選舉研究
第十二卷第二期（94/11），pp.149-187

Is Weighting a Routine or Something that Needs to be Justified?

Tsung-wei Liu* • Kuang-hui Chen**

Abstract

Survey research as a method of collecting sample data is supposed to produce sample statistics which can estimate the corresponding population parameters if the sampling design is appropriate. However, for reasons such as unit non-response, survey data is usually weighted by the institutes that collect the data or by researchers who analyse the data in order to correct or diminish the discrepancies between sample and population. Sample statistics based on weighted data are more representative of the population parameters than unweighted data in terms of some demographic characteristics. Therefore, to some extent, it seems legitimate to weight data and this manipulation has become a routine when dealing with survey data.

It is true that to weight data could be helpful, but this manipulation needs justifications. This paper therefore tries to argue that to weight data is no panacea and should not be taken for granted when considering the examples in Taiwan's Election and Democratization Studies (TEDS) surveys. The first section discusses why weighted data is not necessarily representative of the population. As the TEDS surveys show, the turnout, the vote shares of parties, and marital status become more deviant from the population parameters after weighting the data.

* Assistant Professor, Department of Political Science, National Chung-Cheng University.
** Ph.D. candidate, Department of Political Science, University of California, Santa Barbara.

If the focus is the relationships between variables, the correlations may be changed by weighting the data in bivariate or multivariate analysis. However, it is not clear whether we manufacture relationships which do not exist or if weighting the data actually helps us approximate the relationships that already exist in the population. Besides, it should be noted that to weight data set as a whole only deals with the problem of unit non-response, but does not solve the problem of item non-response.

The third section discusses why most efforts should be devoted to examining and improving questionnaires, sampling designs, and interviewers' training and supervision, instead of simply appealing to post-weighting. If everything necessary has been tried, weighting data may be the last resort to improve the estimates. But the justifications for the selection of auxiliary variables and the methods of calculating weight factors should be provided rather than doing it without any explicit considerations. It is also important to consider whether the consequence of weighting is positive or negative.

Keywords: weighting, unit non-response, item non-response, TEDS

I. Univariate Statistics

TEDS 2001 was conducted from January to April, 2002, after the 2001 year end Legislative Yuan election. This survey tried to access 5702 selected Taiwanese citizens over age 20, among which 35.5% were successfully interviewed, 19.6% refused to be interviewed and most of the rest were non-contacts. The sample size is 2022 and the sampling error is 2.18% (Hsu, Hung and Huang, 2002). TEDS 2002 for Taipei and Kaohsiung mayoral elections was conducted from January to April, 2003 after the elections were held at the end of 2002. TEDS 2002 successfully interviewed 1216 respondents out of 4478 selected Taipei citizens (27.2%), and 1227 respondents out of 4001 selected Kaohsiung citizens (30.7%). The maximum sampling error is 2.8% (Huang, 2003). TEDS 2003 was conducted from August to September, 2003. This survey tried to access 3893 selected Taiwanese citizens over age 20, among which 29.9% were successfully interviewed and 25.1% refused the interviews. The sample size is 1164 (Chu, 2004). Therefore, TEDS 2001, 2002 and 2003 are supposed to provide acceptable statistics to estimate population parameters. However, the estimation based on the respondents could be biased unless the characteristics and attitudes between respondents and non-respondents are identical. Whether this is the case can only be examined by (a) comparing statistics obtained from respondents with that from non-respondents[1] or by (b) comparing respondents' statistics with population parameters. But neither of these is possible. Non-respondents are not interviewed, so we know little about them. In most cases, population parameters are not available, which is why surveys are conducted to estimate them.

There are four variables in TEDS surveys whose population parameters are available, including gender, age, education and area strata, against which the sample statistics can be examined preliminarily to see whether the samples are representative of the population.[2] In TEDS 2001 and 2002, gender and area strata pass goodness-of-fit tests at the 0.05 level of significance while age and education do not (Hsu, Hung and Huang, 2002). In TEDS 2003 gender and age pass the tests and education and area strata do not.

In order to make the samples more representative of the population, TEDS surveys use gender, age, education, and area strata to produce the final weight factors by the raking procedure, and the final weights are used to correct the discrepancies between the sample and the popula-

tion. After the data set is weighted by the final weight factors, the sample and the population are consistent in terms of the four demographic variables.

However, the question we have to conjecture further is the extent to which the post-weighting procedure has contributed to the improvement of descriptive inferences of other variables in TEDS surveys.

If the relationships among the four variables and each of the rest of other variables in TEDS surveys are uniform, the post-weighting procedure using the four demographic variables to adjust the sample data will make all the sample statistics approximate to the population parameters. That is, if the four demographic variables are representative of other variables, using the four demographic variables to adjust the sample data set produces an equal effect upon all the variables within these data sets. Thus, the correspondence of demographic variables between the sample and population manufactured by weighting represents the same correspondence of other variables between sample and population. However, according to our knowledge with regard to the relationships between the four demographic variables and other variables, the assumption is not true. The four demographic variables are either unrelated or related at varying degrees with other variables. Thus, the post-weighting procedure is not guaranteed to produce a perfect congruence between sample statistics and population parameters.

In order to verify our speculation about the likely influence of post-weighting in terms of descriptive inference, we select two political variables (turnout and vote shares) and one non-political variable (marital status) to examine the effects of post-weighing. The reason we focus on these three variables is that their population parameters are available, in addition to the four demographic variables that have been used to weight the data.

1. Turnout

TEDS 2001 unweighted data shows that 81.3% of respondents voted in the 2001 Legislative Yuan election and 17.3% did not. But the actual turnout is only 66.2%. The turnout was overreported by 15.1%. If sample statistics can be improved as close as possible to population parameters by weighting data, the discrepancy between the estimate based on weighted data and the actual turnout will be noticeably smaller. However, the turnout was overreported by 14.5% even though the weighted data is used. This gap is much larger than the 3% of sampling error no matter whether data is weighted or not.

TEDS 2002 witnesses the same pattern. The unweighted data shows that 87.2% of respon-

dents in Taipei and 88.8% in Kaohsiung voted in the 2002 mayoral elections. But the actual turnouts are only 70.6% and 71.4% in Taipei and Kaohsiung respectively. The discrepancies are 16.6% and 17.4% for Taipei and Kaohsiung even though the data is weighted. The problem of overreported turnout is not ameliorated by weighting the data.[3]

It is true that respondents tend to overreport voting. Respondents may claim that they did vote but in fact they did not for reasons of social desirability or memory failure (Presser, 1990; Abelson, Loftus and Greenwald, 1992). It is also possible that the gap is caused by the fact that the survey is based on the voting-age population while the official turnout is based on the voting-eligible population (Martinez, 2003). Those explanations make sense and are supported by the empirical evidence. The implication is that the estimation of population parameters cannot be improved by simply weighting data, at least in this case. Weighting data can do nothing about the discrepancy caused by misleading information provided by the respondents, because it cannot pick out what people are lying and correct them. The discrepancy caused by the different definitions of the population could be corrected by weighting data. But when the gap is greater than 15%, weighting data based on the voting-eligible population offers little help.

One of the suggestions for minimizing the discrepancy is to improve the question wording. A better questionnaire design can make respondents more willing to reveal whether they voted and help them to recall how they voted (Belli et al., 1999). In other words, what is needed in the first place is to examine whether the questions asked are capable of measuring what are intended to be measured.

The estimation of turnout rates is a strict test for surveys, because it is one of the few estimates that can be examined against the population parameters. The discussion of turnout is not only meaningful in its own right, but also reminds us of the fact that the same problem could happen to other variables even though the "correct" answers are not available against which we can see how far/close the estimates are to their corresponding population parameters.

2. Vote Shares and Party Support

In the 1992 British general election, 54 national voting intention polls with more than 1000 samples in average were carried out within one month of the polling day. All polling companies predicted that the Labour would win. But the Conservatives were 7.6% ahead when the ballot boxes were opened. No special event suddenly appeared in the last stage of that campaign and the voting intentions were quite stable over the campaign, so "late swing" cannot explain why

polls went wrong (Jowell et al., 1993).

Some of the inquiries into the 1992 polling disaster point out that it is the sampling design that caused the bias that favoured the Labour (Curtice and Sparrow, 1997; Jowell et al., 1993). In the U.K., the traditional face-to-fact quota sampling leaves plenty of flexibility for interviewers to select their respondents as long as they can meet the quota based on a set of characteristics such as gender, age, social class, work status, and housing tenure. When interviewers can choose respondents, it is no surprise that they tend to interview the persons who agree easily to be interviewed and refrain from spending more time on persuading difficult persons. Therefore, the bias will occur no matter how perfect the quotas are controlled and how many additional auxiliary variables are used to weight data.

