Survey Research Methods (2010)
Vol.4, No.3, pp. 139-149
ISSN 1864-3361
http://www.surveymethods.org

© European Survey Research Association

Elaborate Item Count Questioning: Why Do People Underreport in Item Count Responses?

Takahiro Tsuchiya
The Institute of Statistical Mathematics

Yoko Hirai
Tokyo Metropolitan University

The item count technique, used often to investigate illegal or socially undesirable behaviours, requires respondents to indicate merely the number of applicable items from among a list. However, the number of applicable items indicated via the item count question tends to be smaller than when it is calculated from the direct 'applies/does not apply' responses to each item. Because this inconsistency, which we refer to as the underreporting effect, often disturbs proper item count estimates, the causes of this effect are explored in this paper. Web survey results revealed that the order of the response alternatives is irrelevant to the underreporting effect, and that the underreporting effect is caused by the response format in which the item count question requests merely the number of applicable items and not the number of non-applicable items. It is also shown that the magnitude of the underreporting effect decreases when the respondents are asked to indicate the numbers of both applicable and non-applicable items, which we refer to as elaborate item count questioning.
Keywords: indirect questioning technique, item count technique, underreporting effect, order effect, check-all-that-apply

Introduction

Underreporting effect in item count response

For sensitive research topics, such as involvement in illegal or socially undesirable behaviours, we cannot expect every respondent to honestly answer 'applies' or 'does not apply' to ordinary direct questions. Indirect questioning techniques, including the randomized response (Warner, 1965), are often employed in such situations as alternatives to the ordinary direct questioning (hereafter, DQ) technique. Because the indirect questioning technique basically conceals each respondent's actual status as to the key sensitive topic, it is supposed to elicit more truthful answers than the ordinary DQ. Among the indirect questioning techniques, the item count technique (Droitcour et al., 1991), which is sometimes referred to as the unmatched count technique (Dalton, Wimbush, and Daily, 1994; Dalton, Daily, and Wimbush, 1997; Wimbush and Dalton, 1997) or the list experiment (Sniderman and Grob, 1996), has recently attracted much attention. The procedure of the item count (hereafter, IC) technique is as follows. Let the target key item be 'having driven after drinking alcohol'. A set of non-key items, which usually comprises three to five items such as 'having travelled to Egypt' or 'having gotten a speeding ticket', is prepared in addition to the target key item. A group of respondents is divided into two homogeneous subgroups. One subgroup is asked to mentally count the number of items that apply to them from among the key and non-key items. Because the respondents are requested to indicate only the number of applicable items, it remains unknown whether a respondent has driven after drinking alcohol, except when his or her answer is 'none' or 'all'. The other subgroup is also asked to count the number of applicable items from among merely the non-key items. The proportion of people who have driven after drinking alcohol is estimated from the difference in the mean numbers of the applicable items between the two subgroups.

Although the IC procedure is simple and easy to understand for most respondents, its estimates are often unstable. Droitcour et al. (1991) found smaller prevalence estimates for drug use by the IC technique than by the DQ technique. Biemer and Wright (2004) also found the IC estimates to be smaller than the DQ estimates for cocaine abuse. Tsuchiya, Hirai, and Ono (2007a) achieved a negative estimate for the proportion of shoplifters.

The factors which might affect the performance of the IC include the selection of non-key items and the sample size. When the number of non-key items is small, it is insufficient to conceal the respondent's true status as to the key item. On the other hand, when the number of non-key items is large, a comparatively larger sample size than the DQ is required because the variance of the mean response increases. In addition, Tsuchiya, Hirai, and Ono (2007b) advocated that the response effect, which we refer to as the underreporting effect in this paper, could be another cause for the IC instability, as discussed below. With regard to this effect, the mean number of applicable items in the IC responses is smaller than that calculated from the conventional 'applies/does not apply' responses to each item in the list. Biemer and Wright (2004), who compared the IC and DQ responses of the same respondents, found that a non-negligible number of respondents selected 'none is applicable' to the IC question, although they selected 'applies' to all the items in the list. Tsuchiya et al. (2007b) pointed out that the underreporting effect could ren-

Contact information: Takahiro Tsuchiya, The Institute of Statistical Mathematics, e-mail: taka@ism.ac.jp

der the IC technique useless if the absolute magnitude of underreporting increases as the item list becomes longer. The IC estimate is calculated by subtracting the mean responses of the short list, which excludes the key item, from the mean responses of the long list, which includes the key item. If the absolute magnitudes of underreporting are the same in both short and long lists, the IC estimates will not be influenced by the underreporting effect. However, if the absolute magnitude of the underreporting effect in the long list is larger than that in the short list, the IC estimate will shrink, or worse, become negative. Moreover, although a longer non-key item list is desirable to conceal the respondents' true status as to the key behaviour, a longer list is more likely to destabilise the IC estimator because of the underreporting effect. Empirically, Tsuchiya et al. (2007a, 2007b) found that the underreporting effects actually increase in longer lists, although at the same time, Tsuchiya et al. (2007b) arrived at the perplexing result that the underreporting effect is smaller in the long list than in the short list.

