.94)	-.43	.001	**
O3	62.30(49.09)	47.15(34.19)	-.35	.008	**
O4	69.78(48.16)	55.63(46.76)	-.30	.027	*
O5	58.73(70.80)	43.71(38.46)	-.27	.056	+
O6	66.53(46.67)	48.97(37.49)	-.42	.002	**
O7	56.13(42.66)	37.98(24.60)	-.53	.000	**
O8	55.04(44.12)	37.95(23.62)	-.49	.001	**

+ p<.10 , * p<.05 , ** p<.01 .

long survey are presented in <*Table 17*>, but none of the reliability change were significant. More will be discussed in the General Discussion section.

Person Fit Index: The IRT-based person fit index for Experiment 1 did not show any significant result.

Analysis of open-ended questions: Total of 20 open-ended questions were used in the long survey, but 12 added questions did not show any significant pattern so they were not included in the analysis. The results are shown in <*Table 18*>. Consistent with Experiment 1, open questions O1 to O8 all showed decrease in the t-test analysis. Except for O5, which were marginally significant (p=.056), all the rest of the items were significant.

Table 19. Yes/No ratio of stem questions.

	First Half (Yes/No)	Second Half (Yes/No)	χ2	*p*
B1	86 / 14	81 / 19	0.714	0.597
B2	83 / 17	75 / 25	1.501	0.597
B3	65 / 35	69 / 31	0.338	0.673
B4	41 / 59	42 / 58	0.000	1.000
B5	78 / 22	66 / 34	2.683	0.597
B6	76 / 24	70 / 30	0.743	0.597

Analysis of stem-and-branch questions: The ratio of yes and no responses were shown in <*Table 19*>. Although the results were not statistically significant, the general pattern showed that there were increase in the "no" responses. Items B3 and B4 showed slight decrease in the "no" responses.

Average of scales by individuals: In <*Table 20*>, the change in average of each scale is shown. The first 6 scales in the table are the pairs of scales being compared; TA1 and TA2, SI1 and SI2, and HS1 and HS2. They are arranged in the order of the distance between each pair; TA1 and TA2 were the farthest apart, with 393 Likert-scale items, 20 open questions, and 6 stem questions; next in line are the SI1 and SI2, with 325 Likert-scale items, 20 open questions, and 6 stem questions; last but not the least, HS1 and HS2, with 257 Likert-scale items, 20 open questions, and 6 stem questions.

Table 20. Average of scales of individuals by location.

	# of questions in between	First Half (s.d.)	Second Half (s.d.)	Effect Size	*p*	FDR- *p*
TA1	L:393, O:20, S:6	2.62(0.66)	2.59(0.78)	-.05	.742	.938
TA2	L:393, O:20, S:6	2.86(0.69)	2.82(0.74)	-.06	.655	.917
SI1	L:325, O:20, S:6	2.71(0.26)	2.75(0.27)	.12	.377	.881
SI2	L:325, O:20, S:6	2.75(0.28)	2.79(0.37)	.12	.362	.881
HS1	L:257, O:20, S:6	2.78(0.27)	2.83(0.27)	.16	.234	.819
HS2	L:257, O:20, S:6	1.99(0.76)	2.14(0.78)	.19	.155	.722
SE	L:211	3.12(0.32)	3.04(0.35)	-.24	.078	.549
FT	L:189	3.08(0.50)	2.94(0.50)	-.29	.033	.462
CG	L:167	3.07(0.33)	3.06(0.33)	-.02	.871	.938
PO	L:124	2.89(0.50)	2.88(0.47)	-.03	.853	.938
C	L:117	3.04(0.47)	3.00(0.42)	-.09	.522	.913
CC	L:87	3.03(0.23)	3.01(0.22)	-.10	.463	.913
N	L:69	2.90(0.49)	2.90(0.45)	.01	.953	.953
SD	L:9	3.01(0.35)	3.04(0.36)	.06	.644	.917

Note. L is for Likert scale item, O is for open-ended question, S is for stem question.