It could be argued that voting intentions are changeable, so they are not always equal to actual vote shares. However, the polls conducted by MORI, Gallup and NOP after the 1992 general election in which respondents were asked how they voted "found" that the Conservatives should lose the 1992 election by a margin of 5% to 7%. Even though the data was weighted to "estimate" the 1992 election result, the correct answer still did not come up (Curtice, 1997). When the bias was caused by the fact that the Conservative supporters either did not reveal their preferences or refused to be interviewed in both pre- and after-election polls, weighting the data seems to be an irrelevant solution. What count are how to sample and who are interviewed, not how to weight the data.[4]

TEDS 2001 was conducted after the 2001 Legislative Yuan election, so it should not be bothered by the problem of changeable voting intentions. How respondents had voted was asked and the distribution is shown in Table 1.[5] The DPP loses 2.5% of votes in the survey, and it loses 4.3% when the data is weighted. The KMT gains 0.1% of votes in the survey, and it loses 1.2% when the data is weighted. As the last two columns of Table 1 show, the differences between the actual vote shares and the distribution obtained by the survey become larger after the data is weighted. Weighting the data seems to make the problem worse rather than to correct the bias.

Table 1 Vote Shares of Parties in the 2001 Legislative Yuan Election

	unweighted data		weighted data		vote shares	difference 1	difference 2
	frequency	percentage	frequency	percentage	percentage	percentage	percentage
KMT	376	28.7	353	27.4	28.6	0.1	1.2
DPP	470	35.9	486	37.7	33.4	2.5	4.3
NP	29	2.2	27	2.1	2.6	0.4	0.5
PFP	222	16.9	211	16.4	18.6	1.7	2.2
TSU	78	5.9	76	5.9	7.8	1.9	1.9
Independents and Small Parties	136	10.4	136	10.5	9.0	1.4	1.5
Total	1311	100.0	1290	100.0	100.0	7.8	11.6

Note: difference 1= absolute value of 'vote shares' minus 'unweighted', difference 2= absolute valute of 'vote shares' minus 'weighted'

Respondents' voting behaviours were asked in TEDS 2002 too. As Table 2 shows, the gap between the result of Taipei mayoral election and survey result is 6.2% for both the KMT and DPP. The discrepancies are improved to 3.9% by weighting the data. The gaps between the result of Kaohsiung mayoral election and survey result are 8.1% for the KMT and 7.9% for the DPP. The discrepancies become worse reaching 9.7% for the KMT and 9.5% for the DPP after weighting the data. The estimates of vote shares are improved by weighting data in the case of Taipei while they become worse in the case of Kaohsiung.

Table 2 Vote Shares of the 2002 Taipei and Kaohsiung Mayoral Elections

Taipei

	unweighted data		weighted data		vote shares	difference 1	difference 2
	frequency	percentage	frequency	percentage	percentage	percentage	percentage
KMT	655	70.4	617	68.0	64.1	6.2	3.9
DPP	276	29.6	291	32.0	35.9	6.2	3.9
Total	931	100.0	908	100.0	100.0	12.5	7.7

Note: difference 1= absolute value of 'vote shares' minus 'unweighted', difference 2= absolute valute of 'vote shares' minus 'weighted'

Kaohsiung

	unweighted data		weighted data		vote shares	difference 1	difference 2
	frequency	percentage	frequency	percentage	percentage	percentage	percentage
KMT	344	38.7	329	37.1	46.8	8.1	9.7
DPP	515	57.9	528	59.5	50.0	7.9	9.5
Independent	30	3.4	30	3.4	3.1	0.2	0.2
Total	889	100.0	887	100.0	100.0	16.2	19.4

Note: difference 1= absolute value of 'vote shares' minus 'unweighted', difference 2= absolute valute of 'vote shares' minus 'weighted'

The British experience is not only helpful when thinking about how come we cannot get the vote shares right, but also suggests that weighting the data is not a solution. TEDS surveys did not use quota sampling, so interviewers did not select whom they would like to interview. However, how samples were replaced when the first set of samples could not be contacted or refused to be interviewed could be a problem. In addition to the first set of samples, there were several sets of replacement samples. If the first set of samples cannot be interviewed, interviewers can find replacements from other sets of samples. It is true that interviewers' choices were limited. But it is also true that it could be one of the sources for bias if interviewers did not try hard enough to contact or persuade the people from the first set of samples.

It is also possible that the discrepancy is caused by the fact that it is sensitive to ask people whom they vote for in Taiwan, so Taiwanese tend to either hesitate to answer the question or make up an answer. TEDS surveys show that the losing parties receive less votes in surveys than in elections, so it is plausible to take seriously the effect of bandwagon when looking for explanations for the discrepancy. No matter what the reasons for the discrepancy are, how samples are replaced, whether it is a sensitive question, or the possibility of bandwagon, what they have in common is that these problems cannot be solved by weighting the collected data.

If we cannot even get the vote shares right in the survey irrespective of whether the data is weighted or not, it is doubtful that we can be confident of the measurements of other related variables. For example, TEDS 2001 shows that the percentages of supporters for the KMT, DPP, NP, PFP, and TSU were 14.6%, 29.9%, 0.4%, 13.2% and 1% respectively (unweighted data). When the discrepancy between the actual vote shares and vote shares obtained in surveys is noticeable, it is difficult to take the distribution of party support seriously, even only as a snapshot.

3. Marital Status

The variables discussed above, turn out and vote shares, are political variables, so we next examine a non-political variable, marital status, to see the effect of weighting data. The population presented in the first part of Table 3 consists of people who were age twenty or above in the year of 2001, which is in agreement with the definition of population in TEDS 2001. It is evident that when the distribution of the unweighted sample data is compared with that of the population, significant discrepancies between sample and population appear. Whereas the percentage of married, compared with that of the population, is overestimated in the unweighted sample data set, the percentages of single, divorced, and widowed are underestimated, especially in the last two categories. We go on and consider the weighted sample statistics. The change of sample statistics before and after post-weighting is barely noticeable. The statistics of Taipei and Kaohsiung in 2002 and that of Taiwan in 2003 presented in Table 3 also have similar patterns.

Table 3 Distributions of Marital Status: Population, Unweighted, and Weighted Survey Data

	Single	Married	Divorced	Widowed	Nonresponse	Total
2001 Taiwan						
Popultaion	4,270,868	9,946,330	802,475	955,664		15,975,337
%	26.7	62.3	5.0	6.0		100.0
Sample						
Unweighted	496	1470	20	32	4	2022
%	24.5	72.7	1.0	1.6	.2	100.0
Weighted	506	1463	19	32	4	2022
%	25.0	72.3	.9	1.6	.2	100.0
2002 Taipei						
Popultaion	577,432	1,172,815	114,787	102,789		1,967,823
%	29.3	59.6	5.8	5.2		99.9
Sample						
Unweighted	324	863	15	11	3	1,216
%	26.6	71.0	1.2	.9	.2	100.0
Weighted	305	874	16	17	3	1215
%	25.1	71.9	1.3	1.4	.2	99.9
2002 Kaohsiung						
Popultaion	324,925	648,023	72,367	59,606		1,104,921
%	29.4	58.6	6.5	5.4		99.9
Sample						
Unweighted	315	862	16	32	2	1227
%	25.7	70.3	1.3	2.6	.2	99.9
Weighted	326	850	16	33	2	1227
%	26.6	69.3	1.3	2.7	.2	100.1
2003 Taiwan						
Popultaion	4,515,178	10,029,342	928,539	1,003,823		16,476,882
%	27.4	60.9	5.6	6.1		100.0
Sample						
Unweighted	255	864	18	25	2	1164
%	21.9	74.2	1.5	2.1	.2	99.9
Weighted	244	873	18	27	2	1164
%	20.9	75.0	1.6	2.3	.2	100.0

Note: Population statistics are cited from the Ministry of Interior website (http://www.moi.gov.tw/W3/stat/home.asp).

Undoubtedly, we should be cautious about drawing a conclusion from this comparison, because the issue of marital status, to some extent, is a sensitive topic, especially for those who are divorced. It seems reasonable that the percentage of divorced people is always underestimated as a result of systematic bias, because people who are divorced, more or less, would not like to admit their divorce status in front of strangers. However, we notice that the percentage of widow does not change after the post-weighting. Because becoming a widow or widower is not as sensitive as divorce, the percentage of widowed in the sample is supposed to approximate that of the population, if the post-weighting does function as it is expected. However, it does not.