One may suspect that the underreporting effect is because of the sensitive contents of the item lists. However, this effect is observed not only in sensitive item lists but also in non-sensitive lists that comprise items such as having a passometer or donating blood (Tsuchiya et al., 2007a). Rather, a larger underreporting effect is observed more frequently in non-sensitive item lists than in sensitive ones. This is because the number of applicable items shrinks in both IC and DQ responses in the case of the sensitive item lists, whereas it does not shrink in the DQ responses in the case of the non-sensitive item lists. Thus, the underreporting effect is not the outcome of the sensitivity of the item lists. Furthermore, because the underreporting effect occurs for non-sensitive item lists, the IC technique yields smaller estimates than the DQ technique even for the non-sensitive key item (Tsuchiya et al., 2007b). The strategy of handling the underreporting effect determines the success and failure of the IC technique.

To address the underreporting effect, one approach is to statistically correct the effect to produce a model-based estimator (Biemer and Brown, 2005, see Cruyff, van den Hout, van der Heijden and Böckenhold (2007) in the case of the randomized response). Another approach is to improve the questionnaire such that it can elicit IC responses that are comparable to the ordinary DQ responses. In this paper, we address the underreporting effect through the latter approach. That is, we conduct experimental surveys to detect the causes of the underreporting effect and to develop methods for suppressing it.

Before proceeding to the next section, we would like to remark on the term 'underreporting effect'. Until now, it has not been confirmed which responses are more similar to the true status between DQ and IC using external criteria. It is possible that the respondents overreport the applicable items in the DQ technique. However, a difference between the DQ and IC responses is observed even in the list that inquires about facts such as 'having siblings' or 'smoking cigarettes', as shown later. It is unlikely that a large number of respondents answer 'applies' to such well-defined items by mistake. Furthermore, people are familiar with the ordinary DQ questions. In contrast, an IC question is uncommon, and people are not used to answering merely the number of applicable items. An IC question demands more mental processing than a DQ question. Hence, because the difference is more likely to be attributable to the IC response than to the DQ response, we refer to the difference as the underreporting effect. Moreover, because our aim in this study is to explore a method to elicit an IC response that is comparable to the DQ response, we focus on the smallness of the IC response.

Theoretical background

In this paper, we focus on two hypotheses for explaining the underreporting effect. The first hypothesis is that underreporting is attributable to the response order effect. The IC responses are usually selected from alternatives, each of which describes the possible number of applicable items. An open-ended response format is not used in order to avoid inadequate responses such that more than the maximum number of items is applicable. However, when response alternatives are used, it is possible for the response order effect to emerge (Schuman and Presser, 1981). If the primacy effect occurs for the response alternatives sorted in the ascending order, the number of applicable items will be smaller compared to the case in which the primacy effect does not occur. The recency effect for the alternatives sorted in the descending order also causes underreporting. If this hypothesis of the response order effect is true, the underreporting effect will diminish or an overreporting effect will be observed when the alternatives are rearranged in the reverse order.

The second hypothesis is that the wording of the IC question is such that the respondents are not explicitly required to select 'applies' by comparing it to 'does not apply'. Because the question is usually worded like, 'How many of the items listed below are applicable to you?', the IC question calls the respondents' attention to merely the applicability of items and does not demand the respondents' attention to the non-applicability of items. The respondents are not forced to select between 'applies' and 'does not apply'; they may not mentally consider both options equally. Because the wording of the question merely refers to the applicability of items, respondents consider whether the selection of 'applies' is adequate. When the respondents select 'applies', it implies that they have judged that the item certainly applies. However, when they do not select 'applies', it does not mean that they have judged that 'does not apply' is more suitable than 'applies'; it merely implies that they believed that 'applies' is not adequate. The options 'applies' and 'does not apply' are not compared on an equal basis. Thus, in such cases, the non-selection of 'applies' does not necessarily mean the selection of 'does not apply'.

On the other hand, the DQ question requires the respondents to decide which option –'applies' or 'does not apply'– is more appropriate to describe their true status. The answer 'does not apply' to the DQ question implies that the respondent decided that 'does not apply' is more adequate for expressing his/her true status than 'applies'.

Apart from the IC question, one may suppose that ex-

plicit 'check-all-that-apply' responses are logically comparable to 'applies/does not apply' or 'yes/no' responses to a set of items. However, many researchers have suggested avoiding 'check-all-that-apply' questions because they have a number of drawbacks. Sudman and Bradburn (1982) advocate the use of a forced-choice format, i.e. 'applies/does not apply' or 'yes/no' format, instead of the 'check-all-that-apply' format, because it is difficult to interpret what the absence of a check indicates. Dillman (2007:62) and Morrison, Dillman and Christian (2010) also support avoiding 'check-all-that-apply' questions in order to reduce the primacy effects. Furthermore, in the past two decades, it has been repeatedly uncovered that 'yes/no' questions endorse more options than 'check-all-that-apply' questions even when the responses to each item are expressed explicitly (Rasinski, Mingay, and Bradburn, 1994; Mitofsky and Edelman, 1995; Dillman, Smyth, Christian, and Stern, 2003; Smyth, Dillman, Christian, and Stern, 2006; Smyth, Christian, and Dillman, 2008; Thomas and Klein, 2006). All these research studies illustrated that the number of applicable items in the forced-choice format is significantly greater than that in the 'check-all-that-apply' format. The answers to the two types of questions are not comparable. Smyth et al. (2006) argue that this is because the respondents who answered 'check-all-that-apply' questions employed a weaker satisficing response strategy.

