Do You Remember? Proxy Bias and Recall Bias in Social Mobility Studies: Evidence from TEPS and TEPS-B

Yung-Yu Tsai*

Sep. 2018

Abstract

Social mobility studies often use cross-sectional data, therefore, these studies highly rely on proxy-reported and recall data to get information including respondents' family background and their personal education or employment histories. However, this approach has been severely criticized because of its potential proxy bias and recall bias.

Using Taiwan Education Panel Survey (TEPS) and its beyond (TEPS-B), I estimate the gap between parental self-reported data and children's proxy, recall data. Besides, I also estimate the gap between personal first job information answered by respondents in two different years.

The results from this study suggest that proxy bias and recall bias of parental education are slight. Nevertheless, there is a significant, systemic bias when it comes to parental occupations. Furthermore, it has higher validity when using self-reported data than using the proxy one. On the other hand, it seems that the respondents recognized different jobs when surveying their first full-time occupation in two different years.

Even if only take those samples with consistent job start years into consideration, there is still a bias when using recall data to get first job information, but the problem is not serious and can be acceptable. The implications of these findings are discussed.

Keywords— proxy bias, recall bias, social mobility, TEPS

*e-mail: 106256028@nccu.edu.tw

1 Introduction

Researches of social mobility focus on the correlation between social class original (such as family background) and destination. These studies need to collect variables, which happen in different periods of one's life, including parental education, occupation, and income as well as personal education and career history. However, longitude panel data is always costly and scarce. Therefore, researchers turn to cross-sectional data, which require respondents to recall memories about their parental educations and occupations in their childhood (usually at age 15) and their personal experience of schools and jobs. Unfortunately, this approach has been severely criticized because of its potential proxy bias and recall bias (Breen & Jonsson, 1997; Massagli & Hauser, 1983; Plewis & Bartley, 2014).

In the other words, when requiring respondents to recall their childhood memories and to provide proxy answers of their parental socioeconomic status, there would probably form a gap between collected information and real cases. The similar problem happens in another situation of respondents' personal education and career history. As a result, based on these data, the inference and analysis would be highly doubtful.

Though the problem could be solved by using longitude panel data, there still has some insoluble challenges, such as high cost of money and time, mortality from samples, and untimely inappropriateness of questionnaires. Hence, using proxy and recall data is still an inevitable option for researchers of social science. Accordingly, it's important to understand the character and seriousness of these biases.

By using Taiwan Education Panel Survey (TEPS) and its beyond (TEPS-B), this article merges questionnaire from parents in 2001 and interview surveys from students in 2010 and 2015. Then, I compare the answers provided by parents and students in different years, analyzing and discussing the existence and seriousness of proxy bias and recall bias. To be more specific, this article seeks to answer the following questions:

1. Is there any bias in children's proxy, recall responses of their parental education and occupation? What's the main types and extents of these biases?
2. Is there any bias in personal recall responses of his or her first job information? What's the main types and extents of these biases?
3. How would the above types and extents of biases affect statistical inference in research of social mobility? Is there any way to avoid or reduce this problem?

This article aims at dealing with the above research questions and looks forward to improving the quality of data and validity of research result in the future. In section 2, I discuss some important literature relating to bias and social mobility. In section 3, details of using data and research method are elaborated. In section 4, I would present results and findings. Finally,

the main conclusions and the following suggestions are provided in the last section.

2 Literature review

2.1 Proxy bias and recall bias

social science research pays attention to examining and confirming causal relationship, so any existence of bias would hamper researchers' intention of observing the correlation between independent variables and dependent variables, and this would become a threat to internal validity of research. Sackett (1979) and Choi (2000) classify bias into three types: selection bias, informational bias and confounding bias. Informational bias refers to bias happening in process of data collecting. This is, the observation value obtained by the researcher cannot reflect the reality of respondents' situations.

Proxy bias and recall bias belong to kind of informational bias. The former results from data collected from others beside the observed one (such as proxy answers about the income of heads of the household from other family members, but not the heads themselves), while the later caused by information relies on respondents' personal memories about their past (Hassan, 2006; Massagli & Hauser, 1983).

We should notice that "bias" is different from merely "error" from ignorance or amnesia. Bias means systematical, unrandom error, and will have a seriously negative effect on the estimation of research. The bias may come from the fact that a respondent intentionally or unintentionally revises personal memories or adjusts provided answers, since he or she seeks to rationalize his or her personal experience or to conform the expectation of the researcher. For example, other study finds that in research of voting behavior, researchers tend to overestimate the voting consistency of individuals while using recall data from respondents (Dassonneville, 2013; Himmelweit, Biberian, & Stockdale, 1978). Epidemiologists also find that when a respondent has a certain disease, he or she usually tends to seek for a rational explanation of this situation, and he or she would exaggerate the certain event happening in past (Delgado-Rodriguez & Llorca, 2004; Pannucci & Wilkins, 2010; Raphael, 1987). Researches of social mobility also conclude that while asking fathers to provide information about their children's class of occupation, they tend to minimize the gaps between their children and themselves; in contrast, while asking children to offer answers about their fathers' class of occupation, they tend to maximize the gaps between their fathers and themselves (Broom, Jones, McDonnell, & Duncan-Jones, 1978).

On the other hand, even if the bias comes from ignorance or amnesia, there could also be a systematical error. To be more specific, the possibilities and extents of wrong cognition and forgetfulness may have correlations with certain demographic variables that concerned by researchers. For instance, a study indicates that while surveying personal career changing, recall bias

will be aggravated if the respondent has a complex, varied career. Moreover, if a person ever be unemployed, he or she would probably miss providing this experience (Manzoni, Vermunt, Luijkx, & Muffels, 2010). In addition, when the researcher requires other family members to provide information on personal wage of the original respondent, the gender and marital status of this original respondent affect the extent of bias (Tamborini & Kim, 2013). In other words, errors of measurement are not random, and they would be related to some demographic characters.