4. An Estimate or Estimator?

The discrepancy between sample statistics and population parameters is not necessarily labelled as survey errors. It depends on whether a survey is regarded as a single case or one of hypothetical replications. An estimate obtained from a sample may be biased, but the estimator could be unbiased, i.e. the mean of sample statistics is equal to population parameters (Groves, 1989). Therefore, if what in question is the estimator and the sampling design is appropriate, post-weighting is not needed even though the discrepancy is observed.

If an accurate estimate is what we are looking for, post-weighting may be one of the methods to correct the biased estimate. However, according to what we have observed from TEDS surveys, the post-weighting procedure does not narrow down the gap between sample and population satisfactorily. In some occasions, weighting the data make the estimates even worse.

To weight data is justifiable in some occasions. But it is no panacea, because it is an irrelevant solution to some biases. When the auxiliary variables do not correlate to other variables, weighting data makes no difference. When the variables that the researchers are interested in are sensitive or difficult such as the examples of voting behaviours and marital status discussed above, devoting the effort to designing a better questionnaire is something that needs to take place first, not weighting the data as a short cut. The U.K. example demonstrates that weighting the data is no remedy for problematic sampling design. A biased estimate from a problematic sample cannot be corrected by weighting, so it is much more important to examine sampling design than to appeal to weighting the data readily.

How to define and correct survey errors not only depends on what in question are estimates or estimators, but also depends on whether the goal of surveys is to describe specific population or to develop models (Groves, 1989). The discussion so far have argued that post-weighting is

not always legitimate in univariate description. What if survey data is used to multivariate analysis or model building? We turn to this issue in the section to follow.

II. Relationships between Variables

The goal of post-weighting is to minimize the discrepancies between sample and population in order to make the sample more representative of the population. By doing so, it is expected to draw "better" inferences from the sample data in terms of both univariate as well as multivariate analysis. In this section, we examine the influence of post-weighting on bivariate analysis, that is, the examination of the relationships between variables. The questions we focus on are: Does post-weighting alter the relationships between variables? If so, how strong is the influence? Why are the relationships altered? In addition to the issue of post-weighting for unit non-response, we also take the issue of item non-response into account, which makes the sample data incomplete. Dealing with these questions elicits our concern with the applicability of post-weighting to multivariate analysis.

1. Multivariate Analysis

In multivariate analysis, the implementation of post-weighting means that respondents of different strata have varying degrees of contributions to the relationships between variables according to their corresponding weights. If the auxiliary variables selected to produce the weight factors do not relate to the variables in question, there should exist no systematic influence of post-weighting on the multivariate analysis. If these weight factors are associated with these variables under examination, the multivariate analysis should be affected by post-weighting. Four demographic variables, gender, education, age, and area strata, are used to produce the weight factors in TEDS surveys. Because the four demographic variables are assumed to be related to a wide array of variables regarding individuals' opinions, attitudes, and behaviours, the implementation of post-weighting would make a difference regarding the relationships between variables in TEDS data.

For convenience of analysis, among the four variables employed to produce the weights, we focus only on education to illustrate the likely impacts resulting from post-weighting upon the relationships between variables. As it is listed on Table 4, on average, either respondents with lower levels of education are assigned higher weight factors in order to compensate for their

under-representation in the sample relative to the distribution in the population, or those with higher education are given lower weight factors to counteract their over-representation in the sample. That is to say, respondents with lower levels of education will be counted more than those with higher education as a result of the implementation of post-weighting. Thus, compared with the unweighted data, weighted data discloses more characteristics of those with less education.

Table 4 Average Weights and Item Nonresponse across Levels of Education

(TEDS 2001)

Education Level	Mean of Weight	Number of Nonresponse	N
Elementary School or Below	1.22	2.85	549
Junior High School	1.26	1.48	265
High School	.94	1.12	582
Young College	.86	.86	279
University	.66	.80	340
Non-response	.97	2.71	7
Total	1.00	1.55	2022

Number of nonresponse refers to the mean of item nonreponse out of the 8 questions regarding the relationship between Taiwan and China (see Appendix A).

(TEDS 2002 Taipei)

Education Level	Mean of Weight	Number of Nonresponse	N
Elementary School or Below	1.91	1.90	137
Junior High School	1.74	.66	102
High School	1.02	.42	326
Young College	.80	.43	217
University	.62	.42	430
Non-response	1.06	.25	4
Total	1.00	.61	1216

Number of nonresponse refers to the mean of item nonreponse out of the 6 questions regarding the relationship between Taiwan and China (see Appendix A).

(TEDS 2002 Kaohsiung)

Education Level	Mean of Weight	Number of Nonresponse	N
Elementary School or Below	1.07	2.01	265
Junior High School	1.11	.60	149
High School	1.08	.70	376
Young College	1.02	.55	181
University	.73	.48	248
Non-response	1.01	2.75	4
Total	1.00	.91	1223

Number of nonresponse refers to the mean of item nonreponse out of the 6 questions regarding the relationship between Taiwan and China (see Appendix A).

(TEDS 2003)

Education Level	Mean of Weight	Number of Nonresponse	N
Elementary School or Below	1.00	1.57	300
Junior High School	1.08	.79	162
High School	1.02	.49	342
Young College	1.06	.40	149
University	.86	.43	205
Non-response	.98	2.50	6
Total	1.00	.80	1164

Number of nonresponse refers to the mean of item nonreponse out of the 6 questions regarding the relationship between Taiwan and China (see Appendix A).

The existing literature has demonstrated the relationship between education and opionionation (Converse, 1976; Reese and Miller, 1981; Neuman, 1986). Specifically, well-educated people are more cognitively sophisticated, and hence, are better capable of collecting information and putting it into use than those with less education in survey responses (Sniderman et al., 1991). Poor-educated respondents should be less likely to propose valid responses to survey questions than their well-educated counterparts. Hence, poor-educated respondents are more likely to be non-respondents than well-educated ones. In addition, poor-educated respondents are supposed to have less coherent opinions and attitudes than well-educated ones because the levels of cognitive sophistication and information accessibility between the two groups are different. Thus, weighted data should have higher rates of item non-response and lower levels of

association between variables which are theoretically supposed to be correlated. Consequently, we hypothesize that the relationships between variables will become weaker once the data is weighted. There are a series of questions with regard to the relationship between Taiwan and China in TEDS 2001 data (see Appendix A for the eight questions). These questions and the correlation coefficients between each of them are used to examine the effects of post-weighting upon bivariate analysis.

Firstly, we compare the means of non-response rates (including don't know, no opinion, refuse to answer, depends, etc.) in the eight questions across five levels of education. As we mentioned above, Table 4 shows that education and non-response are negatively correlated. That is, the lower the level of education is, the higher the rate of non-response is.

Furthermore, we examine whether the implementation of weighting would attenuate the correlation coefficients as a result of the greater weights assigned to the respondents with lower levels of education. The same eight questions produce twenty-eight correlation coefficients in TEDS 2001 (see Appendix B for the details of these Spearman correlation coefficients). We compare the absolute values of the correlation coefficients before and after weighting. Although the differences between weighted and unweighted correlation coefficients are slim, a pattern emerges: among the twenty eight absolute values of correlation coefficients, twenty six of them decrease after the implementation of weighting, while the other two remain unchanged. In addition, all the chi-square values between pairs of these variables decrease after weighting (see Appendix C). We also find similar patterns regarding changed correlation coefficients in TEDS 2002 as well as TEDS 2003.[6] It is evident, as we hypothesized, that the post-weighting does attenuate the relationships between these variables due to the greater weights assigned to those with lower levels of education.

However, the slight difference between weighted and unweighted correlation coefficients may in part reflect the possible influence resulting from item non-response. Specifically, because respondents who did not offer a valid answer to the survey question are excluded from analysis, these drop-outs also produce some impacts on correlation coefficients between these variables. Respondents who say don't know or no opinion, refuse to answer, or offer ambiguous answers are usually treated as item non-response and are dropped out as missing value, and hence, are excluded from the multivariate analysis. Given that those drop-outs are more likely to be those with less education, it appears that, compared to the whole sample data, respondents who are included in the multivariate analysis, on average, are better-educated. Thus, weighting

and drop-out are two opposite forces affecting the relationships between variables. Whereas weighting increases the contribution of those with less education, the item non-response decreases their contribution. According to the comparison of weighted and unweighted correlation coefficients, it seems that because the weights of those with less education are enlarged, the relationship between variables attenuates. That is to say, once the factors which are used to produce the weights are associated with the variables under examination, the implementation of post-weighting does produce a certain amount of systematic influences upon relationships between variables. But it is still uncertain whether the effect of post-weighting makes the correlation coefficients closer to the true relationships or farther away from them.