Our hypothesis is that the IC question requires the respondents to mentally check all that apply and to express the number of items they mark. This is because the IC question requests merely the number of applicable items and not the number of non-applicable items. Because the number of applicable items is fewer in the 'check-all-that-apply' format, the underreporting effect emerges when we compare the IC responses to the DQ responses that are reported in a forced-choice format. If this hypothesis is true, the IC response should be equal to the number of marked items in the explicit 'check-all-that-apply' format. Otherwise, the IC response should be equal to the number of applicable items in the forced-choice DQ format when the IC question requires the number of both applicable and non-applicable items, which we refer to as elaborate item count questioning.

Research design

Preparing item lists

Estimation of the proportion of sensitive behaviours is not our aim in this study, and hence we do not use sensitive item lists. As mentioned above, the underreporting effect emerges even for non-sensitive item lists. If we use sensitive ones, the DQ responses will be disturbed, and it will be more difficult to examine the two hypotheses. The DQ responses to the non-sensitive item lists are more reliable than those to the sensitive lists as bases for comparison. Furthermore, we use not only well-defined item lists but also ill-defined item lists, i.e. we include such item lists as those on which respondents wavered in deciding their responses; this is because ill-defined item lists are more appropriate for our experiments. To estimate the proportion of key behaviour in the IC technique, non-key items should be well-defined, clearly decidable ones. Although well-defined item lists also produce the underreporting effect, the magnitude of this effect is supposed to be smaller than that in the ill-defined item list. When the underreporting effect is small, relatively large samples are necessary to detect statistically significant differences between various experimental conditions. Ill-defined ambiguous item lists are supposed to produce a larger underreporting effect. Hence, the cause of the underreporting effect is expected to be specified more easily in the ill-defined item lists than in the case of well-defined item lists.

We prepared forty items as candidates. Half of the items comprised statements that most people would respond to without hesitation, and the other half comprised statements that would cause most people to think carefully about before responding. The questionnaire first asked about the applicability of an item with the 'applies/does not apply' format. Next, the questionnaire asked respondents to rate the degree to which they wavered in their decision, using a three-point scale: 3, wavered; 2, wavered a little; and 1, did not waver at all. This procedure was repeated for over 40 item candidates.

We conducted an online survey to prepare the item lists. Respondents were randomly selected from the database of the same web survey company as in survey 1 below. As the database comprises people who registered by themselves, the respondents are not representative of the entire Japanese population or of Japanese adults. However, this does not influence the implications elicited from the present research results because the purpose is not to estimate the population parameters but to examine the occurrence of the underreporting effect based on various conditions. The selected people were invited via an e-mail to access the questionnaire website. The website was accessible until the predetermined number of respondents completed the survey. A total of 504 respondents completed the survey between 6 and 9 February 2009.

Based on the mean ratings of the degree of waver for each item, we selected 16 well-defined and 16 ill-defined ambiguous items. We then arranged them into eight item lists as shown in Table A1 of the Appendix. Four lists, from 1 to 4, comprised well-defined items—the respondents replied with certainty. The other four lists, from 5 to 8, comprised ambiguous items—the respondents were hesitant in responding.

Survey 1: Investigating order effect

The first experimental survey aimed to investigate the existence of the order effect. It was conducted online between 26 February and 3 March 2009. We prepared three types of questionnaires: the DQ, the ascending IC and the descending IC questionnaires. The DQ questionnaire shows each item list on one screen, and requires an 'applies/does not apply' response for each item with the wording of the question being 'Does each item below apply to you or does it not apply to you?' The ascending IC questionnaire also presents each item list on one screen, and the five response alternatives from 'none is applicable' to 'four items are appli-

cable' are listed downward beneath the item list. The question wording was 'Starting here, we present a few situations. Please answer the number of items that apply to you from among them. First, how many items are applicable to you from among the four items listed below?' The descending IC questionnaire was the same as the ascending IC one except that the order of the response alternatives was reversed. The eight item lists were shown alternately between the well-defined and the ill-defined lists, i.e. list 1, list 5, list 2, list 6, and so forth. Because every questionnaire does not allow the skipping of a response before proceeding to the next item list, there exists no item non-response.

A sample was randomly selected from the database of the survey company. The sample did not overlap with the respondents who answered to the survey of the item list preparation. The sample was randomly divided into three groups, and the respondents were invited to access the corresponding questionnaire website. The website was closed after the number of completed respondents exceeded the predetermined number, 667. The total number of respondents who completed the survey was 669 for the DQ group, 674 for the ascending IC group, and 673 for the descending IC group. Table 1 shows the breakup of respondents by gender and age. Although the percentages are not in complete agreement among the three groups, we could say that these three groups are homogeneous.