Some studies contend that proxy bias and recall bias are not severe, indicating that respondents can clearly remember past experiences and that proxy interviewees can bring out reliable information (Bassett, Magaziner, & Hebel, 1990; Krieger, Okamoto, & Selby, 1998; Manzoni et al., 2010; Pless & Pless, 1995). On the contrary, other studies maintain that proxy bias and recall bias result in serious problems of overestimation or underestimation and systematical error that affect the result of inference (Broom et al., 1978; Himmelweit et al., 1978; Perry & Felce, 2002; Tamborini & Kim, 2013).

Notwithstanding this disagreement, all the studies agree that the existence and degree of proxy bias and recall bias will be affected by the type of data, length of recalling period, the relationship between the proxy interviewees and original respondents. Overall, when it comes to education, which is straightforward, the bias would be slight; conversely, when it comes to the class of occupation, which is more complicated, the bias would be increased (Krieger et al., 1998; Manzoni et al., 2010). Similarly, when the event that the respondent required to recall happened in a recent past, the recall bias would be decreased (Pless & Pless, 1995).

2.2 Concept and measurement of social mobility

The degree of social mobility is defined as the extent of correlation between status origin and destination. In a society, if the class statuses of people are highly related to their parents, then we recognize this situation as class rigidity or class reproduction. In contrast, if there are huge gaps between the statuses of parents and children, it means the social mobility of this society is high.

However, researches use different tools of measurement to measure social status. There are at least three different perspectives (Ganzeboom & Treiman, 2010): socioeconomic status (SES), occupational prestige, and class status. SES represents a combination of knowledge and economic power of certain occupation. Generally, researchers define SES as the average of educational level and wage of a certain occupation, so this approach is comparatively objective. On the contrary, occupational prestige refers to judgments of certain occupation by the general population, including moral impressions, social contributions, and socioeconomic status. Lastly, class status, which is related to Marxist class theory, categories occupations into different classes by criteria including whether possessing capital, whether acquiring skills, whether employed by others and whether dominating subordinates, etc. The results of

these three measurements are related but still different. Therefore, researchers should choose suitable approach base on their research designs and purposes.

Take the variety of measurement of occupation into consideration, this article discusses both International Standard Classification of Occupations (ISCO) and The EGP Class Scheme, which is initiated by Erikson and Goldthorpe (1992). The former can be transcoded into SES or occupational prestige, and the later is a kind of criterion of class status.

Table 1 shows the comparison between this two scale. The main distinctions are: first, the categories of service class and routine non-manual workers in EGP scheme are not as meticulous as ISCO; secondly, EGP scheme categories petty bourgeoisie into independent type, while ISCO categories these people into others group basing on real working content; lastly, although both ISCO and EGP distinct farmers from other laborers, ISCO cuts farmers into two groups by whether possessing skill but EGP cuts farmers into two groups by whether being an employer.

Table 1: Comparison between ISCO and EGP

ISCO		EGP	
code	category name	code	category name
1000	Managers	I+II	Service class
2000	Professionals		
3000	Technicians and associate professionals	III	Routine non-manual workers
4000	Clerical support workers		
5000	Service and sales workers		
X	Category into other groups	IV	Petty bourgeoisie
6000	Skilled agricultural, forestry and fishery workers	IVc	Farmers
7000	Craft and related trades workers	V+VI	Skilled workers
8000	Plant and machine operators, and assemblers		
9000	Elementary occupations	VIIa	Non-skilled workers
		VIIb	Agricultural labourers

*N*otes: This article uses ISCO-2008 (without armed forces) and EGP Scheme (the version with 7 categories).

*S*ource: Connelly, Gayle, and Lambert (2016); Ganzeboom and Treiman (2010)

3 Research method

3.1 Data

This article uses Taiwan Education Panel Survey (TEPS) and its beyond (TEPS-B). The subjects of these surveys are students who attended grade 11 (senior-high school) in 2001 in Taiwan, referring to the cohort that born in 1984 or 1985. This article merges the questionnaire from parents in 2001 and the interview surveys from children in 2010 and 2015. Hence, by holding different types of data, including self-report or proxy-report and recall or non-recall data, at the same time, this article compares the inconsistence between these data and discusses the existence of bias. The details of used variables and its characters are showed in table 2.

Table 2: Survey year, respondent, and data type of variables

Variable	Survey year	Respondent	Proxy	Recall
parental highest education	2001	parents	self-report/ spouse proxy	non-recall*
parental highest education	2010	children (subjects)	proxy	non-recall*
parental occupation in 2001	2001	parents	self-report/ spouse proxy	non-recall
parental occupation in 2001	2010	children (subjects)	proxy	recall
Personal first job	2010	children (subjects)	self-report	recall, but shorter
Personal first job	2015	children (subjects)	self-report	recall, and longer

* although the event of attending a school occurred in the past, the highest education is a situation exists in the present.

The numbers of valid samples in the 2001 questionnaire, the 2010 survey and 2015 survey are 16,266, 3,977 and 9,011. Since this article needs to combine at least two datasets for analysis, the last numbers of observations adopted in research will be reduced. As shown in Figure 1, these numbers are between 2,550 to 3,476.

3.2 Research design

3.2.1 Consistency and correlation of response

First, this article compares the answers with regard to parental education acquiring from 2001 parental questionnaire and 2010 children's survey. Since education is an ordinal scale variable, this article uses descriptive statistics to check consistency and uses Spearman's rank correlation coefficient to examine the correlation between self-report data and proxy data.