2. Item non-response

The existence of item non-response calls our attention to the application of post-weighting to the incomplete data set. As Appendix B shows, among the twenty- eight correlation coefficients, the numbers of respondents included range between 1385 (68.5%) and 1570 (77.6%) out of 2022, the total number of respondents in the TEDS 2001 sample. Varying segments of the whole data set were included in each of the correlation analysis. Given that listwise deletion is the default for most statistical packages, a greater proportion of respondents will be dropped out whenever a larger number of variables are taken into consideration. Namely, the larger the number of variables involved, the smaller the proportion of the original data set is analysed.

The issue of item non-response is worth much more attention than it has been given, especially when weighted data is analysed. Specifically, dropping cases out of analysis means that researchers are analysing only a fraction rather than the complete set of the sample. When different variables are included into analysis, varying segments of the original data set are examined. It should be noted that the procedure of post-weighting is conducted in the way that the sample data is treated as a whole in order to make the sample data representative of the whole population. Thus, when we are analysing a sub-group of the sample data, the weight factors based on the whole sample data are no longer applicable to a sub-group analysis, unless the varying dropouts and the remainders are identical in terms of various variables of interest. However, according to our knowledge about survey response, this is hardly the case. In other words, weighting a sub-group by the factors derived from the whole sample data may lead to a distortion rather than a correction.

Post-weighting is a method designed to correct the discrepancies between sample and

population resulting from unit non-response. In this section, we find that post-weighting does influence the relationships between variables if the auxiliary variables that produce weight factors for each stratum of respondents are associated to other variables. We also find that the effects of post-weighting are affected by the existence of item non-response, which usually is not taken into account when producing the weight factors. In the case we examined, post-weighting attenuates the correlation between variables. It is true that we cannot tell whether the procedure of post-weighting has improved the data in terms of the relationships between variables, because we do not know the "true" relationships in the population. However, the analysis above indicates that we should be cautious when using weighted data no matter if we are analysing the whole data set or a fraction of the data set, unless unit and item non-response are random.

III. When Is Appropriate to Weight the Data?

In sections I and II, it is argued that it could be problematic to weight data in both univariate and multivariate analysis. Therefore, it may be better to either use unweighted data or weight data with justifications. It is even better to think about if there is any possibility that sample data can be representative of the population without weighting the data. We now turn to discuss this below.

Weighting data by post-stratification and raking is easy and low in cost, but very effective. They are therefore worth doing. However, they are controversial in terms of ethics. Researchers could be spoiled by such a remedy rather than devote most of their efforts in pursuit of data quality (Hung, 2000). A set of sample data that is valid and representative of the population comes from a painstaking process of data collection rather than the low cost weighting procedure. It costs a lot to spend time on discussing questionnaires, in considering how to sample, and in training and supervising interviewers. But they are where most efforts should be devoted. Weight the data could be done at the final stage. But this is the last resort that needs to be justified rather than a mindless routine or lazy solution.

1. Questionnaire

Survey textbooks usually suggest that the questionnaire is crucial to whether we can collect the information we intend to receive, so every question should be well designed. It is not appropriate to think about weighting the data, if in fact the questionnaire can be improved. The TEDS

questionnaires are designed by groups of specialists on elections. But when a variety of topics are included, some questions are well designed while the others are not.[7] On one hand, the TEDS committee have to spend much more time in crafting the questionnaire. On the other hand, the researchers not involved in the execution of the surveys also have to examine the questionnaire and propose their suggestions in their research papers. The TEDS committee can then consider whether the original questions need revision. Every question deserves a deep discussion and serious examination. For example, whether the question about party preferences in TEDS surveys is valid and how to improve it needs a piece of research to discuss the definition of the concept, how to operationalise it, and wording of the question. The relevant discussions in research papers will be invaluable to the whole academic community.

It is true that one of the limitations in conducting secondary analysis is that the researchers cannot design the questions for the concepts they are interested in. However, more examination and criticism of the TEDS questionnaire design should be encouraged, without which the validity of TEDS survey cannot be improved in the future. Every researcher can access the TEDS data once they are deposited to the Center for Survey Research, Academia Sinica, where they become public goods. It is good to see more and more researchers with a wide variety of backgrounds analyze the TEDS data, but it is also their obligation to help improve the quality of the public goods.[8] Both the TEDS committee and the researchers who consume the data are responsible for the questionnaire design.

2. Sampling Design

The sampling design of TEDS surveys begins with stratifying cities and townships by conducting factor and cluster analysis. Stratification sampling is the first stage, and towns, Li and respondents are sampled in the following stages based on the principle of probability proportional to size.[9] All prospective respondents were identified by the TEDS committee, so the interviewers were not allowed to interview anyone who were not on the list. If all people sampled were interviewed successfully, the sample statistics thus obtained would be reasonably representative of the population. However, only one third of the original samples were successfully interviewed, so the replacement samples were needed.

The method of sample substitution of TEDS 2001 is to replace the original sample by a similar sample in terms of demographic characteristics, which violates the principle of probability sampling. Twenty sets of samples were provided in TEDS 2002 and 2003, respectively. Sam-

ples were conduced set by set. Before all the prospectus respondents on the list of the first sample set had being attempted to contact and interview, either successful or unsuccessful, the next set of sample was not allowed to be interviewed. For example, if district A is selected and 20 samples are assigned to this district, the first set of samples identifies 20 prospective respondents. Unless all the 20 person are successfully interviewed, the second set of samples is needed. If the goal of 20 successful interviews is not achieved in the second set of samples, the third set of samples will be interviewed. This replacement rule will be continuously carried out until the cumulated number of successful interviews reaches 20 in this district. Because interviewers cannot stop contacting perspective respondents until the whole set has been tried, it is possible that more respondents are interviewed than the number of samples assigned. For example, if 15 persons of the first set and 10 persons of the second set of samples are interviewed, then the total, 25, exceeds 20, the number of samples assigned to the district A.

Even though the TEDS surveys restricted the interviewers' flexibility regarding selecting respondents, the surveys were in favour of the people who were easier to contact and more willing to be interviewed. The interviewers did not have to try hard to find and persuade the original samples, because plenty of substitutes were ready for them. The flexibility of choosing respondents is limited, but is one of the resources of bias.

The other problem caused by the sample substitution policy is that the distribution could be distorted. TEDS 2002 intended to finish 1212 interviews, and ended up with 1227 (15 extra samples). TEDS 2003 intended to finish 1112 interviews, and finished with 1164 (52 extra samples). The extra samples, especially in TEDS 2003, is one of the reasons for the discrepancy between the sample and population.

If the characteristics and attitudes of respondents and non-respondents are identical, then how many original samples are interviewed and how to replace samples are not the issues.[10] If the differences do exist and the method of sample substitution could distort the distribution, then what is needed after designing a good questionnaire is to reconsider how to sample and how to replace samples. The suggestion to do oversampling at the beginning based on the estimated rate of successful interviews should be considered seriously (Hung, 2003). This sample design can improve the problem that interviewers may select respondents and decrease the distortion caused by extra samples. Without examining and improving the sampling design, it is not appropriate to think about weighting the data.

3. Interviewers Training and Supervision

All TEDS interviewers are required to take a one-day training course before they are dispatched to their districts. The system of supervision is well established. The district supervisors are responsible to the committee of execution, and will instruct the interviewers and check whether there is any cheating. Even when interviewers were trained and supervised, it did happen that interviewers cheated or made mistakes in TEDS surveys. Some were detected before the end of the survey and some were not detected until the data was released to the public.

The bias caused by the interviewers cannot be corrected by weighting the data, because weight factors cannot make the invalid cases become valid. The one day training courses are helpful, but could be improved (Liu, 1996). For example, it is worth considering whether more courses such as interview practice are needed. It is also worth considering whether to have senior interviewers go with inexperienced interviewers at the early stage of the survey.

The system of supervision in TEDS worked well. The district supervisors did their best to preserve the quality of data. However, when time was limited and the workload was extremely heavy,[11] it's quite possible that some interviewers' cheating and mistakes could not be detected as early as possible.

The cost of training and supervising interviewers is high, and it costs even more when some interviewers need to be substituted. Compared with weighting data, the cost is extremely high and the payoff of the effort is hardly "observable" or "measurable". However, if there is any possibility that the training courses can be improved and the number of district supervisors can be increased to prevent the cheating and mistakes from happening, then it is inappropriate to focus on weighting data.