Survey 2: Investigating the effect of response format

The second experimental survey aimed to investigate the effect of response format. It was also conducted online between 12 and 22 February 2010. We prepared five questionnaires, and among them, two were DQ-type questionnaires and three were IC-type questionnaires. The first DQ-type questionnaire was the same as the one used in survey 1, which requests respondents to answer with 'applies/does not apply' options. We refer to this ordinary DQ as DQ. The second DQ-type questionnaire used a check-all-that-apply format with the question wording 'Please select all that apply to you among the affairs listed below'. Checkboxes were prepared at the left of each item. Because non-response was not allowed, an item 'none is applicable' was appended beneath the item list. We refer to this second DQ questionnaire as 'check-all'.

The first IC-type questionnaire was the same as the one in survey 1 except that the response alternatives were arranged in ascending order from left to right beneath the item list. It requested the respondents to answer merely the number of applicable items. We refer to this ordinary IC as IC. The second IC-type questionnaire was the same as the first IC questionnaire except that the question was 'How many items are non-applicable to you among the four items listed below?' The word 'non-applicable' was shown in red letters to get the attention of respondents. We refer to this second questionnaire as 'reverse IC'. The response alternatives displayed were exactly the same as those in the first ordinary IC questionnaire, though their meaning was completely reversed. The third IC questionnaire asked respondents to answer both the number of applicable and non-applicable items with the question wording 'How many items are applicable to you and how many items are non-applicable to you among the four items listed below?' The word 'applicable' was shown in blue letters, and the word 'non-applicable' was shown in red letters to get the respondents' attention. We refer to this questionnaire as 'elaborate IC'. The response alternatives were arranged in a matrix form with the number of applicable items in the first row. Although the sum of the number of applicable and non-applicable items should be four, we did not reject the irregular responses, the sum of which is not equal to four, because we wanted to know the proportion of such irregular responses.

A sample is randomly selected from the database of the survey company that is different from the survey 1. Hence, the samples of survey 1 and survey 2 are not comparable. In addition, because the database comprises persons who registered online by themselves, the sample of survey 2 is again not representative of the Japanese people. A sample for survey 2 is randomly divided into five groups, each of which was invited via an e-mail to access the corresponding questionnaire website. The respondents who accessed the website were first asked about their gender and age. The website was accessible until the predetermined numbers of respondents, which were set according to gender and age, accessed the websites. For each of the two DQ questionnaires, the minimal number was 50 for every five-year and gender group from ages 20 to 69, and 20 for every gender group of the teens. For each of the three IC questionnaires, the minimal numbers were 70 and 30 respectively.[1]

Results

Effect of response orders

Table 2 compares the mean number of applicable items among three groups of survey 1. Both IC responses are smaller than the DQ responses in every list. Among eight lists, four lists showed statistically significant differences between the DQ and the ascending IC groups, whereas five lists showed significant differences between the DQ and the descending IC groups. We could say that the underreporting effect also emerged in this survey. As expected, the magnitudes of underreporting are larger in the ill-defined ambiguous item lists than in the well-defined item lists. However, even in the case of the well-defined item lists, a maximum difference of $-.168$ was observed between the DQ and the IC, which might disturb the IC estimates if the list were actually used for the IC technique.

Statistically significant differences were never found between the ascending and descending IC groups in every list. The ascending IC group showed larger responses in four lists than the descending IC group, whereas in the other four lists the descending IC group showed larger responses than the

[1] Because some respondents aborted the survey, the numbers of completed respondents are slightly different from the predetermined numbers.

Table 1: Gender and age of respondents in survey 1

	DQ			Ascending IC			Descending IC		
Age group	Male	Female	Total	Male	Female	Total	Male	Female	Total
under 20	2 (0.3%)	4 (0.6%)	6 (0.9%)	4 (0.6%)	4 (0.6%)	8 (1.2%)	0 (0.0%)	3 (0.4%)	3 (0.4%)
20 - 29	16 (2.4%)	71 (10.6%)	87 (13.0%)	31 (4.6%)	68 (10.1%)	99 (14.7%)	24 (3.6%)	63 (9.4%)	87 (12.9%)
30 - 39	84 (12.6%)	133 (19.9%)	217 (32.4%)	67 (9.9%)	171 (25.4%)	238 (35.3%)	66 (9.8%)	159 (23.6%)	225 (33.4%)
40 - 49	96 (14.3%)	120 (17.9%)	216 (32.3%)	86 (12.8%)	113 (16.8%)	199 (29.5%)	96 (14.3%)	113 (16.8%)	209 (31.1%)
50 - 59	60 (9.0%)	46 (6.9%)	106 (15.8%)	49 (7.3%)	40 (5.9%)	89 (13.2%)	65 (9.7%)	52 (7.7%)	117 (17.4%)
over 59	26 (3.9%)	11 (1.6%)	37 (5.5%)	28 (4.2%)	13 (1.9%)	41 (6.1%)	18 (2.7%)	14 (2.1%)	32 (4.8%)
Total	284 (42.5%)	385 (57.5%)	669	265 (39.3%)	409 (60.7%)	674	269 (40.0%)	404 (60.0%)	673

ascending IC group. The survey did not support that the order effect exists for the IC responses. The hypothesis that the order effect evokes the underreporting effect is not supported by our experimental survey.