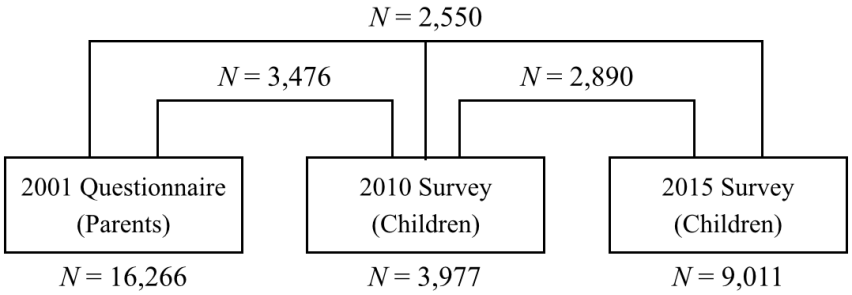


Figure 1: Number of observations

Second, this article compares the responses to parental occupation from 2001 parental questionnaire and 2010 children's survey. I measure occupation in both ISCO and EGP scheme. The 2010 survey used ISCO to measure occupation, so there is nothing need to be adjusted for this research. However, in the original version of 2001 questionnaire, the survey institution didn't adopt ISCO, so, this article recodes this variable basing on other questions, such as working status, whether an employer or not, industrial classification of their job, the position in their firm, etc. On the other hand, to classify occupation into EGP scheme, I also refer the questions mentioned above. Since occupation is also be taken as an ordinal scale variable in ISCO and EGP scheme, this article uses descriptive statistics and Spearman's rank correlation coefficient to analyze.

Finally, discussing the variable of the personal first job, this article compares subjects' self-report answers in 2010 and 2015, meaning that both results are recall data, but the recall period of the former data is 6 years shorter than the latter. Once again, this article uses descriptive statistics and Spearman's rank correlation coefficient to analyze. While the 2010 survey asked the respondents to recall all of their career histories and provide information about occupational classification, the beginning year and average work hours per week, the 2015 survey only required the subjects to provide information of their first full-time job. Given this different questionnaire design, when dealing with 2010 survey, this article captures the first job that its average work hours per week is above 30 hours to represent the first full-time job of a certain subject.

3.2.2 Valid evaluation

Besides checking the consistency and correlation of data acquired from different sources, this article also aims to further understand whether self-report, non-recall data are more reliable than the proxy ones, recall data. Accordingly, this article analyzes the validation in both kinds of data. In theory, the level of education of an individual should be related to his socioeconomic status, and the educational level and occupation of a husband and wife should be related to each other. At the same time, the occupational socioeconomic status of this couple should predict family income (Ganzeboom & Treiman, 2010).

Therefore, this article will convert the level of parental education into the number of years of education, and recode occupation into SES according to the International Professional Social Status Scale (ISEI), which was proposed by Ganzeboom and Treiman (2010). Then, the correlation analysis and regression model are used to test the above relationship for the aim that understanding whether self-report, non-recall data have a better construction validity.

4 Result

4.1 Consistency and correlation of response

4.1.1 Parental education

Table 3 shows the result of comparing the answers about parental education acquired from the 2001 questionnaire (filled by parents) and the 2010 survey (answered by children). In 2001, since the questionnaire didn't send to both fathers and mothers but was filled by one of the couple, the following table discriminates the statistical result of self-report data from spouse proxy data. As shown in the table, the proxy data provided by children no matter compare with self-report data or spouse proxy data, the ratios of consistency are all above 85%. This indicates that using children proxy data on parental education won't cause unacceptable serious bias. On the other hand, although this article cannot directly compare spouse proxy responses to self-report data, the fact that the ratios of consistency of self-report data with children proxy data and spouse proxy data with children proxy data are close implies that there is only a small gap between self-report data and spouse proxy data.

Table 3: Consistency of parental education

	Father		Mother		Total
	self-report	spouse proxy	self-report	spouse proxy	
Consistency with children proxy	83.76%	83.83%	86.69%	84.86%	84.84%
Inconsistency with children proxy	16.24%	16.17%	13.31%	15.14%	15.16%
N=	1,459	1,911	1,893	1,453	6,716

Table 4 presents the Spearman's rho of parental report data and children proxy data. The result indicates that all the coefficients are significant and greater than 0.85, which shows a very strong positive correlation. In brief, when it comes to parental education, children proxy data is little different from parental self-report data. Therefore, it will be acceptable to adopt proxy data, if the acquisition of self-report data is somewhat limited.

Table 4: Spearman's Correlation of parental education

	Father		Mother	
	self-report	spouse proxy	self-report	spouse proxy
Spearman's rho	0.8791***	0.8886***	0.8938***	0.8598***
N=	1,459	1,911	1,893	1,453

*$p<.05$ **$p<.01$ **$p<.001$

Furthermore, table 5 is a parental answers by children answers table, in which the cell values are percentages of row marginal totals: i.e. the table shows how in each case of given parental answers of their highest education are distributed by children proxy responses. For example, in the case that parents report their highest education are "junior high school or below", there are 94.38% of children provides the same answers as their parents; in contrast, 4.82% of children wrongly maintain that their parents had attended senior high school, and 0.12% of children misrecognize their parents have junior college or professional school degrees, and so on.

Table 5: Distribution of parental answers by children answers of parental education

Parental answers		Children's answers					
		1	2	3	4	5	Total
1	Junior high school or below	94.38%	4.82%	0.12%	0.65%	0.04%	100.00%
2	Senior high school	11.84%	81.38%	5.79%	0.80%	0.19%	100.00%
3	Junior colleges or Professional school	2.68%	15.73%	70.05%	9.79%	1.75%	100.00%
4	Academic bachelor	0.92%	2.22%	9.43%	80.22%	7.21%	100.00%
5	Graduate school	1.80%	4.50%	2.70%	10.81%	80.18%	100.00%
N=		6,716					

Notes: Cells shaded grey indicates that children's answers are consistent with their parents

The table points out that the most serious bias happens in the case that parental highest educations are "Junior colleges or Professional school", and the percentage of consistency is only 70.05%. Moreover, the type of bias mainly results in which children misclassify their parents into the groups of "Senior high school" or "Academic Bachelor". In Taiwan, though it takes

students three years to graduate from senior vocational high school and five years for junior college, both of these degrees are vocational and are opening to students who have already graduated from junior high school. Therefore, the children probably understand that their parents attended a vocational school and possess a degree above junior high schools level, but they don't really know the exactly educational credentials that their parents received.