The pairs of scales were placed at the beginning and end of the survey to see if the participants answer less attentively than before when the same contents are asked again later, but only with different wording. The p-value was corrected with FDR p-value to correct type I errors. Unlike the results found for the open questions, nothing was significant. It was interesting to note that p-value before FDR correction for SE and FT from the Academic Motivation scale were close

Table 21. Standard deviation of scales of individuals by location.

	# of questions in between	First Half (s.d.)	Second Half (s.d.)	Effect Size	*p*	FDR- *p*
TA1	L:393, O:20, S:6	0.98(0.25)	0.86(0.32)	-.40	.005	.033*
TA2	L:393, O:20, S:6	1.03(0.29)	0.96(0.31)	-.24	.082	.335
SI1	L:325, O:20, S:6	1.21(0.26)	1.15(0.33)	-.21	.127	.356
SI2	L:325, O:20, S:6	1.11(0.26)	1.07(0.30)	-.15	.279	.424
HS1	L:257, O:20, S:6	1.29(0.37)	1.23(0.44)	-.15	.261	.424
HS2	L:257, O:20, S:6	0.66(0.35)	0.67(0.38)	.02	.870	.937
SE	L:211	1.15(0.32)	1.08(0.34)	-.23	.096	.335
FT	L:189	1.06(0.30)	0.93(0.30)	-.43	.002	.024*
CG	L:167	1.04(0.32)	1.02(0.30)	-.09	.526	.670
PO	L:124	1.00(0.26)	0.96(0.28)	-.15	.276	.424
C	L:117	1.09(0.27)	1.06(0.25)	-.14	.303	.424
CC	L:87	1.22(0.33)	1.20(0.32)	-.03	.798	.931
N	L:69	1.05(0.37)	1.05(0.40)	.00	.999	.999
SD	L:9	1.11(0.24)	1.07(0.29)	-.15	.264	.424

* $p<.05$.

to p=.05 level (.078, and .033, respectively). Later in the analysis, SE and FT scales often show significance along with the six main scales being compared.

Standard deviation of scales by individuals: *Table 21* shows the change in the standard deviation of each scale in long survey. The average of standard deviation increased as the items were located in the later part of survey. The change between the beginning and end of TA1, TA2, SI1, SI2, HS1, HS2 were -

0.40 (p=.033), -0.24(p=.335), -0.21 (p=.356), -0.15 (p=.424), -0.15 (p=.424), 0.02 (p=.937), respectively. Unlike the results from Experiment 1, where most of the scales showed significant change, only TA1 was statistically significant. Interestingly, Failing Tolerance sub-scale of Academic Motivation Inventory showed significant change (-0.43, p=.024), although the Self Efficacy sub-scale from the same inventory did not show statistical significance.

Entropy of scales by individuals: In <*Table 22*>, the average entropy of each scale is shown. Entropy, the disorderliness of the individual's responses, generally increased as the distance between the scales increased. For the pairs of scales being compared, except for HS2, all 5 scales showed highly significant differences between the responses in the beginning and the end.

Average, standard deviation, and entropy of scales by items: The analyses for average, standard deviation, and entropy of scales by item level did not show any significant result.

Average of missing values by individuals: *Table 23* presents average of missing values of individuals. The differences were significant, except for CG, N, PO, C. CC, which was marginally significant (p=.100). For the Test Anxiety scales, which were the farthest, the difference in the missing values for TA1 was 4.63 (p=.001), and TA2 was 5.12 (p=.000). The difference in the missing values

Table 22. Entropy of scales of individuals by location.