Effect of response formats

We compared the four groups of the survey 2 in Table 3 except for the reverse IC group. In comparison with the DQ, the check-all group exhibits smaller numbers of applicable items except list 2, though statistically significant differences were found in four lists. This result supports the previous findings of Rasinski et al. (1994), Thomas and Klein (2006), and Smyth et al. (2006, 2008). Table A1 in the Appendix shows the percentages of respondents who answered for the item that applied to them. The check-all group showed smaller percentages in 26 items among the total 32 items, and statistically significant differences were found in all the 16 ill-defined items among them. Although the DQ group showed smaller percentages than the check-all group in six items, none of them had significant differences. For reference, the results of the DQ group in survey 1 are also shown in the table. Statistically significant differences were found in merely three items between the two DQ groups even though the two samples are not homogeneous.

In Table 3, the IC group also yields smaller responses than the DQ group in all eight lists, and significant differences were detected in four lists. This result is similar to that of survey 1, although the absolute figures are different between the two surveys. We could say that the underreporting effect is again observed in Table 3. However, the check-all group in Table 3 shows even more smaller numbers than the IC except in list 2. The differences between the check-all and IC groups are statistically significant in the four ill-defined item lists. In summary, the DQ, the check-all and the IC responses are not comparable to each other, and the mean numbers of applicable items decrease in the order of DQ, IC and check-all.

As to the comparison of the elaborate IC with the DQ, though six lists gave smaller numbers of applicable items, merely one of them showed a statistically significant difference. In particular, the absolute differences between the elaborate IC and the DQ in lists 1, 2 and 4 are less than .01, whereas those between the ordinary IC and the DQ exceed .02. Because the change of .01 in the mean IC responses results in the difference of one percentage point in the IC estimate, the difference of more than .02 could not be of negligible size in an actual IC survey. Compared with the ordinary IC, the elaborate IC yielded larger responses in all the lists except list 3, and the statistically significant differences between the ordinary and the elaborate ICs are found in the four ill-defined item lists. We could say that the elaborate IC questions are effective in reducing the underreporting effect, and that the result of the elaborate IC is reasonably close to that of the DQ.

In the elaborate IC group, not all the respondents chose two alternatives such that the sum of the numbers of applicable and non-applicable items is equal to four. Table 4 shows the number of elaborate IC respondents through the sum of the two alternatives. The proportion of unqualified respondents whose answers do not add to four is from two to four percent. Most of the unqualified respondents selected alternatives that add to less than four. However, we have to note that the sample is identical among the eight lists. Hence, we cannot say in general that most of the unqualified respondents select such alternatives that the sum of the numbers is less than the list length.

The qualified elaborate IC shown in Table 3 is the result of the elaborate IC respondents who selected two alternatives so that the number of applicable and non-applicable items add to four. Although the differences between the elaborate IC and the qualified elaborate IC are small, the results of the qualified elaborate IC are more similar to those of the DQ than the results of the elaborate IC are. In particular, the differences from the DQ responses reduced in six lists by limiting the elaborate IC respondents to the qualified ones, and the differences between the qualified elaborate IC and DQ groups are less than .01 in four lists. We could say that the elaborate IC question accompanied by qualifying responses is an effective way to reduce the underreporting effect.

Table 5 compares the mean numbers of non-applicable items between the DQ and the reverse IC groups. Statistically significant differences were found in seven lists. In six lists among them, the reverse IC respondents selected fewer numbers of non-applicable items. The reason why not all the differences are negative would be that the reverse IC question demands a highly complicated cognitive process before the respondent answers. The items are described in affirmative sentences. Some respondents might mentally re-describe each item in a negative sentence, and check whether it applies or not. However, such a process is highly confusing. It is possible that some respondents erred in reversing the contents of the items. Although the two lists exhibit positive differences, we could say that the underreporting effect is also observed even in the case that the number of non-applicable items is requested. Moreover, the underreporting effect is not a result of an overreporting of non-applicable items. This result supports the hypothesis that the underreporting effect in ordinary IC responses is due to one-sided attention to the applicable items.