4. To Weight Data

It is no surprise that even though the questionnaire and sampling are well designed and the interviewers are trained and supervised by the highest standard, the sample statistics are still not equal to the population parameters. The goodness-of-fit test is usually conducted after the data is cleaned in order to see how much confidence we can have in the estimates based on the sample. Because gender, age, education and area strata are the variables whose population parameters are available, they are not only used in goodness-of-fit tests, but also in post-weighting.

It is plausible to argue that if we cannot get the four demographic variables right, it seems

not convincing to claim that other estimates are close to population parameters. TEDS data was therefore weighted by the weight factors derived from the auxiliary variables with the method of raking.[12]

However, the reasons for why those four demographic variables are selected are not provided. There is no justification for it if the only reason is because these four variables' population parameters are readily available. If some auxiliary variables are dropped or added, the weights obtained may change in various extents. In other words, the statistics depend on where we stop the procedure of raking. For example, the final weights for TEDS 2001 are derived by twelve iterations until all of the four demographic variables pass the goodness-of-fit tests, so there are twelve sets of weight factors. As Table 5 shows, how many minutes the respondents spent in reading election news everyday depends on which weight factors are used. 28.13% of the respondents spent less than 30 minutes in reading election news if the data is not weighted and decreases to 26% if the data is weighted by the final set of weight factors. The percentages fluctuate during the process of raking. The category of "'less than 30 minutes" is not an exception, the fluctuation is observed in every category. Therefore, unless it is justified explicitly why the four auxiliary variables are all we need in raking, it is reasonable to doubt whether the weight factors thus obtained are helpful or not.

Table 5 How much time was spent in reading election news in the 2001 Legislative Yuan election

	w0	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	max	min	range
less than 30	28.13	28.13	28.65	25.87	26.16	26.15	27.01	25.94	25.99	25.99	26.45	25.98	26.00	28.65	25.87	2.78
31-60	10.42	10.42	10.54	9.56	9.71	9.75	9.94	9.53	9.54	9.55	9.67	9.49	9.49	10.54	9.49	1.05
61-90	3.82	3.82	3.81	3.41	3.52	3.53	3.53	3.35	3.36	3.37	3.37	3.29	3.30	3.82	3.29	0.53
91-120	1.34	1.34	1.33	1.18	1.19	1.19	1.19	1.13	1.13	1.13	1.13	1.11	1.11	1.34	1.11	0.23
more than	2.53	2.53	2.49	2.38	2.42	2.44	2.39	2.33	2.34	2.35	2.33	2.30	2.31	2.53	2.30	0.23
seldom	9.98	9.98	10.05	10.41	10.50	10.49	10.59	10.57	10.57	10.57	10.64	10.63	10.63	10.64	9.98	0.66
never	41.01	41.01	40.36	44.34	43.65	43.59	42.50	44.28	44.17	44.15	43.51	44.31	44.28	44.34	40.36	3.98
revusal	0.20	0.20	0.19	0.21	0.20	0.20	0.18	0.19	0.19	0.19	0.18	0.18	0.18	0.21	0.18	0.03
depends	2.27	2.27	2.29	2.34	2.36	2.36	2.39	2.39	2.40	2.40	2.41	2.41	2.41	2.41	2.27	0.14
don't know	0.30	0.30	0.29	0.31	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.31	0.29	0.02
total	100	100	100	100	100	100	100	100	100	100	100	100	100			

If the weight factors obtained by raking are identical, whatever auxiliary information are

considered, it is fine to weight the data without any justification. However, how many variables and which variables are used in raking have various impacts on the sample statistics.

In addition, the procedure of raking reminds us of the potential problems of weighting. The raking starts with one auxiliary variable whose population parameter is available. The second auxiliary variable is used to calculate the weights by which all cases are weighted. The second auxiliary variable must pass the goodness-of-fit test while it is not necessarily the case for the first auxiliary variable. The first auxiliary variable is distorted by the second auxiliary variable, so it must be corrected until both variables pass the tests. When the third auxiliary variable is included, it is possible that the previous variables are distorted again, so they must be corrected until all three variables pass the test of goodness-of-fit. The procedure goes on and on until all auxiliary variables are used to calculate the weight factors and all auxiliary variables pass the test of goodness-of-fit.

The iterations of the adjustment process means that it is not just a suspicion that some sample statistics could be distorted when the data is weighted, but it is exactly what is expected to occur when the weight factors are calculated by raking. It is possible that the estimates of the target variables are representative of the population parameters without weighting, but are distorted after weighting. For example, when raking is in progress, gender is representative of the population at the beginning, but its statistics are distorted after education is used to weight the data. If gender is distorted, there is no point to assume or pretend that other target variables such as respondents' attitudes and opinions are immune from any distortion. A certain variable such as gender can be corrected, but nothing can be done about the unobservable distortion of other target variables. Correcting what is known does not necessarily mean that what is unknown is also corrected. We are still uncertain of the influence of post-weighting upon a wide range of variables.

Ⅳ. Conclusions

Weighting survey data by raking with demographic variables is widely used by political scientists and institutes in Taiwan. However, whether post-weighting is a relevant issue depends on if the goal is to describe specific population or to build models. It also depends on whether what is judged is an estimate or estimator. Therefore, not every consumer of survey data has to

weight data. On one hand, weighting data could be an advanced manipulation to improve estimates of population parameters if it is not a mindless routine. On the other hand, it could be one of the sources of bias and uncertainty if data is weighted without thorough consideration and convincing justification. In this paper, we examined the effects of post-weighting in terms of both univariate and multivariate analysis. The limited effects resulting from post-weighting in univariate analysis indicates that weighting can do little to improve the estimation of parameters. It is disclosed that post-weighting does produce a certain degree of influence upon multivariate analysis. The relationships between variables are altered due to the implementation of post-weighting. It should be noted that when post-weighting has noticeable effect on either univariate or multivariate analysis, whether the effect is in the right direction is uncertain, i.e. it may make estimates worse rather than better. We also pointed out that sub-group analysis could be problematic as long as the weight factors are produced based on the whole data set. If unit non-response is a justifiable reason for weighting data, then it follows that the problem of item non-response should be considered when thinking about whether and how to weight data.

It is likely that weighting can improve the estimates based on sample data. But it may not be the best or relevant solution to correct bias when the bias is caused by the questionnaire, sampling design, or interviewers. Post-weighting cannot produce good estimates when the questionnaire is poorly designed, and cannot be the remedy for problematic sampling design either. Weighting also surely has nothing to do with correcting bias caused by interviewers. We do not oppose weighting survey data completely, but argue that it should not be done as a routine. It must be done with thorough considerations such as why to do it, how to do it right, and what the consequences may be. Finally, post-weighting had better to be regarded as the last method to make the sample representative of the population, instead of the first thing to do as a shortcut solution.

* * *

Received : 93.10.07 ; Revised : 93.12.20 ; Accepted : 94.01.12 ◦

Appendix A: Questions Regarding the Relationship between Taiwan and China in TEDS 2001-2003

K5a: Regardless of how backward China is, I believe that being Chinese is something to be extremely proud of.

*K5b: Mainlanders eat Taiwanese rice and drink Taiwanese water. If they don't identify with Taiwan, they should go back to China.

*K5c: "Taiwanese are not Chinese." This kind of attitude is unforgiveable.

K5d: In order to control Taiwan's destiny, we must cut all ties with China and build a society of 23 million people with one common fate.

K5e: No matter how much difference there is in the standard of living between Taiwan and China, we must have patience and try to overcome it so that our country can be unified.

K5f: China is China; Taiwan is Taiwan. If Taiwan wants to seek autonomy and independence, China has no right to get involved.

K5g: Taiwan only has a future if it unifies with China.

K5h: China's history belongs to China. We want to create a history which belongs to Taiwan.

Note: * indicates that the question was not asked in TEDS 2002 and 2003.