Another type of bias catches our eyes is that children tend to underestimate their parental education. As shown in table 5, in the case that parents got degrees above senior high schools, the percentages of underestimation are all higher than the overestimation one. Overall, the ratio of underestimation is 8.29%, while the ratio of overestimation is 6.86%. As a result, though the proxy bias is slight when it comes to parental education, the research of social mobility will probably make a mistake of overstating the intergenerational mobility of education or understating parental education's effect on children's status attainment. This problem should be taken into consideration when using proxy data of parental education from children.

4.1.2 Parental occupation

As revealed in the section of literature review, the most common measurements of occupation in research of social mobility are ISCO and EGP scheme, so in the following section, I discuss these two approaches separately.

International Standard Classification of Occupations (ISCO)

Table 6 illustrates the comparing result of the answers about parental occupation from parents and children. The parents gave their answer about their "present" occupation, which means their occupation in 2001 (when their children are around age 16), and the children provided their responses of recall information about their parental occupation when they were 15 in 2010 (which means they have to recall memory 10 years backward). The result indicates that only one-third of the subjects provided the same answers comparing with their parents (diff=0). 20% of the samples show small differences, which are equal to one scale of class (diff=1), between parents and children. However, there are around half of the samples exist a serious bias, that the differences are above two scales of class (diff>1).

Table 6: Consistency of parental occupation (measuring in ISCO)

	Father		Mother		Total
	self-report	spouse proxy	self-report	spouse proxy	
diff=0	28.15%	26.87%	36.74%	33.48%	31.41%
diff=1	22.66%	21.02%	19.81%	18.42%	20.47%
diff>1	49.19%	52.12%	43.45%	48.09%	48.12%
N=	1,112	1,418	1,459	1,102	5,091

On the other hands, the difference of parents self-report data from children's answers is slightly smaller than spouse proxy data from children's answers. Although this article cannot directly compare self-report data with spouse proxy ones, the indirect information provided by table 6 can help infer that spouse proxy data are different from self-report one. In other words, even if there is no recall bias (since the spouse only be required to provide his or her spouse's present occupation at that time) and the proxy person is someone closely (such as their spouse), there will still exist a bias. This inference underlines the complexity of measurement in occupation, lots of information relating to occupation classification can only be provided by the subjects themselves.

In addition, we have to notice that the bias of maternal occupation is slighter than paternal one. It can be explained that mothers are usually more closely with their children and willing to share their experience in work. Another alternative possibility is that the labor force participation rate of female is lower, and if a mother didn't work, it's less possible for her child to give a wrong answer.

Table 7 presents the Spearman's rho of parental report data and children proxy data. It's shown that although the answers between parents and children have a significant correlation, the Spearman's rho are only around 0.34 to 0.47, which are merely moderately correlated, revealing that the bias is conspicuous.

Table 7: Spearman's Correlation of parental occupation (measuring in ISCO)

	Father		Mother	
	self-report	spouse proxy	self-report	spouse proxy
Spearman's rho	0.4640^{***}	0.4747^{***}	0.4297^{***}	0.3366^{***}
$N=$	1,112	1,418	1,459	1,102

$^{*}p<.05$ $^{**}p<.01$ $^{**}p<.001$

Table 8 is a parental answers by children answers table, in which the cell values are percentages of row marginal totals. For instance, in the case that parents reported their occupation were "managers" (1), 43.2% of their children provide the same answers as their parents, while 6.51% wrongly indicate that their parents were "professionals" (2) and 15.35% of children say that their parents were "Technicians and associate professionals" (3), and so on. It's shown in the table that when it comes to "unemployed" (0), "managers" (1) or "professionals" (2), the bias is much slighter. This is because these three categories of occupation are clearly and precisely defined. the class "managers" means people who plan, direct, coordinate and evaluate the overall activities of enterprises, governments and other organizations, and the titles of these kinds of people are usually managers. "Professionals" indicates people who increase the existing stock of knowledge, apply scientific or artistic

concepts and theories, teach about the foregoing in a systematic manner, or engage in any combination of these activities, and these kinds of people usually received some graduate degree, passed certain national examinations or possess professional certifications.

Table 8: Distribution of parental answers by children answers of parental occupation (measuring in ISCO)

Parental answers	Children's answers										
	0	1	2	3	4	5	6	7	8	9	Total
0	59.18%	1.46%	0.29%	2.33%	5.54%	12.24%	4.08%	8.16%	3.50%	3.21%	100%
1	5.58%	43.02%	6.51%	15.35%	10.93%	5.35%	0.47%	3.95%	7.44%	1.40%	100%
2	5.95%	12.32%	41.89%	18.48%	10.06%	5.13%	0.41%	3.08%	2.05%	0.62%	100%
3	8.24%	9.62%	7.69%	25.55%	4.95%	11.26%	0.27%	23.35%	7.97%	1.10%	100%
4	16.38%	4.74%	2.16%	28.30%	28.74%	5.60%	0.57%	4.02%	6.03%	3.45%	100%
5	11.50%	5.18%	1.51%	12.58%	9.20%	36.59%	1.01%	10.21%	7.55%	4.67%	100%
6	0.00%	16.67%	0.00%	0.00%	16.67%	50.00%	0.00%	0.00%	16.67%	0.00%	100%
8	7.10%	12.65%	0.31%	8.64%	14.81%	2.78%	0.00%	17.28%	32.72%	3.70%	100%
9	14.86%	1.43%	0.10%	2.48%	6.48%	4.76%	2.00%	26.10%	32.38%	9.43%	100%
N=	5,091										

*N*otes: Cells shaded grey indicates that children's answers are consistent with their parents.
0 = Unemployed, 1 = Managers, 2 = Professionals, 3 = Technicians and associate professionals, 4 = Clerical support workers, 5 = Services and sales workers, 6 = Skilled agricultural, forestry and fishery workers, 7 = craft and related trades workers, 8 = Plant and machine operators and assemblers, 9 = Elementary occupations.