Concluding remarks

This paper investigated the causes of the underreporting effect that could disturb the IC technique. Two hypotheses were examined on the basis of experimental web surveys. Results that would support the first hypothesis – the response order effect raises the underreporting effect – were not found. In contrast, it is reasonable to say that the second hypothesis was supported by the obtained results. That is, the underreporting effect emerges because the IC question requests that the respondents pay attention merely to the applicability of each item and not to address the non-applicability. So far, many other surveys have revealed that the 'check-all-that-apply' format yields fewer responses than does the 'applies/does not apply' format, even if marked in an explicit

Table 2: Mean numbers of applicable items with standard errors in parenthesis: Survey 1 for the response order effects

List	DQ	Ascending IC	Descending IC	Ascending IC − DQ	Descending IC − DQ	Ascending IC − descending IC
1	1.486 (.041)	1.441 (.039)	1.431 (.039)	−.045	−.055	.010
2	0.369 (.026)	0.325 (.025)	0.340 (.026)	−.044	−.029	−.015
3	1.215 (.035)	1.117 (.032)	1.140 (.032)	−.098	−.076	−.022
4	2.284 (.034)	2.116 (.036)	2.152 (.039)	−.168 ‡	−.132 *	−.036
5	2.224 (.039)	2.006 (.036)	2.067 (.037)	−.218 ‡	−.157 †	−.061
6	2.286 (.039)	2.144 (.037)	2.105 (.038)	−.142 *	−.180 ‡	.038
7	2.716 (.039)	2.184 (.042)	2.155 (.041)	−.532 ‡	−.561 ‡	.029
8	2.462 (.042)	2.362 (.043)	2.299 (.042)	−.100	−.163 †	.063
n	669	674	673			

Note: Symbols *, † and ‡ represent the results of Bonferroni adjusted *t*-tests which test whether the mean numbers of applicable items are the same.
*$p < .10$ †$p < .05$ ‡$p < .01$

manner. Because the IC respondents mentally check merely the items that apply to them, the IC responses are fewer than the DQ responses. The IC respondents do not select 'applies' by comparing it to 'does not apply'. It can be said that the cognitive process of responding to the IC question is relatively different from that of answering the DQ question. This consideration is also supported by the result that the reverse IC responses are fewer than the number of non-applicable items in the DQ. However, the IC responses were more similar to the DQ responses than were the explicit 'check-all-that-apply' responses. We suppose that this is because the tacit marking of items demands a more careful examination of item lists than does the explicit marking scheme. Some careful respondents would check whether they have miscounted their IC responses before giving their answers. They might re-examine the list from the first item and find that they have missed counting some items as applicable. On the other hand, the explicit 'check-all-that-apply' question does not request the respondents to go over the item list again because the response to each item is usually recorded during the first examination on the list.

One way to evade the underreporting effect is to request the respondents to answer the numbers of both applicable and non-applicable items, which we refer to as the elaborate IC question. Some respondents may count the applicable items on their left hand and the non-applicable items on their right hand. Even if they do not use their fingers, such a process requires a cognitive process that is similar to that used in answering the DQ. The respondents of the elaborate IC questions would compare 'applies' and 'does not apply' on an equal basis. The experimental survey in this paper demonstrated that this elaborate IC question certainly decreased the magnitude of the underreporting effect. We suggest the use of the elaborate IC format for the IC technique, although it appears to be a redundant task. One may think that the ordinary IC question using well-defined non-key items is sufficient because the differences between the ordinary and elaborate IC questions were non-significant when the well-defined item lists were used. The well-defined item list certainly suppresses the underreporting effect in comparison to the ill-defined item list. Hence, it is obviously necessary to carefully prepare well-defined non-key items. However, this does not imply that the elaborate IC question is unnecessary. We suppose that the non-significant results are merely because of a lack of power and that the difference of .073 between the ordinary and elaborate IC questions shown in the list 4, for example, is not negligible in the calculation of the IC estimates. We propose the use of the elaborate IC question because the cognitive process of the elaborate IC response can be considered to be more comparable to that of the forced-choice DQ response than is the cognitive process of the traditional IC response. Although the IC technique should be used as an alternative to the forced-choice question, the ordinary IC question is regarded as an alternative to the 'check-all-that-apply' question.

The problem of whether the elaborate IC question actually yields more appropriate IC estimates for sensitive key items than the conventional IC question does should be left as a challenging topic for future research. It is a challenge because the way in which the sensitivity of item lists affects the respondents' behaviour when the elaborate IC questions are used is unpredictable. In our study, the underreporting effect seems to exist marginally, even when the elaborate method is used. It is possible that there exist other causes of the underreporting effect, which might adversely affect the IC responses. Alternatively, smart respondents might evade counting the number of non-applicable items by subtracting the number of applicable items from the list length. For the IC technique to yield good estimates consistently, more experience must be gained by performing more experiments.

Other remaining problems include the treatment of the ineligible cases in which the sum of the numbers of applicable and non-applicable items is not equal to the list length. When the survey is conducted online, it is possible to accept merely the qualified answers before proceeding to the next question. However, this constraint is impossible in other survey modes such as mail surveys. The experimental survey in this paper illustrated that the qualification of the elaborate IC respondents yields results more similar to those of the DQ respondents. It is not confirmed whether the qualification yields better IC estimates of the sensitive key items.