	# of questions in between	First Half (s.d.)	Second Half (s.d.)	Effect Size	*p*	FDR- *p*
TA1	L:393, O:20, S:6	1.14(0.26)	0.89(0.39)	-.76	.000	.000**
TA2	L:393, O:20, S:6	1.21(0.27)	1.04(0.39)	-.49	.000	.001**
SI1	L:325, O:20, S:6	1.40(0.20)	1.22(0.35)	-.62	.000	.000**
SI2	L:325, O:20, S:6	1.32(0.19)	1.19(0.35)	-.46	.000	.001**
HS1	L:257, O:20, S:6	1.25(0.25)	1.10(0.34)	-.51	.000	.001**
HS2	L:257, O:20, S:6	0.78(0.41)	0.74(0.46)	-.09	.490	.686
SE	L:211	1.24(0.30)	1.13(0.34)	-.37	.006	.015*
FT	L:189	1.16(0.32)	1.05(0.34)	-.35	.010	.020*
CG	L:167	1.16(0.31)	1.15(0.31)	-.05	.716	.835
PO	L:124	1.14(0.30)	1.08(0.34)	-.19	.167	.291
C	L:117	1.24(0.31)	1.22(0.29)	-.07	.616	.784
CC	L:87	1.26(0.32)	1.27(0.30)	.02	.873	.940
N	L:69	1.00(0.33)	1.00(0.32)	.01	.953	.953
SD	L:9	1.27(0.29)	1.23(0.31)	-.14	.302	.470

* p<.05 , ** p<.01 .

for SI1, SI2, HS1, HS2 were 8.07 (p=.000), 12.34 (p=.000), 6.87 (p=.003), 5.33 (p=.000), respectively. Also, some scales placed as “fillers” showed significance, as SE was -3.23 (p=.002), FT was 3.62 (p=.002), and SD was 3.47 (p=.010).

Table 23. Average of missing values of individuals by location.

	First Half	Second Half	*p*	First Half	Second Half	Effect Size	*p*
	Ratio of completed surveys	Ratio of completed surveys		Avg. of missing data	Avg. of missing data		
TA1	0.93	0.89	.369	1.12	5.75	4.63	.001**
TA2	0.95	0.85	.211	1.60	6.72	5.12	.000**
SI1	0.89	0.83	.273	1.15	9.22	8.07	.000**
SI2	0.89	0.84	.849	1.55	13.89	12.34	.000**
HS1	0.97	0.92	.867	1.25	8.12	6.87	.003**
HS2	0.96	0.85	.949	1.00	6.33	5.33	.000**
SE	0.92	0.84	.073	8.70	5.47	-3.23	.002**
FT	0.89	0.92	.403	5.00	8.62	3.62	.002**
CG	0.87	0.89	.635	6.27	7.64	1.37	.200
PO	0.92	0.90	.752	10.30	9.70	-0.60	.700
C	0.84	0.81	.522	9.26	8.15	-1.11	.200
CC	0.92	0.89	.435	10.40	8.42	-1.98	.100+
N	0.92	0.92	.832	5.12	4.30	-0.82	.400
SD	0.86	0.91	.233	10.53	14.00	3.47	.010**

+ p<.10 , * p<.05 , ** p<.01 .

Straight-line response (Herzog and Bachman, 1981): Using the method described above in Experiment 1, the non-response items were included as responses in the analysis. As can be seen in <*Table 24*>, the differences by the

Table 24. Straight-line response by location.

	# of questions in between	First Half (s.d.)	Second Half (s.d.)	Effect Size	*p*	FDR- *p*
TA1	L:393, O:20, S:6	4.62(2.95)	7.14(4.63)	.66	.000	.000**
TA2	L:393, O:20, S:6	4.13(2.93)	5.63(4.70)	.38	.004	.019*
SI1	L:325, O:20, S:6	4.45(4.40)	7.44(9.82)	.40	.005	.019*
SI2	L:325, O:20, S:6	4.03(1.98)	6.50(9.30)	.36	.005	.019*
HS1	L:257, O:20, S:6	2.93(1.91)	4.08(4.05)	.37	.009	.026*
HS2	L:257, O:20, S:6	7.89(5.68)	8.13(6.18)	.04	.771	.981
SE	L:211	5.36(3.90)	6.65(4.81)	.30	.030	.071+
FT	L:189	4.71(2.84)	5.36(3.39)	.21	.130	.261
CG	L:167	5.08(4.37)	5.32(4.34)	.06	.679	.951
PO	L:124	5.45(4.68)	6.35(5.89)	.17	.212	.371
C	L:117	6.13(7.55)	6.13(7.81)	.00	.985	.985
CC	L:87	4.62(5.11)	4.60(4.65)	-.01	.969	.985
N	L:69	3.38(2.39)	3.34(2.37)	-.02	.903	.985
SD	L:9	6.16(7.67)	6.68(8.27)	.06	.629	.951

+ p<.10 , * p<.05 , ** p<.01 .

placement in the survey showed significant results for the tests being compared, except for HS2. The p-values for the scales TA1, TA2, SC1, SC2, and HS1 were .000, .019, .019, .019, and .026, respectively.