Appendix B: Spearman Correlation Coefficients between Variables Regarding Relationship between Taiwan and China (TEDS 2001)

Unweighted

	K5A	K5B	K5C	K5D	K5E	K5F	K5G	K5H
K5A	1.000	-.234	.277	-.279	.415	-.189	.293	-.228
p(2-tailed)	.	.000	.000	.000	.000	.000	.000	.000
n	1699	1570	1504	1502	1454	1504	1437	1519
K5B	-.234	1.000	-.053	.416	-.199	.325	-.207	.377
p (2-tailed)	.000	.	.040	.000	.000	.000	.000	.000
n	1570	1725	1529	1506	1450	1513	1448	1532
K5C	.277	-.053	1.000	-.178	.284	-.174	.249	-.197
p (2-tailed)	.000	.040	.	.000	.000	.000	.000	.000
n	1504	1529	1622	1460	1411	1455	1401	1467
K5D	-.279	.416	-.178	1.000	-.305	.418	-.339	.509
p(2-tailed)	.000	.000	.000	.	.000	.000	.000	.000
n	1502	1506	1460	1606	1426	1473	1409	1477
K5E	.415	-.199	.284	-.305	1.000	-.290	.394	-.235
p (2-tailed)	.000	.000	.000	.000	.	.000	.000	.000
n	1454	1450	1411	1426	1560	1431	1385	1429
K5F	-.189	.325	-.174	.418	-.290	1.000	-.444	.427
p (2-tailed)	.000	.000	.000	.000	.000	.	.000	.000
n	1504	1513	1455	1473	1431	1629	1441	1504
K5G	.293	-.207	.249	-.339	.394	-.444	1.000	-.349
p (2-tailed)	.000	.000	.000	.000	.000	.000	.	.000
n	1437	1448	1401	1409	1385	1441	1557	1440
K5H	-.228	.377	-.197	.509	-.235	.427	-.349	1.000
p (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.
n	1519	1532	1467	1477	1429	1504	1440	1638

Weighted

	K5A	K5B	K5C	K5D	K5E	K5F	K5G	K5H
K5A	1.000	-.204	.255	-.263	.397	-.156	.273	-.201
p (2-tailed)	.	.000	.000	.000	.000	.000	.000	.000
n	1716	1581	1504	1512	1461	1518	1441	1538
K5B	-.204	1.000	-.039	.409	-.172	.325	-.207	.361
p (2-tailed)	.000	.	.126	.000	.000	.000	.000	.000
n	1581	1737	1530	1513	1452	1524	1451	1549
K5C	.255	-.039	1.000	-.153	.255	-.155	.222	-.182
p (2-tailed)	.000	.126	.	.000	.000	.000	.000	.000
n	1504	1530	1620	1453	1409	1455	1403	1469
K5D	-.263	.409	-.153	1.000	-.283	.394	-.333	.498
p(2-tailed)	.000	.000	.000	.	.000	.000	.000	.000
n	1512	1513	1453	1610	1425	1476	1413	1485
K5E	.397	-.172	.255	-.283	1.000	-.264	.379	-.213
p (2-tailed)	.000	.000	.000	.000	.	.000	.000	.000
n	1461	1452	1409	1425	1565	1439	1386	1435
K5F	-.156	.325	-.155	.394	-.264	1.000	-.430	.407
p(2-tailed)	.000	.000	.000	.000	.000	.	.000	.000
n	1518	1524	1455	1476	1439	1641	1444	1514
K5G	.273	-.207	.222	-.333	.379	-.430	1.000	-.339
p (2-tailed)	.000	.000	.000	.000	.000	.000	.	.000
n	1441	1451	1403	1413	1386	1444	1562	1449
K5H	-.201	.361	-.182	.498	-.213	.407	-.339	1.000
p (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.
n	1538	1549	1469	1485	1435	1514	1449	1655

Appendix C: Chi-square Values between Variables Regarding Relationship between Taiwan and China (TEDS 2001)

Unweighted

	K5A	K5B	K5C	K5D	K5E	K5F	K5G
K5B χ^2	333.98						
df	9						
p	.000						
K5C χ^2	343.02	248.79					
df	9	9					
p	.000	.000					
K5D χ^2	363.21	575.35	336.53				
df	9	9	9				
p	.000	.000	.000				
K5E χ^2	486.35	241.85	392.88	433.87			
df	9	9	9	9			
p	.000	.000	.000	.000			
K5F χ^2	255.34	427.34	242.47	569.91	376.45		
df	9	9	9	9	9		
p	.000	.000	.000	.000	.000		
K5G χ^2	335.61	304.90	305.33	453.57	605.29	680.57	
df	9	9	9	9	9	9	
p	.000	.000	.000	.000	.000	.000	
K5H χ^2	319.58	508.65	244.45	857.20	324.28	747.51	518.67
df	9	9	9	9	9	9	9
p	.000	.000	.000	.000	.000	.000	.000

Weighted

	K5A	K5B	K5C	K5D	K5E	K5F	K5G
K5B χ^2	305.51						
df	9						
p	.000						
K5C χ^2	338.19	232.92					
df	9	9					
p	.000	.000					
K5D χ^2	349.99	571.14	297.26				
df	9	9	9				
p	.000	.000	.000				
K5E χ^2	438.12	217.54	364.51	395.64			
df	9	9	9	9			
p	.000	.000	.000	.000			
K5F χ^2	239.59	419.17	210.75	544.53	347.88		
df	9	9	9	9	9		
p	.000	.000	.000	.000	.000		
K5G χ^2	335.25	297.19	294.06	449.28	595.58	661.78	
df	9	9	9	9	9	9	
p	.000	.000	.000	.000	.000	.000	
K5H χ^2	299.92	459.92	232.03	854.62	305.65	722.99	514.27
df	9	9	9	9	9	9	9
p	.000	.000	.000	.000	.000	.000	.000

Notes

1. TEDS 2001 successfully followed up 100 non-respondents with a short questionnaire. Their demographic characteristics and political attitudes are more or less different from the respondents', so the estimations of population parameters based on the respondents exclusively are biased (Hung, 2003).
2. It should be noted that the parameters of education used in goodness-of-fit tests and post-weighting are in fact not population parameters. Instead, they are estimates estimated by Professor Hung Yung-Tai (Chu, 2004).
3. TEDS 2003 did not ask the respondents whether they voted in the Legislative Yuan election. The relevant question asked is which party they voted for in the 2001 Legislative Yuan election. The survey data shows that the percentages of non-voters are 11% (unweighted data) and 10.9% (weighted data), so to weight the data does not make the estimate closer to the actual percentage of non-voters (33.8%).
4. The British Election Studies are face-to-face interviews and use random sampling. After the 1992 disaster, some opinion poll institutes such as Gallup and ICM started to use random sampling while others still prefer traditional quota sampling. They therefore provide solid evidence to demonstrate that what counts is how to sample, not how to weight the data (Curtice and Sparrow, 1997).
5. Respondents were asked whether they voted in the 2001 Legislative Yuan election. Those who answered yes were asked which party they voted for. 332 out of the 1643 respondents who claimed they voted in the 2001 Legislative Yuan election either refused or were unable to answer which party they voted for. They are assumed to be randomly distributed, so are recoded as missing value. It is possible that respondents' voting behaviours and whether they reveal to whom they support are correlated. But how the relationship looks like needs theoretical arguments and empirical evidence, which is beyond the scope of this research.
6. Six of the eight questions regarding the relationship between Taiwan and China were repeated in TEDS 2002 and TEDS 2003, and hence produce fifteen correlation coefficients respectively for both Taipei and Kaohsiung data sets in TEDS 2002 and for TEDS 2003. In TEDS 2002, it is found that nine out of fifteen and fourteen out of fifteen correlation coefficients attenuate

after weighting in Taipei and Kaohsiung respectively. In TEDS 2003, eleven out of fifteen correlation coefficients decrease after weighting.

7. Before each survey was conducted, the TEDS committee had to draft, discuss, and revise about 300 questions, and then select and organise about 200 questions within about six three-hours meetings, so it is impossible to consider each question in depth.
8. There have been at least 18 journal articles, 57 conference papers, and 5 MA dissertations use TEDS 2001 and 2002 data so far (Novermber, 2003). But most of them just pick out the questions related to their research without discussing in detail whether their definitions and operationalisations are different from the TEDS committees', whether the gap is acceptable, and how to adjust it given the gap, not mention how the TEDS questionnaire should be improved.
9. In TEDS surveys, every person has the same probability of being drawn as samples. If the probabilities are not equal, then the data should be weighted by the reciprocal of corresponding probabilities.
10. The difference between respondents and non-respondents is difficult to demonstrate, because the data about non-respondents is usually not available. TEDS 2001 followed up 408 non-respondents from the original samples and successfully finished 100 short interviews, which can be used to represent non-respondents. It is demonstrated that there is no significant difference between complete interviews and unit non-responses regarding political attitudes. On the other hand, the political attitudes of unit non-respondents and the substitutes are different (Hung, 2003).
11. Every district supervisor is responsible for 80, 100, and 120 successfully interviewed respondents in TEDS 2001, 2002, and 2003 respectively.
12. To weight data by raking with demographic variables is wildly used. However, raking is not the only method of weighting data. Some methods have been existed and some have been developing. For example, minimum-discrimination-information method of weighting could be better than raking under certain circumstances (Huang and Chang, 2003).