The IC technique could keep the respondents' true status as to the key item unknown to others more certainly when the item list becomes longer. The length of each item list used

Table 3: Mean numbers of applicable items with standard errors in parenthesis: Survey 2 for the response format effects

List	DQ	Check-all	IC	Elaborate IC	Qualified elaborate IC	Check-all − DQ	IC − DQ	Check-all − IC	Elaborate IC − DQ	Elaborate IC − IC	Qualified elaborate IC − DQ
1	1.445 (.033)	1.369 (.032)	1.422 (.027)	1.448 (.027)	1.452 (.027)	−.076	−.022	−.053	.004	.030	.008
2	0.345 (.020)	0.369 (.021)	0.318 (.017)	0.345 (.017)	0.341 (.017)	.024	−.027	.051	−.000	.023	−.004
3	1.117 (.026)	1.048 (.026)	1.093 (.021)	1.071 (.021)	1.078 (.021)	−.070	−.025	−.045	−.047	−.015	−.039
4	2.099 (.028)	2.016 (.028)	2.033 (.024)	2.107 (.023)	2.106 (.023)	−.084	−.067	−.017	.007	.073	.006
5	2.051 (.031)	1.581 (.029)	1.868 (.024)	1.990 (.024)	1.994 (.024)	−.469 ‡	−.183 ‡	−.286 ‡	−.061	.126 ‡	−.057
6	2.282 (.029)	1.885 (.029)	2.112 (.024)	2.269 (.024)	2.282 (.025)	−.398 ‡	−.171 ‡	−.227 ‡	−.014	.170 ‡	−.001
7	2.628 (.032)	1.791 (.034)	2.148 (.029)	2.328 (.027)	2.350 (.028)	−.836 ‡	−.480 ‡	−.357 ‡	−.300 ‡	.202 †	−.278 ‡
8	2.355 (.035)	1.968 (.036)	2.190 (.030)	2.319 (.030)	2.333 (.030)	−.386 ‡	−.165 ‡	−.222 ‡	−.035	.143 ‡	−.021
n	1,066	1,068	1,459	1,459	1,398-1,431						

Note: Symbols *, † and ‡ represent the results of Bonferroni adjusted *t*-tests which test whether the mean numbers of applicable items are the same. * $p < .10$ † $p < .05$ ‡ $p < .01$

Table 4: Number of respondents in the elaborate IC group shown through the sum of the numbers of applicable and non-applicable items with percentages in parenthesis

List	Less than four	Four	More than four
1	42 (2.9%)	1,413 (96.8%)	4 (0.3%)
2	33 (2.3%)	1,424 (97.6%)	2 (0.1%)
3	28 (1.9%)	1,427 (97.8%)	4 (0.3%)
4	37 (2.5%)	1,420 (97.3%)	2 (0.1%)
5	43 (2.9%)	1,409 (96.6%)	7 (0.5%)
6	49 (3.4%)	1,406 (96.4%)	4 (0.3%)
7	53 (3.6%)	1,398 (95.8%)	8 (0.5%)
8	25 (1.7%)	1,431 (98.1%)	3 (0.2%)

Table 5: Number of non-applicable items in DQ and reverse IC with standard errors in parenthesis: Survey 2 for the response format effects

List	DQ	Reverse IC	Reverse IC − DQ
1	2.555 (.033)	2.420 (.029)	−.135 †
2	3.655 (.020)	3.303 (.032)	−.352 ‡
3	2.883 (.026)	2.681 (.025)	−.201 ‡
4	1.901 (.028)	1.914 (.024)	.014
5	1.949 (.031)	1.661 (.025)	−.289 ‡
6	1.718 (.029)	1.446 (.025)	−.271 ‡
7	1.372 (.032)	1.551 (.028)	.179 ‡
8	1.645 (.035)	1.339 (.028)	−.307 ‡
n	1,066	1,459	

Note: Symbols *, † and ‡ represent the results of Bonferroni adjusted *t*-tests which test whether the mean numbers of non-applicable items are the same.
* $p < .10$ † $p < .05$ ‡ $p < .01$

in this paper was four. It is necessary to demonstrate that the elaborate IC method is not affected by the length of the item list.

Our samples in this paper are not representatives of the Japanese people. However, it has been repeatedly found in various survey conditions that 'check-all-that-apply' responses produce fewer applicable items than forced-choice responses (for example, Thomas and Klein, 2006). Hence, we believe that our findings in this paper have a certain amount of validity in other survey conditions.

References

Biemer, P., & Brown, G. (2005). Model-based estimation of drug use prevalence using item count data. *Journal of Official Statistics*, *21*, 287-308.

Biemer, P. P., & Wright, D. (2004). *Estimating cocaine use using the item count methodology: Preliminary results from the national survey on drug use and health.* Paper presented at the annual meeting of the American Association for Public Opinion Research. Phoenix, AZ.

Cruyff, M. J. L. F., van den Hout, A., van der Heijden, P. G. M., & Böckenholt, U. (2007). Log-linear randomized-response models taking self-protective response behavior into account. *Sociological Methods and Research*, *36*, 266-282.

Dalton, D. R., Daily, C. M., & Wimbush, J. C. (1997). Collecting "sensitive" data in business ethics research: A case for the unmatched count technique (UCT). *Journal of Business Ethics*, *16*, 1049-1057.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology*, *47*, 817-828.