Middle response (3 on 5-point Likert scale) frequency: As can be seen in <*Table 25*>, there is a pattern that when the test is placed in the end, twice as

Table 25. Number of individuals who answered response 3.

	All "3"		Response 3 (>10)		*p*
	First Half	Second Half	First Half	Second Half	
TA1	1	1	4	8	.001**
TA2	1	0	7	13	.286
SI1	1	1	9	20	.029*
SI2	0	0	15	20	.099+
HS1	1	1	5	10	.255
HS2	1	1	9	7	.148

+ p<.10 , * p<.05 , ** p<.01 .

many participants choose response 3. Except for HS2, which actually decreased in the number of individuals, and SI2, which only increased by 33%, all the other inventories being compared showed doubling pattern. The significant results were observed for TA1 (p=.001) and SI1 (p=.029), and SI2 was marginally significant (p=.099).

Middle response (2-3-4 on Likert scale) frequency: The results for the in-the-middle response frequency including 2, 3, and 4 on Likert scale for Experiment 2 is shown in <*Table 26*>. All inventories being compared showed differences as predicted. In particular, in TA1 (p=.075), SI1 (p=.095), and HS1 (p=.005) the participants who answered 2-3-4 responses were twice more when they were located in the first part of the survey than in the second part.

Table 26. Number of individuals who answered responses 2-3-4.

	All 2-3-4		*p*
	First Half	Second Half	
TA1	13	26	.075+
TA2	10	30	.007**
SI1	6	11	.098+
SI2	3	16	.218
HS1	15	27	.005**
HS2	18	26	.354

+ p<.10 , * p<.05 , ** p<.01 .

TA2 (p=.007) and SI2 (p=.218) had three times more participants, and HS2 (p=.354) had 44% increase.

Discussion

Unlike the results for experiment 1, the results for experiment 2 included all the inventories used in the survey to see how the change occurs. The assumption was that, the bigger the differences in the locations of items between Type A and B are, the bigger differences in response in response quality will be observed. In other words, the closer the items are positioned from Type A to Type B, the difference will be less..

The result for missing values was the most impressive since the gradual change was very clear. All the scales for comparison, 2 test anxiety scales, 2 self

identity scales, and 2 hopelessness scales, showed highly significant differences($p<.01$). Even some inventories used as fillers, such as academic motivation scale (SE and FT) and acceptance desire scale (SD) showed differences that were highly significant ($p<.01$).

The reason for these scales to yield significant differences is because of their location in the survey. Although they acted as fillers, they were quite far apart from type A to type B, and although they did not show much noticeable differences in other indices, for missing values analysis the difference was very clear. And to prove that the placement in the survey matters, the scales which did not have much distance from type A to type B did not show any significant differences (see *Table 24*).

Also, the academic motivation scale showed some interesting pattern throughout all the analyses. Both the self efficacy scale (SE) and failing tolerance scale (FT) from academic motivation scale showed significant decrease in entropy along with other tests for comparison (TA1, TA2, SI1, SI2, HS1), and significant increase in missing values along with all 6 tests being compared.

Failing tolerance scale showed a significant decrease in standard deviation along with test anxiety scale (TA1), and self-efficacy scale had a decrease for standard deviation and an increase for straight-line response which

were marginally significant. It is probably because it was placed right after the 6 tests being compared in Type A, that it has showed some meaningful differences. However, acceptance desire scale (SD) did not show such differences even though it was placed in the same area as SE and FT in type B.