Another considerable bias is that children confound "Technicians and associate professionals" (3) with "Clerical support workers" (4). The main difference between these two categories is "Technicians and associate professionals" assist in professional works under the supervision of professionals, while "Clerical support workers" participate in non-professional works. However, these two kinds of people both usually have position title "assistant", or their duties mix professional and non-professional works. Therefore, it's not surprising that these two categories are easy to be confounded.

Although the bias in "skilled agricultural, forestry and fishery workers" (6) is also high, the number of samples is limited (where N=6). Hence, statistical inference of this category is meaningless. In addition, children also tend to provide inconsistent answers in the situation that their parents were "Craft and related trades workers" (7), "Plant and machine operators, and assemblers" (8) or "Elementary occupations" (9). These three occupations all belong to the class of laborer or blue-collar worker. The key distinction between these are "Craft and related trades workers" need specific technical and practical knowledge and skills, while "Plant and machine operators, and assemblers" only possess low-level

technic, and "Elementary occupations" just involve the performance of simple and routine tasks, which are deskilling. Nevertheless, there is no clear criteria for whether possessing technic and knowledge or the level of technic. Children would probably know that their parents were laborers working in factories, but don't understand their actual job detail. Another thing needs to be noticed is that in the situation that parents belong to "Plant and machine operators, and assemblers", a high portion of children recognize their parents were "managers". It can be speculated that the title of their parents were "supervisors" or "team leaders", so the children think their parents belong to the position of managers.

In conclusion, the cause of the huge gap between answers provided by parents and children may result from the complication of classification of occupation. To classify a certain occupation according to ISCO, the respondents must know the industry, position, title and job content of the original subjects. Therefore, this information can only be provided by the subjects themselves, and proxy data may result in bias.

EGP Class Scheme

Table 9 illustrates the comparing result of the answers about parental occupation from parents and children, measuring in EGP Scheme. The result indicates that around 40% of the subjects provided the same answers comparing with their parents (diff=0). a quarter of the samples show small differences, which are equal to one scale of class (diff=1), between parents and children. However, there is one-third of the samples exist a serious bias, that the differences are above two scales of class (diff>1).

Table 9: Consistency of parental occupation (measuring in EGP Scheme)

	Father		Mother		Total
	self-report	spouse proxy	self-report	spouse proxy	
diff=0	50.90%	48.38%	39.03%	37.39%	43.96%
diff=1	23.14%	23.47%	27.66%	25.64%	25.04%
diff>1	25.96%	28.14%	33.31%	36.98%	31.00%
N=	1,275	1,606	1,609	1,217	5,707

Moreover, no matter only taking the case of self-report or spouse proxy into consideration, the results are similar. However, it's obvious that children understand their fathers' occupation better. This finding is contrary to the case in ISCO. Checking the detail of comparison, I find that in the situation when mothers maintain themselves as "Petty bourgeoisie", a high portion (three-quarters) of children provide inconsistent

answers with their parents. In Taiwan, it's common that a couple run a small business together (such as small and medium enterprises or street vendors). In these cases, the couple may tend to recognize each other as the boss, while their children may think their fathers are employers and mothers as employees.

Table 10 presents the Spearman's rho of parental report data and children proxy data. The result indicates that all the coefficients are significant but all below 0.5, which are merely moderately correlated, revealing that the bias is conspicuous.

Table 10: Spearman's Correlation of parental occupation (measuring in EGP Scheme)

	Father		Mother	
	self-report	spouse proxy	self-report	spouse proxy
Spearman's rho	0.4920^{***}	0.4685^{***}	0.4137^{***}	0.3280^{***}
$N=$	1,275	1,606	1,609	1,217

$^{*}p<.05$ $^{**}p<.01$ $^{**}p<.001$

Table 11 is a parental answers by children answers table, in which the cell values are percentages of row marginal totals. For instance, in the case that parents reported their occupation were "Service class" (1), 55.9% of their children provide the same answers as their parents, while 23.83% wrongly indicate that their parents were "Routine non-manual workers" (2) and 7.07% of children say that their parents were "Petty bourgeoisie" (3), and so on.

As shown in the table, in the case that parents were "Unemployed" (0), "Service" (1), "Routine non-manual workers" (2), "Petty bourgeoisie" (3), or "Skilled workers" (5), the bias is relatively small. Since EGP Scheme is rougher than ISCO, children can provide the right answers even if they don't understand their parental occupation completely. At the same time, EGP Scheme separates Petty bourgeoisie from other occupations also help avoid misclassifying these people into managers.

On the other hand, a greater bias happens in the situation that parents belong to "Farmers" (4), "Non-skilled workers" (6), or "Agricultural labourers" (7). As the class of Petty bourgeoisie, Farmers also own the means of production and don't employ a huge number of employees. The main difference between these two class is their industry. The same distinction also can be found in Non-skilled workers and Agricultural labourers. Therefore, the bias shown in these cases reveals that even

when it comes to industry, a relatively clear criterion, proxy data provided by children still are inconsistent with their parents. Besides, children also tend to misclassify their Non-skilled workers parents into the category of skilled workers, indicating that it's hard to confirm whether a worker possesses professional technic.