References

Abelson, Robert P., Elizabeth F. Loftus and Anthony G. Greenwald

1992 "Attempts to Improve the Accuracy of Self-Reports of Voting." In Judith M. Tanur (ed.). *Questions about Questions.* New York: Russell Sage Foundation.

Belli, Robert F., Michael W. Traugott, Margaret Young and Katherine A. Mcgonagle

1999 "Reducing Vote Overreporting in Surveys." *Public Opinion Quarterly,* 63: 90-108.

Chu, Yun-Han

2004 *Taiwan's Election and Democratization Study, 2002-2004 (II): Taiwan's Democratization and Political Change Survey.* (in Chinese). National Science Council Project Report.

Converse, Jean M.

1976 "Predicting No Opinion in the Polls." *Public Opinion Quarterly,* 40: 515-530.

Curtice, John

1997 "So how well did they do? The polls in the 1997 election." *Journal of Market Research Society,* 39: 449-461.

Curtice, John and Nick Sparrow

1997 "How accurate are traditional quota opinion polls?" *Journal of the Market Research Society,* 39: 433-448.

Groves, Robert M.

1989 *Survey Errors and Survey Costs.* New York: John Wiley & Sons.

Hsu, Sheng-Mao, Yung-Tai Hung and Chi Huang

2002 "The Research Method of TEDS 2001." (in Chinese). In Chi Huang, *Taiwan's Election and Democratization Study 2001: The Legislative Yuan Election.* National Science Council Project Report.

Huang, Chi

2003 *Taiwan's Election and Democratization Study 2002-2004 (I): The Taipei and Kaohsiung Mayoral Elections.* (in Chinese). National Science Council Project Report.

Hung, Chi and Yu-Tzung Chang

2003 "On Minimum-Discrimination-Information (MDI) Method of Weighting: An Appli-

cation to the 2001 Taiwan's Election and Democratization Study (TEDS)." (in Chinese). *Journal of Electoral Studies,* 10(2): 1-35.

Hung, Yung-Tai

2000 "The Methods of Weighting Survey Data." (in Chinese). In John Fuh-Sheng Hsieh and Shing-Yuan Sheng (eds.). *The Scope and Methods of Political Science.* Taipei: Wu-Nan.

2003 "The Nonresponse Problems of the 2001 TEDS Survey." (in Chinese). *Journal of Electoral Studies,* 10 (2): 37-58.

Jowell, Roger, Barry Hedges, Peter Lynn, Graham Farrant and Anthony Heath

1993 "The 1992 British Election: The Failure of the Polls." *Public Opinion Quarterly,* 57: 238-263.

Liu, I-Chou

1996 "Unmeasurable Error: The Bias Caused by Interviewers." (in Chinese) *Survey Research,* 2: 35-58.

Martinez, Michael D.

2003 "Comment on 'Voter Turnout and the National Election Studies'." *Political Analysis,* 11: 187-192.

Neuman, W. Russel

1986 *The Paradox of Mass Politics.* Cambridge: Harvard University Press.

Presser, Stanley

1990 "Can Context Changes Reduce Vote Overreporting?" *Public Opinion Quarterly,* 54: 586-593.

Reese, Stephen D. and Mark M. Miller

1981 "Political Attitude Holding and Structure." *Communication Research,* 8(2): 167-188.

Sniderman, Paul M., Richard A. Brody and Philip E. Tetlock

1991 *Reasoning and Choice.* Cambridge: Cambridge University Press.

抽樣調查資料之加權：正當的處理方法或是一種迷思？

劉從葦*、陳光輝**

《本文摘要》

經由抽樣設計恰當的調查研究所蒐集到的樣本資料應該能夠準確估計母體參數。但是因為單位無反應的問題，執行調查的單位或分析資料的學者通常會以加權的方式來減少樣本統計量與母體參數之間的差距。加權後的資料在人口學變項上比未加權資料較為接近母體參數，因此加權似乎是一個合理處理樣本資料的做法。

然而，即使加權是可行的解決方法，也絕非萬靈丹。在加權前也必須提出事後操弄資料的理由，而不是將加權視為理所當然。本文以台灣選舉與民主化調查為例，首先說明加權後的資料不必然較接近母體參數的原因。投票率、各政黨得票率、與婚姻狀況在加權後反而和母體參數有較大的差距。

除了單一變數分析之外，當討論的主題是變數間的關係時，加權可能增加也可能減少相關性的強度。雖然加權似乎會影響相關性，但其影響究竟是更接近真實的關係，抑或是扭曲真正的相關性則不得而知。此外，通常對整筆資料作加權只處理了單元無反應的問題，但仍然沒有解決多變量分析一定會遇到的項目無反應問題。

不論是單一變數分析或是多變量分析，在加權之前應該先嘗試其他增加樣本代表性與提高資料品質的方法。如果沒有先投入更多時間與心力在問卷設計、抽樣設計、訪員訓練與監督上，加權只是低成本

* 國立中正大學政治學系助理教授
** 加州大學聖塔芭芭拉分校政治學系博士候選人

的取巧做法。最後，假使一定要加權，必須說明與討論為什麼要加權、以哪些變數加權、如何加權以及加權所產生的影響，而非不加思考地將加權當作例行公事。

關鍵詞：加權、單位無反應、項目無反應、台灣選舉與民主化調查

審查意見

審委意見（一）

Although I have no disagreement with the author about the conclusion that weighting is no panacea, I do have some reservations about the purpose of the manuscript and the way its argument is made. First of all, since similar qualifications on weighting are abundant in the textbooks of survey methodology, I wonder why bother to add another paper to this familiar topic. Secondly, the tone of this manuscript leaves a wrong impression that investigators of TEDS treated weighting as panacea and paid little attention to question wording, sample replacement, and interviewer training, etc. However, a careful reading of the TEDS codebooks reveals just the opposite, nowhere in the TEDS documents ever claims that the weight variable should be adopted regardless of the research purpose. Instead, it is presented as a post-survey measure that incorporates unit-nonresponse adjustments and poststratification, as many multi-purpose surveys do. Trying to sell a paper by beating a straw man is pointless.

If the paper is to be publishable, its foci have to be redirected from sounding alarms to offering concrete and constructive solutions. I would like to see the author revise the manuscript in the following directions.

1.Delineate clear criteria of comparison between sample statistics and known population parameters. What are the acceptable ranges of discrepancies given the sampling errors ?

2.Choose appropriate variables to compare with. As the author keenly admits, the voter turnout and party vote shares are bad choices. It makes no sense to compare sample results of sensitive/normative questions to the population. There are simple too many sources of discrepancies and no way to differentiate the effect of weighting from others. Besides, no one claims that weighting can cure systematic bias, measurement error, or ever item nonresponse.

Propose pragmatic alternatives to the current rules and policies adopted not only by the TEDS project but other large-scale surveys. For example, the audience will be much more interested in finding a feasible a alternative to the current sample replacement policy than simply reviewing its disadvantages.

審委意見（二）

這篇文章的作者下了不少功夫指出一些抽樣調查資料的問題，然後以「加權沒有解決這些問題」爲理由來質疑加權的必要性和恰當性。

任何研究人員面臨類似這樣的論文主題第一件要做的事情應該是對「加權」有較深入的理論和應用的認知，然後再來探討加權權解決了什麼問題、沒有解決什麼問題，爲什麼？可惜作者並沒有從這個研究途徑下手。

作者以調查資料經過加權之後「投票率」、「各政黨得票率」、和「婚姻狀況」三項調查結果偏離母體參數爲例子，說明加權沒有解決問題，誠然加權是沒有解決問題，但是造成「投票率」、「各政黨得票率」、和「婚姻狀況」三項調查結果偏離母體參數的原因很多，涵蓋率、訪問失敗率、資料衡量的信度與效度至少都是主要嫌疑犯，以這個不易課責的現象要加權來負擔後果並不合邏輯。

從 TEDS 的執行報告可以看出這項調查資料是以地區、性別、年齡、與教育程度四個變數進行加權，這是通盤性、一般性的資料彌補措施，並不針對任何調查項目刻意做調整，作者如果要檢討加權的功能應該從這個角度出發，選擇各種有代表性的調查項目，比較各種彌補資料缺失的作法（其實選擇很少），然後再做評估。

從一篇研究論文的貢獻程度來看，本文結論是建議「先嘗試其他增加樣本代表性與提高資料品質的方法」，「投入更多時間與心力在問卷設計、抽樣設計、訪員訓練與監督上」，並指出「加權只是低成本的取巧做法」，這個結論和建議更加凸顯文章薄弱的科學基礎和貢獻，除了重複一些眾所皆知的事情之外，如果有其他更合理可行的選擇，作者爲什麼不放在論文裡面進行研究和評估呢？