It was unexpected for the Hopelessness scale (HS2) not to show significant difference along with all the other tests for comparison. It may be due to its location in the survey. In type A, participants encounter 88 Likert scale questions, 10 open questions, 3 stem questions, 237 Likert scale questions, 10 open questions, and 3 stem questions before doing 20 Likert scale items on hopelessness scale (HS2). When the participants encounter HS2 after doing 10 open questions and 3 stem questions, perhaps they become “refreshed” from doing 237 Likert scale items and therefore do not respond less carelessly. Even a few minutes of “refreshment” can help attention span. However, that still does not explain why HS1 showed some meaningful differences while HS2 didn’t, since HS1 was also placed in the same area as HS2 in type B. Some unbalanced observations were observed, which needs further studies.

Both for experiment 1 and experiment 2, the standard deviation and entropy of scales by individual level showed decrease when items were located at the second part of survey. It means that the responses showed less variability. These results may be related to straight-line responses. The straight-line

responses for both the short survey and the long survey increased toward the latter part of the survey, which means the participants responded in a “straight-line” fashion, which may signify fatigue or boredom. It is safe to assume that the participants who had longer straight-line responses had less standard deviation and entropy.

Lastly, results of entropy analyses are consistent between Experiment 1 and 2. The farther the pairs of the scales were from each other, the effect size was bigger. For example, in Experiment 1, TA1 had bigger effect size than SC1, and TA2 had bigger effect size than SC2. TA1 and TA2 both showed significant differences while only SC2 showed significant difference. In Experiment 2, TA1 had bigger effect size than SI1, and SI1 had bigger effect size than HS1. Likewise, TA2 had bigger effect size than SI2, and SI2 had bigger effect size than HS2. The 2 Test Anxiety scales and the 2 Self Identity scales showed significant differences while only HS1 showed significant difference. It suggests that the shorter the distance between the pair of tests being compared, lesser the effect size, and less likely for the difference of the entropy to be significant. Same pattern emerged for straight-line responses as well.

It would be safe to argue that TA1 and TA2 showed significant difference between the locations in the survey because TA1 was located in the beginning of the Type A survey at the end of the Type B survey, which were the

farthest apart in the counterbalanced design. The location of the survey matters as SC1 and SC2 did not show such consistent result.

General Discussion

To summarize, the purpose of the present study was to test the effects of the survey length on the response quality. Many indices were applied to probe these effects. The reliability coefficient, Cronbach α, did not show any effect, stem-and-branch questions showed predicted pattern but was not statistically significant, open-ended questions showed significant difference between two locations of items, along with standard deviation and missing values analyses.

This study proposed some new measurements: the Person fit index, entropy and straight-line response and middle response analyses. They showed meaningful differences in response quality between two different locations by individual level. However, average of scales by individuals did not show any significant result. .

The open-ended questions were found to be effective in proving that the quality of responses decreases as the survey length increases. Stem-and-branch questions need more probing since the attenuation mentioned by Duan's team was not replicated in the current research (Duan et al., 2006, Jensen et al., 1999, Shaffer et al., 1996). Some studies on medical diagnosis questionnaires contain stem-and-branch questions with more than 10 branch questions and the survey

length had effect on those stem-and-branch questions (Duan et al., 2006). Perhaps a variety of stem-and-branch questions with different sets of branch questions, such as 3, 5, 7 branch questions, can be presented to study what effect stem-and-branch questions have in various places in the surveys.

The most unexpected finding was the increased reliability coefficient. Though the increase was not statistically significant, the fact that reliability increased even at the end of the survey is difficult to understand. One speculation is that these results may be due to the content of item. The items in the survey are all about the self. According to Helgeson and Ursic (1994), the content of the questions matters: items related to self and personality do not decrease in reliability as the survey length increases, but items with issues such as politics and religions may lose the interest of the test-takers and therefore decrease reliability.

Looking closely at the middle responses for Experiment 1 and 2, some patterns emerge that are similar to the standard deviation, entropy, and straight-line responses. If we combine the middle response analyses of response 3 and response 2-3-4, we find that TA1, TA2, and SC2 have attained significantly increased number of middle responses as the items were located at the later part of survey. The same pattern follows for standard deviation, entropy, and straight-line analyses in Experiment 1. For Experiment 2, when the results for the

response 3 and the response 2-3-4 analyses are combined together, TA1, TA2, SI1, and HS1 show significant increase.