Table 11: Distribution of parental answers by children answers of parental occupation (measuring in EGP)

Parental answers	Children's answers								
	0	1	2	3	4	5	6	7	Total
0	62.54%	2.48%	9.60%	11.46%	0.93%	9.60%	2.79%	0.62%	100%
1	5.70%	55.99%	23.83%	7.07%	0.11%	5.93%	1.03%	0.34%	100%
2	16.10%	10.11%	48.04%	9.58%	0.27%	12.38%	3.19%	0.33%	100%
3	7.54%	8.44%	19.30%	53.38%	0.54%	8.62%	2.05%	0.12%	100%
4	8.93%	21.43%	1.79%	28.57%	16.07%	8.93%	5.36%	8.93%	100%
5	6.80%	12.62%	21.36%	7.28%	0.00%	46.60%	5.34%	0.00%	100%
6	15.98%	2.54%	11.01%	7.72%	1.27%	51.53%	9.10%	0.85%	100%
7	10.07%	3.60%	10.07%	13.67%	30.22%	12.23%	7.19%	12.95%	100%
N=	5,707								

*N*otes: Cells shaded grey indicates that children's answers are consistent with their parents.
0 = Unemployed, 1 = Service class, 2 = Routine non-manual workers, 3 = Petty bourgeoisie, 4 = Farmers, 5 = Skilled workers, 6 = Non-skilled workers, 7 = Agricultural labourers

Comparing ISCO with EGP Scheme

The inconsistency between parental answers and children responses are both huge in ISCO and EGP Scheme. However, the inconsistency is higher when measuring occupation in ISCO than EGP, because that EGP Scheme is relatively easy and rough.

Furthermore, checking the type of this bias, as shown in table 12, no matter using which scheme or considering fathers or mothers, children always tend to underestimate their parental occupation. In other words, researchers would overstatement the rate of upward mobility when relying on children's proxy data.

In addition, comparing the difference between fathers and mothers, it's much more common for children to underestimate their mothers' occupation than fathers. On the other side, children perform better in proxy their mothers than fathers when measuring in ISCO, while they

perform poorly in proxy their mothers than fathers when measuring in EGP scheme.

Table 12: the ratio of underestimation and overestimation of children's answers about parental occupation

	ISCO			EGP Scheme		
	Father	Mother	Total	Father	Mother	Totla
Underestimate	41.26%	48.38%	44.84%	28.81%	46.14%	37.39%
Consistency	27.43%	35.34%	31.41%	49.50%	38.32%	43.96%
Overestimate	31.30%	16.28%	23.74%	21.69%	15.53%	18.64%
N=	2,530	2,561	5,091	2,881	2,826	5,707

To conclude, this finding reveals that the gap between answers provided by two generation is not merely random errors, but systematical biases. The extent and type of these biases vary from different measurements and subjects. Hence, researchers should take these into consideration when using proxy, recall data.

4.1.3 Personal first job

We can obtain information about personal first jobs in both TEPS-B 2010 survey and 2015 survey. In fact, not all the respondents just begin their first job at the time being surveyed, so they have to recall their memories to some degree. However, since 2010 is former than 2015, the recall period would be shorter, and the bias may become slighter in 2010.

Table 13 shows the beginning year of the first job provided in two survey years. We have to notice that the narration of questions in two survey years are different, so the same respondent would recognize different first jobs in two years. therefore, I present these result separately in the following table.

As shown in the table, in 2010 survey, the average beginning year of the first job is 2005 to 2006; since the survey year is 2010, the respondents are required to recall memories happened 4-5 years ago on average. In 2015 survey, the average beginning year of the first job is 2008 to 2009, and this means that the respondents should recall memories happened 6-7 years ago. In two survey years, the longest recall period is 13 years and 16 years.

If only take samples with same begin year into consideration (account for one-third of all samples), the average beginning year of the first job is 2007 to 2008, and the average recall period is 2.5 years in 2010 and 7.5 years in 2015.

Table 13: Descriptive statistic of begin year of personal first job

		N=	Mean	SD	Min.	Max.
All samples						
2010	beginning year	2,645	2,005.74	2.76	1,997	2,010
	Recall year	2,645	4.26	2.76	0	13
2015	beginning year	2,490	2,008.53	2.56	1,999	2,015
	Recall year	2,490	6.47	2.56	0	16
Samples with same beginning year						
2010	beginning year	766	2,007.50	1.87	2,000	2,010
	Recall year	766	2.50	1.87	0	10
2015	beginning year	766	2,007.50	1.87	2,000	2,010
	Recall year	766	7.50	1.87	5	15

Table 14 illustrates the result of comparing the answers of first jobs provided by subjects in 2010 and 2015 survey. Only one-third of samples recognize the same beginning year in two surveys. Even treat one year former or later as reasonable error, there are still half of the all samples show a difference above two years. This is, the respondents extremely possible refer to different jobs in these two years.

Table 14: Consistency of personal first job in two survey years

Consistency	Beginning year	ISCO		EGP	
		All samples	Samples with same beginning year	All samples	Samples with same beginning year
diff=0	33.58%	41.50%	60.70%	66.63%	79.68%
diff=1	17.05%	23.62%	25.27%	15.05%	12.57%
diff>1	49.36%	34.89%	14.04%	18.32%	7.75%
N=	2,281	2,511	748	2,511	748

As mentioned in the section on literature review, the questionnaire design of 2010 is different from 2015. The 2010 survey asked the respondents to recall all of their career histories, including full-time and

part-time job, and this article determines whether a certain job is full-time based on whether the average work hours per week of this job is above 30 hours. In contrast, the 2015 survey only required the subjects to provide information about their first full-time job. If being asked by respondents, the interviewer would explain that the definition of full-time job is the job with average work hours per week above 30 hours. However, if not being asked, the interviewer would not provide this information actively.

Supposing that a subject had worked as a clerk in a fast food restaurant when he was a student, and his wage was calculated hourly, but his working hours exceeded 30 hours per week (it's possible). In this case, he would honestly provide the above information in the 2010 survey, and this article would take this job as his first job. However, in the 2015 survey, he may not take this job as his first full-time job and would turn to the other job he obtained after graduation. As a result, an inconsistency forms.