加權基本上是處理不等機率抽樣調查的標準程序，也是彌補個案無反應缺失的常模，作者以這篇文章的邏輯和內容要質疑或挑戰這個作法，顯然是嚴重不足的。

審委意見（三）

一、中文摘要已把整篇的精華濃縮，但英文的摘要部份似乎未能把中文摘要的精華全部顯現，如中文摘要部份提到項目無反應問題；中文摘要的最後一句話，「加權所產生的影響」，這都是整篇文章想要探討的幾個主要主題，如能補上，相信可以讓英文摘要部份更能顯現整篇文章的重點探討所在。

二、在 p.151，第一段倒數第八行 "provide acceptable statistics to estimate population parameters" ，在本頁中，作者也提到TEDS 2001，35.5 %成功訪問率，19.6 %拒訪率；TEDS 2002，27.2 %的台北市成功率，30.7 %的高雄市成功訪問率；TEDS 2003，

29.9 %成功訪問率，25.1 %拒訪率。在成功訪問率 1/3，拒訪率約占訪問訪問成功率的一半或將近一樣，不知這樣的數據是否可以提供 "acceptable statistics" ，或者應該說「只供參考」而已。

三、在 p.151，作者引用 TEDS 2001、2002、2003 的資料，其中用 gender、age、education 及 area strata，其中 gender、age 及 area strata 有母體的資料，但 education 引用朱雲漢教授的推估資訊。然後作者在這之後的論點指出加權與否，並不能把未誠實回答的投票率成功地修正回來，但因爲作者引用的資料，其中有一個是推估的母體資料-education，所以不太能知道未能修正的原因，是否會因 education 這個加權變數的影響，或者是加權步驟本身就未能發揮作用。如果可以的話，作者可以將 education 這個變數拿掉，只用 gender、age 及 area strata 去作加權，再跑一次分析，相信可以更有說服力，也可消除因 education 只是推估的不確定因素，雖然結果可能跟目前作者的結論是一樣。

四、在 p.159，做與不做加權，一個很重要的假設就是回答與不回答的意見是沒有差異的。把不回答者視爲 random mission，所以作者在討論一些敏感議題，如 Vote Shares and Party Support 或 Marital Status 時，回答與回答之間，應該就有很大的差異性存在，所以加權本身當然是沒辦法有多大的作用，錯不在加權這個方法上，錯是在使用者並沒有好好檢查它的假設是否合乎。

五、在 p.165，問卷設計，作者在檢討 TEDS 的整個問卷設計過程，的確，要在六次的三個小時會議中，刪掉 100 題的題目及檢視剩下的 200 題題目，是天方夜譚，是值得檢討，但要把責任要分給使用者來分擔也無可厚非，但前提是要 TEDS Committee 肯聽，肯虛心檢討，從問卷上的題目，有些是直譯外國問卷，一字不改，美其名是與國際接軌，但與台灣國情不合，又如何能檢視台灣選舉現狀呢？作者提議是一個很不錯的對於現狀改進意見，但如果 TEDS Committee 不肯接納，又何必浪費眾多使用 TEDS 資料的學者寶貴時間呢？建議要可行，而不是說說喊「爽」而已。

六、在 p.166，抽樣設計，既然洪永泰教授在 TEDS 2001，已發現回答與不回答之間有差異，那就不太懂爲什麼在 TEDS 2002 及 2003 要有 20 個 set 的預備樣本呢？建議作者可在此項改進探討上，加上「邀請國內有實務經驗的抽樣調查的統計學者」加入 TEDS Committee，不要一群政治學者關起門來自我欣賞，一個更開放，肯接納多方意見的學術研究，才是進步的原動力，畢竟國科會這樣一個大規模的政治議題調查，再也不是單一學門的學術議題研究，而是多個學門的綜合研究。

七、在 p.165，加權與否，其實，統計在一般蒐集調查資料時，有一個很 Ratio Estimation 的統計方法，可幫助研究者在做推估時做更準確的推估，它不是一個抽樣方法，但

要找出一個與主要推估母體參數相關的變數，在蒐集資料時，加上此資料的蒐集，以便未來對母體參數做更精確的推估，也許也是一個不錯的選擇。

作者利用TEDS 2001、2002及2003的一個國內大型政治議題的學術研究，來探討是否加權的主題研究，文章指出像這樣一個有規模的學術界的調查，加權在大多數的推估上，似乎都不太有太大的作用，更何況國內每年有太多類似的不同調查，其規模或者嚴謹的態度，恐怕也不及TEDS，那加權所得到的結果，來作爲他們的推估，進而得出的結論，其可靠性值得深思。文章也提出幾個改進的調查上的議題，不僅對TEDS本身，對於其它的調查研究也應該是可以適用的。這篇文章的登出，相信對於國內不僅是選舉議題的調查，對於其它各式各類的調查研究，也是一個很好的警示及提醒，很樂意推薦此篇文章的刊登。

審查意見答覆

感謝三位審查人的悉心指教，在針對審查人的意見分別提出說明之前，首先再次說明撰寫本文的原因。加權可以是非常有用的工具，也可能是造成更大偏差的原因，端視研究者如何使用。就像刀子可以是有用的工具，但不分人事時地物地拿著刀亂揮，必定是危險的行爲。本文的目的不是贊成或反對加權，而是強調研究者必須問自己爲什麼要加權？要怎麼加權？加權產生的效果是什麼？假使沒有問過這些問題就盲目加權，那就只是個加權迷思，而非資料處理的進階做法。本文以TEDS爲例來檢驗加權的效果，結果發現加權有時反而使得本已偏差的統計量更加偏離母體參數，這也更加支持必須要認真思考上述問題的正當性。

審委意見（一）

1.It is idiosyncratic to researchers about what are the acceptable ranges of discrepancies between sample statistics and known population parameters. We think judgments should be made by researchers given their specific research topics.

2.In our opinions, the turnout and party vote shares are not bad choices, but tough tests for surveys. However, no matter the issues in question are simple facts or sensitive attitudes, sample statistics are all expected to be as close to population parameters as possible. Some variables are much more difficult to be measured than others, so what needed is to design instruments of measurements carefully rather than take the discrepancy for granted.

3.It is quite important to propose pragmatic alternatives to the rules or policies adopted by the

TEDS surveys and we did have some suggestions for the TEDS team. However, this piece of research is about the legitimacy of weighting data, not how to improve the TEDS surveys. It deserves another paper to discuss this issue in detail.

審委意見（二）

本文以「投票率」、「各政黨得票率」、「婚姻狀況」三個有母體資料的變數來比較加權前後的變化，發現根據未加權資料的估計偏離母體參數。筆者完全同意審查人所提的可能原因，筆者在文中也提出許多造成偏差的可能原因。例如受訪者傾向符合社會期待而使投票率的估計偏高，但這些原因都不是加權所能解決的，要解決問題，需要的是對症下藥，例如如何在問卷設計上讓來自社會期待的偏差降到最低，而不是以加權的方式讓性別、年齡、區域、教育等變數與母體一致後，就彷彿「投票率」、「各政黨得票率」、「婚姻狀況」也會自然地與母體一致。TEDS 的資料甚至顯示，加權讓本已偏差的估計更加偏離母體參數。

性別、年齡、區域、教育的確是少數可以彌補一般資料缺失的加權變數，但如果加權只是無意識的行爲或是讓資料看起來很漂亮的低成本取巧做法，比較沒有加權的資料反而更能區隔執行單位與資料品質的好壞。例如在電話調查中，不論是採用任意成人法或是做戶中抽樣，事後只要加權，兩筆資料看起來是一樣漂亮的。但不加權才能使粗劣的執行過程現出原形。

先嘗試其他增加樣本代表性的方法、投入更多心力在問卷設計上等等建議的確是眾所週知，但這些教科書上的叮嚀卻是知易行難，也難以抵擋低成本取巧的加權誘惑。本文的建議確實是一般常識，但當這麼簡單的一般常識還需要大費周章地寫成論文時，思考所隱含的現象與意義也許更爲重要。

原文註 9 即已說明，如果是不等機率抽樣的調查，加權是標準的處理程序。本文討論的焦點是針對單位無反應所做的加權，所以並沒有反對針對不等機率抽樣所做的加權。

審委意見（三）

筆者完全同意審查人的建議與看法。審查人在問卷設計、抽樣設計、與 ratio estimation 的建議皆十分寶貴，所以審查意見的刊登對往後包括 TEDS 在內的所有抽樣調查必定有相當大的幫助。唯有不斷地反省、發聲與行動才能使學術研究持續地進步。