It would be safe to argue that distance has effect on the individual's responses. For example, in Experiment 1, TA1 and TA2 showed significant difference between the locations in the survey because TA1 was located in the beginning of the Type A survey at the end of the Type B survey, which were the farthest apart in the counterbalanced design. The location of the survey matters as SC1 and SC2 did not show such consistent result. Only SC2 showed significant difference while SC1 did not.

Likewise, for Experiment 2, same patterns emerge as the distance between the scales become farther apart. TA1 and TA2 both show significant increase while only one of the pair, SI1 and HS1, show significant increase in the middle responses. They are both located in the beginning of Type A survey after TA1. Perhaps bigger sample size is required to investigate whether this is just a coincidence that one type shows significance while the other doesn't.

The contribution of this research was that it proposed some new statistical indices into the field of psychology for measuring psychological wariness, and they are entropy of scales, straight-line responses and middle responses analyses. Different indices seem to detect difference aspects of

response quality. The use of multiple indices that reflect different aspects of response quality will help understand the characteristics of responses when survey is very long.

The weakness of this research is that the items in the survey were not diverse, both in the contents of issues presented and the types of stem questions. For future research, a variety of contents related to different issues can be presented to investigate the effect of survey length on the reliability of the survey as well.

As mentioned by Cape (2010) in his presentation at ARF Re:think, data quality suffers as survey length increases, although it does not necessarily cause more drop-out. It would be interesting to see the drop-out rate by time component on future research as well.

References

Adams, G.R. (1985). Family correlates of female adolescents' ego identity development. *Journal of Adolescence, 8*, 69-82.

Adams, G.R., Bennion, L.D., & Hugh, K.S. (1989). *Objective Measure of Ego Identity Status: A Reference Manual.* Logan: Utah State University.

Berry, D. T., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review, 11*, 585-598.

Beatty, P. & Herrmann, D. (2002). "To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse." in Groves, R. M., Dillman, D. A., Eltinger J. L., and R.J.A. Little. (eds.), *Survey Nonresponse* (pp. 71-85). New York: Wiley.

Beck, A. T., Weissman, A., Lester, D., and Trexler, L. (1974). The Measurement of Pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology, 42*, 861-865.

Bennion, L.D. & Adams, G.R. (1986). A revision of Extended Version of the Objective Measure of Ego Identity Status: an identity instrument for use with late adolescent. *Journal of Adolescents Research, 1*, 183-198.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57,* 289-300.

Biemer, P.P. (1991). *Measurement errors in surveys.* New York: Wiley.

Birenbaum, M. (1986). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement, 45,* 523-534.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bogen, K. (1996). "The Effect of Questionnaire Length on Response Rates – A Review of the Literature." Proceedings of the Survey Research Methods Section of the American Statistical Association, 1020-1025.

Burchell, B. and Marsh, C. (1992). The Effects of Questionnaire Length on Survey Response. *Quality & Quantity, 26*, 233-244.

Cape, P. (2010). Questionnaire length, fatigue effects, and response quality revisited. *Survey Sampling International*. Presented at ARF Re:think 2010.

Choi, Jeong Hoon & Lee Jung Yoon. (1994). Irrational beliefs and situational factors in social anxiety. *Korean Journal of Counseling and Psychology, 6*, 21-47.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Deutskens, E., De Ruyter, K., Wetzels, M., and Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: an experimental study. *Marketing letters, 15,* 21-36.

Duan, N., Alegria, M., Canino, G., McGuire, T. G., and Takeuchi, D. (2007). Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats. *Health services research, 42,* 890-907.

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Fagerland, M. W., and Sandvik, L.. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary clinical trials, 30,* 490-496.

Ferrando, P., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement*, *61,* 997-1012.

Galesic, M., and Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73,* 349-360.

Goetz, E. G., Tyler, T. R., and Cook, F. L. (1984). Promised incentives in media research: a look at data quality, sample representativeness, and response rate. *Journal of Marketing Research, 21,* 148-154.