A question here is which one is the real first job cared for by researchers of social mobility? Based on the definition set by Organization for Economic Co-operation and Development (OECD), the term full-time job is opposite of the term part-time job, these terms have no relation with the contract of employment, position in an organization or real work content. Nevertheless, take Status Attainment Model initiated by Blau and Duncan (1967) as an example, the first job they refer to means "the jobs obtained by subjects after they leave school". Therefore, in their model of path analysis, they set education as causes and first job as results. Accordingly, these two definitions are both reasonable, researchers should clarify which approach they want to adopt before questionnaire designing.

Backing to this article, as shown in table 14, when taking all samples into consideration, there are only 42% (for ISCO) and 67% (for EGP) of samples give same answers between two survey years. However, we can not conclude that it is a bias. It probably results from the questionnaire designing, which make the respondents identify two different jobs in two surveys. Accordingly, if we only take samples with the same beginning year of their first jobs into analysis (although there are exceptions, these cases have higher possibilities of referring to the same job), the percentages of consistency come to 61% (for ISCO) and 80% (for EGP), which are much more acceptable.

Table 15 presents the Spearman's rho of answers provided in two survey years. For all samples, the rho is only moderately correlated. In contrast, for samples with the same beginning year, the rho rises to 0.6, which is nearly highly correlated. Even though this degree of correlation

is not good enough, the bias is much smaller when comparing to the case of parental occupation.

This finding also implies that, even though recall bias has already threatened the quality of data, the problem would aggravation when combining with proxy bias. Therefore, it's better for researchers to interview the subjects themselves to acquire variables about their personal occupation. However, when accessing longitude data is difficult, and researchers seek to turn to recall data, the problem is relatively slighter.

Table 15: Spearman's Correlation of personal first job

	first job (measure in ISCO)		first job (measure in EGP)	
	all samples	samples with same begin year	all samples	samples with same begin year
Spearman's rho	0.3704^{***}	0.6515^{***}	0.3755^{***}	0.6305^{***}
$N=$	2,511	748	2,510	748

$^{*}p<.05$ $^{**}p<.01$ $^{**}p<.001$

Table 16 shows the pattern of inconsistency between 2010 answers and 2015 answers. Comparing the answers provided in 2015 to 2010 ones, the percentage of overestimation is higher than underestimation. This is, people tend to identify themselves have higher positions of their first jobs when they grow older. It can be unconsciously, which means since people are more familiar with their current work, they wrongly put some part of their current experiences into past memories. Besides, it can also be intentional, which means that people intend to reduce the gap between their first job positions and current positions and make a false report about their first job. However, no matter where these inconsistencies come from, they are systematical bias. Researchers should be careful, or they would probably overstatement the effect of first jobs on current occupations.

In conclusion, the inconsistency of personal first job information mainly results from different questionnaire designing in two survey years. After excluding this factor and selecting only samples referring to the same jobs, the inconsistency still exists, but far slighter. However, the pattern of this bias shows a systematical overestimation, which may come from intentional or unintentional correction of memories of respondents. Researchers should mind these bias, and infer the conclusion from their result carefully.

Table 16: the ratio of underestimation and overestimation of 2015 answers comparing to 2010 answers

	ISCO	EGP
Underestimate	14.57%	8.29%
Consistency	60.70%	79.68%
Overestimate	24.73%	12.03%
N=	748	748

4.2 Valid evaluation

The above results have already shown that proxy data acquired from children are inconsistent with self-report data provided by parents themselves (though in different degree). However, are the self-report data more reliable? Table 17 and table 18 are correlation matrix of variables obtained from of 2001 parental questionnaire and 2010 children survey. I convert the level of parental education into the number of years of education and recode occupation into SES according to the International Professional Social Status Scale (ISEI).

In theory, the level of education of an individual should be related to his socioeconomic status, and the educational level and occupation of a husband and wife should be related to each other. At the same time, the occupational socioeconomic status of this couple should predict family income. As shown in the tables, most of the correlation coefficient is higher in 2001 version than in 2010 version, meaning that parent-report data have a better construction validity.

Table 17: Correlation matrix of parental education, occupation and family income (2001 parent-report data)

		1	2	3	4	5
1	Father education	-				
2	Mother education	0.6779^{***}	-			
3	Father ISEI	0.5420^{***}	0.4352^{***}	-		
4	Mother ISEI	0.4785^{***}	0.5400^{***}	0.5326^{***}	-	
5	Family income	0.4174^{***}	0.3997^{***}	0.4255^{***}	0.4404^{***}	-

$^{*}p<.05$ $^{**}p<.01$ $^{**}p<.001$

On the other hand, Table 19 presents result of regression analysis of family income. The dependent variable is family income and the independent variables are parental education and ISEI. The result shows

that the R-squared is higher in 2001 version (R^2 = .2945) than in 2010 version (R^2 = .2675), but the difference is very little. This may be explained by the rough measurement of family income of the questionnaire, which only cut chosen items of income into 6 groups and has a class interval around 25 thousand to 50 thousand. This measurement cannot accurately reflect the variance between observations.

Table 18: Correlation matrix of parental education, occupation and family income (2010 children-proxy data)

		1	2	3	4	5
1	Father education	-				
2	Mother education	0.6538***	-			
3	Father ISEI	0.5045***	0.4345***	-		
4	Mother ISEI	0.4996***	0.6084***	0.5074***	-	
5	Family income	0.4025***	0.3854***	0.3905***	0.4100***	-