Dillman, D. A. (2007). *Mail and Internet Surveys: The Tailored Design Method* (2nd ed.). New Jersey: John Wiley & Sons.

Dillman, D. A., Smyth, J. D., Christian, L. M., & Stern, M. J. (2003). *Multiple answer questions in self-administered surveys: The use of check-all-that-apply and forced-choice question formats.* Paper presented at the annual meeting of the American Statistical Association. San Francisco, CA.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (p. 185-210). New York: John Wiley & Sons.

Mitofsky, W. J., & Edelman, M. (1995). A review of the 1992 VRS exit polls. In P. J. Lavrakas, M. W. Traugott, & P. V. Miller (Eds.), *Presidential Polls and the News Media* (p. 81-100). Colorado: Westview Press.

Morrison, R. L., Dillman, D. A., & Christian, L. M. (2010). Questionnaire design guidelines for establishment surveys. *Journal of Official Statistics*, *26*, 43-85.

Rasinski, K., Mingay, D., & Bradburn, N. (1994). Do respondents really "mark all that apply" on self-administered questions? *Public Opinion Quarterly*, *58*, 400-408.

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Smyth, J. D., Christian, L. M., & Dillman, D. A. (2008). Does "Yes or No" on the telephone mean the same as "Check-All-That-

Apply" on the web? *Public Opinion Quarterly*, *72*, 103-113.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, *70*, 66-77.

Sniderman, P. M., & Grob, D. B. (1996). Innovations in experimental design in attitude surveys. *Annual Review of Sociology*, *22*, 377-399.

Sudman, S., & Bradburn, N. M. (1982). *Asking Questions*. San Francisco: Jossey-Bass.

Thomas, R. K., & Klein, J. D. (2006). Merely incidental?: Effects of response format on self-reported behavior. *Journal of Official Statistics*, *22*, 221-244.

Tsuchiya, T., Hirai, Y., & Ono, S. (2007a). An empirical study of item count technique based on face-to-face interviews: Some suggestions for its application. *Proceedings of the Institute of Statistical Mathematics*, *55*(1), 159-175. (in Japanese)

Tsuchiya, T., Hirai, Y., & Ono, S. (2007b). A study on the properties of the item count technique. *Public Opinion Quarterly*, *71*, 253-272.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63-69.

Wimbush, J. C., & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology*, *82*, 756-763.

Appendix

Table A1: Item lists with the percentages of applied respondents

		Survey 1	Survey 2	
		DQ	DQ	Check-all
List 1	Having more than one television set at home	70.1	69.5	66.9
	Owning a set of golf clubs	26.8	26.1	22.1
	Having a dishwasher at home	31.4	30.8	28.0
	Keeping a dog	20.3	18.1	19.9
List 2	Having flown in a helicopter	16.0	15.0	15.2
	Having swum with dolphins	5.7	4.7	3.9
	Having climbed to the top of Mt. Fuji[1)]	11.2	10.3	12.1
	Having been to Egypt	4.0	4.5	5.7
List 3	Taking coffee every morning	65.5	62.3	57.4
	Go skiing every winter	12.9 †	7.8	8.5
	Brushing teeth three times a day	27.2	24.6	22.9
	Keeping a diary	16.0	17.1	15.9
List 4	Having siblings	90.7	90.5	86.7
	Smoking cigarettes	26.6	23.4	18.4
	Working on a full-time basis	56.4 †	48.2	45.3
	Living in a detached house	54.7	47.8	51.1
List 5	Niggling over unclear things	74.1	67.7	55.7 ‡
	Making a decision without a second thought	32.4	26.1	16.9 ‡
	Having an extra bit of curiousity about everything	69.1	64.0	48.4 ‡
	Having an inability to make a decision	46.8	47.3	37.2 ‡
List 6	Being less than picky about food	71.4	72.2	65.3 †
	Forgetting something often	44.4	41.2	29.5 ‡
	Staying indoors usually on holidays	60.4	67.2	60.5 †
	Having a dream often	52.3	47.7	33.2 ‡
List 7	Would like to be born again in Japan if possible	73.2	73.1	56.7 ‡
	Prefer an aircraft rather than a ship for a round-the-world trip	57.2	53.1	34.0 ‡
	Would like to go to the future rather than the past if a time machine is realized	63.7	62.2	44.0 ‡
	Prefer summer rather than winter for entry as an olympic athlete	77.4	79.9	72.6 ‡
List 8	Having got the wrong change	86.4 †	79.9	72.6 ‡
	Having forgotten a rendezvous carelessly	30.2	29.0	20.8 ‡
	Having tripped on a snowy street	71.4	65.4	50.8 ‡
	Having left something in a public vehicle such as a train	58.1	61.2	52.6 ‡

Note: [1)] Mt. Fuji is the highest mountain in Japan.

Symbols ∗, † and ‡ represent the results of Bonferroni adjusted chi-squared tests showing that the proportion is equal to that of DQ in survey 2. ∗ $p < .10$ † $p < .05$ ‡ $p < .01$