Heberlein, T. and Baugmartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review, 43*, 447-462.

Helgeson, J.G. and Ursic, M.L. (1994) The Role of Affective and Cognitive Decision-Making Processes during Questionnaire Completion. *Psychology & Marketing 11*, 493-510.

Herzog, A.R., and Bachman, J.G. (1981). Effects of Questionnaire Length on Response Quality. *Public Opinion Quarterly, 45,* 49-59.

Hwang, Kyung Ryul. (1997). Comparison of three training methods of reducing test anxiety: behavioral method, cognitive method, combination of cognitive and behavioral methods. *Korean Journal of Counseling and Psychology, 9*, 57-80.

Jang, Eun Kyung. (2010). Response-faking detection of the Korean version MMPI-2 by IRT person-fit and Markov chain model. Seoul: Department of Psychology, Seoul National University.

Jensen, P. S., Watanabe, H. K., and Richters, J. E. (1999). Who's up first? Testing for order effects in structured interviews using a counterbalanced experimental design. *Journal of Abnormal Child Psychology, 27,* 439-436.

Jeon, Mi Hyun. (2013). Effects of perfectionism, self-efficacy, and conscientiousness on procrastination. Kyungsan: Department of Psychology, Daegu University.

Johnson, W.R., Seiveking, N.A., and Clanton, E.S. (1974). Effects of alternative positioning of open-ended questions in multiple-choice questionnaires. *Journal of Applied Psychology, 59*, 776-778.

Johnson, O., Sejdinovic, D., Cruise, J., Piechocki, R., and Ganesh, A. (2013). Non-Parametric Change-Point Estimation using String Matching Algorithms. *Methodology and Computing in Applied Probability*, 1-22.

Kang, Y., Harring, J. R., and Li, M. (2014). Reexamining the Impact of Nonnormality in Two-Group Comparison Procedures. *The Journal of Experimental Education* (ahead-of-print), 1-28.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277-298.

Keum, Myung Ja. (1984). The effect of counselor's self-disclosure on client's self-disclosure under client's need for approval. Seoul: Department of Psychology, Seoul National University.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association, 107*, 1590-1598.

Killick, R., and Eckley, I. (2013). *Changepoint: An R package for changepoint analysis.* R package version, 1.

Kim, Ah Young. (2002). A study of standardizing Academic Motivation Scale. *Journal of Educational Evaluation, 15*, 157-184.

Kim, Deuk Ran. (1993). A study on the sex typed response styles and related variables in androgynous males and females. Seoul: Department of Psychology, Sung Kyun Kwan University.

Kim, Hwa Ja. (1998). A study of differences in dysfunctional family structure and social problem-solving ability in college students' ego-identity status. Seoul: Department of Psychology, Yonsei University.

Kim, Ji Hye. (1991). Effect of self-focused attention on anxiety. Seoul: Department of Psychology, Korea University.

Kim, Wan Seok. (1994). Korean version of Need for Cognition Scale. *The Korean Journal of Industrial and Organizational Psychology, 7*, 87-101.

Kim, Young Mi. (1995). 문제해결 전략에 의한 문제의 분류 및 문제해결 지도과정 연구. Chuncheon: Department of Education, Unpublished masters thesis, Kangwon National University.

Krosnick, J.A. (1999). Survey Research. *Annual Review Psychology, 50*, 537-567.

Krosnick, J.A., Holbrook, A.L., Berent, M.K., Carson, R.T., Hanemann, W.M., Kopp, R.J., … Conaway, M. (2002). The impact of 'no opinion'

response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly, 45*, 549-559.

Lambert, D., (1992). "Zero-inflated Poisson Regression, With an Application to Defects in Manufacturing." *Technometrics, 34,* 1-14.

Lee, Hoon Jin. (1997). Self-concept and attributional style in paranoia. Seoul: Department of Psychology, Seoul National University.

Lee, Young Ho. (1993). The relations between attributional style, life events, event attribution, hopelessness and depression: A covariance structure modeling approach. Seoul: Department of Psychology, Seoul National University.

Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling error in surveys.* Wiley.

Levine, M. V., and Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4,* 269-290.

Li, M. F., and Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21,* 215-231.

Loevinger, Jane. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51,* 493.

Lord, F.M. and Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.

Marsden, P.V. and Wright, J.D. (2010). *Handbook of survey research.* 2nd ed. Emerald Press.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156.

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics, 33,* 341-365.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective.* MIT Press.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21,* 115-127.

Nichols, D. S., Greene, R. L., and Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology, 45,* 239-250.

Nunnally, J.C. and Bernstein, I.H. (1994). *Psychometric theory.* McGraw-Hill., Inc.

Osborne, J. W. (2008). "Best Practices in Data Transformation" in Osborne, J. W. (ed.), *Best practices in quantitative methods* (pp. 197-204). Sage.

Park, Ah Cheong. (1996). Preliminary development of Korean adolescent Ego-Identity Scale. *The Korean Journal of Clinical Psychology, 15*, 140-162.

Reise, S. P. and Due, A. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217-226.

Reise, S. P. and Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65,* 143-151.

Rossi, P.H., Wright, J.D. and Anderson, A.B. (1983). *Handbook of survey research.* Academic Press.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph, 34,* 100-114.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennigs, D. (1999). Correlates of Person Fit and Effect of Person Fit on Test Validity. *Applied Psychological Measurement, 23,* 41-53.

Shannon, C.E. and Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press.

Song, Sul Hee. (1994). The effects of family type and perceived parental acceptance-rejection on Korean adolescents' ego-identity formation. Daejeon: Department of Psychology, Chungnam National University.

Spielberg, C.D., Gonzalez, H.P., Taylor, C.J., Anton, W.D., Algaze, B., Ross, G.R., and Westberry, L.G. (1980). *Preliminary manual for the test anxiety inventory*. California: Consulting Psychologist Press.

Tarendash, A.S. (2013). *Chemistry: The Physical Setting.* Barron's Educational Series, Inc.

Urbina, S. (2004). *Essentials of Psychological Testing.* Hoboken, NJ: John Wiley.

Weisberg, H. F. (2009). *The total survey error approach: A guide to the new science of survey research*. University of Chicago Press.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika. 29,* 350–362.

Yammarino, F.J., Skinner, S.J., and Childers, T.L. (1991). Understanding Mail Survey Response Behavior. *Public Opinion Quarterly, 55*, 613-639.

Zhang, X. S., Zhu, Y. S., and Zhang, X. J. (1997). New approach to studies on ECG dynamics: extraction and analyses of QRS complex irregularity time series. *Medical and Biological Engineering and Computing. 35,* 467-473.

국 문 초 록

본 연구에서는 설문조사의 비표집 오차의 원인 중 하나인 설문지의 길이가 응답에 미치는 영향을 조사하였다. 기존의 사용된 신뢰도, 응답의 평균, 표준편차, 결측치 비율 지표에 더불어, 문항반응이론에 기반한 개인별 우도, 엔트로피, 최대 한줄 길이, 중간 반응 빈도 지표들이 사용되었다. 실험 1 에서는 짧은 설문지를 사용하였고, 192 문항에 30 분이 소요되었다. 실험 1 의 결과, 후반부의 문항들에 대해 개인별 표준편차, 개인별 엔트로피가 유의미하게 낮아졌고, 주관식 질문에 대한 응답의 길이는 유의미하게 짧아졌다. 최대 한줄 길이와 중간 반응 빈도도 뒤로 갈수록 증가하였다. 실험 2 에서는 긴 설문지를 사용하였고, 505 문항에 60 분이 소요되었다. 실험 2 의 결과, 후반부의 문항들에 대해 개인별 표준편차, 엔트로피, 주관식 질문의 지표들은 유의미하게 낮아졌고, 결측치 비율, 최대 한줄 길이, 중간 반응 빈도 지표들은 유의미하게 증가하였다.

주요어 : 문항 위치, 응답의 질, 설문의 길이, 설문의 응답, 비표집 오차

학 번 : 2009-22833