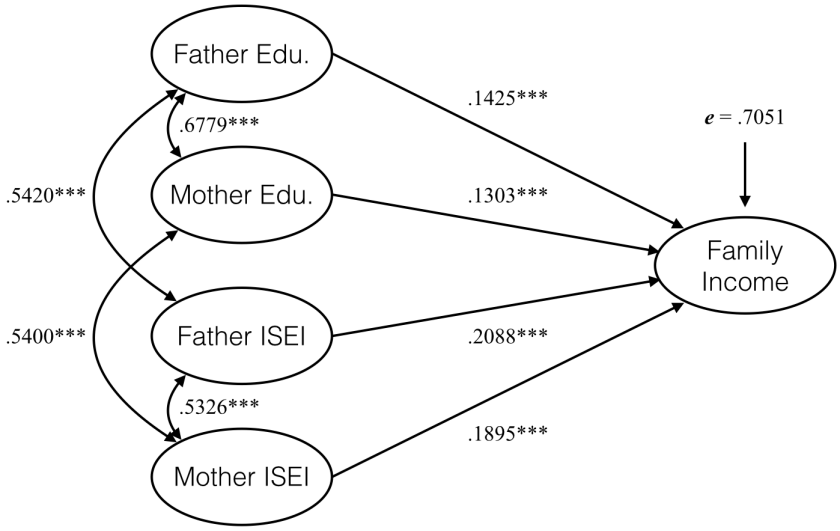
*p<.05 **p<.01 **p<.001

Table 19: result of regression analysis of family income

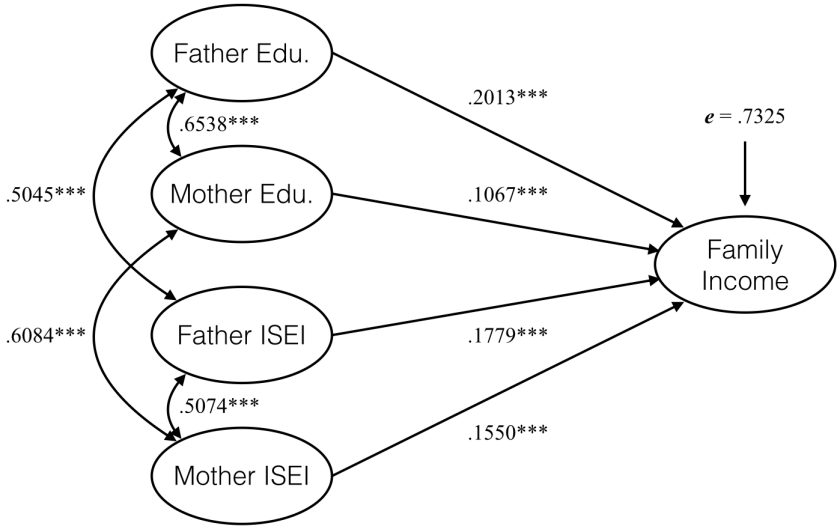
	Parent-report (2001)		Children proxy (2010)	
Variables	Coef.	β	Coef.	β
Father education	2,312.21***	0.1425	3,221.57***	0.2013
Mother education	2,339.88***	0.1303	1,901.23***	0.1067
Father ISEI	549.96***	0.2088	571.80***	0.1779
Mother ISEI	561.64***	0.1895	541.84***	0.1550
cons	-25,268.85		-35,144.21	
N=	2,002		2,199	
R^2	0.2949		0.2675	
adj. R^2	0.2934		0.2662	

*p<.05 **p<.01 **p<.001

Figure 2 graphs the above information into a figure of path analysis. Overall, the construction validity is more acceptable when using 2001 data, revealing that self-report data is more reliable than proxy data.



(a) 2001 parent-report data



(b) 2010 children proxy data

Figure 2: Valid evaluation model

5 Conclusion

By using Taiwan Education Panel Survey (TEPS) and its beyond (TEPS-B), this article merges three surveys initiated in different years and replied by different respondents, comparing these answers, and discussing the existence and seriousness of proxy bias and recall bias.

First, referring to parental education, it's shown that there is only a negligible inconsistency between self-report data and children proxy data. Over 85% of children respondents provide answers as same as their parents, and the answers of the two generations are highly correlated. This finding indicates that for simple variables, such as education,

proxy bias is slight, being in accord with past research (Krieger et al., 1998). However, I also find that children tend to confound their parental education more easy in the case that their parents belong to the group of "junior colleges or professional school", revealing that the bias doesn't originate from wild guesses of proxy respondents, but due to the fuzziness of the classification of the variable. Accordingly, this article suggests that it's acceptable to use proxy data when measuring simple variables, but researchers should double-check their design of chosen items, making sure there is no any confounding or misleading chosen item.

Secondly, referring to parental occupation, the proxy (by children) and recall (about 10 years) data possess a huge gap with parental self-report data. Only 30 to 40% of samples provide consistency answers with their parent, and the answers between tow generations are merely moderately correlated, existing an obvious bias. Moreover, it seems that children tend to systematically underestimate their parental occupation, seriously threatening internal validity of data. Besides, this article concludes that even if the answers are proxy by spouses, who are the closest family members, and without being requested any recall, the bias still exists (though slighter). Furthermore, the extent and pattern of bias vary from different measurements of occupation and different subjects. In brief, since the EGP scheme is simple than ISCO, there is smaller bias when using EGP scheme. Similarly, since female usually are engaged in occupations with clearly distinctive criteriums, there is little bias in maternal occupations than paternal ones.

In addition, when it comes to variables about the personal family background, the finding shows that parent-report data have a better construction validity than children proxy data, indicating that proxy and recall bias threatens the quality of data. Accordingly, it's better for researchers to get self-report, non-recall data. However, if it's necessary to use proxy or recall one, researchers should try to use simple classification to measure variables, allowing proxy respondents to provide answers based on some external information (such as employer or employee, title of position) but not internal privacy (such as real job content, possessing technic or not) of origin subjects.

Finally, referring to personal first jobs, this article finds that respondents will identify different jobs according to different designs of the questionnaires. Therefore, though panel surveys can avoid proxy and recall data, it should be noticed that the design of the questionnaire in the past can not always fit the current need. In short, there is a trade-off, and researchers should weight for both sides. On the other hand, if respondents identify the same first job in two survey years, it's shown that a slight bias still exists. The extent of this bias is reluctantly acceptable. However, the pattern of this bias reveals that respondents tend

to minimize the gap between their first job and current class position. This finding in accord with past research (Manzoni et al., 2010), which concludes that the recall bias aggravates in the case that respondents experience huge changes in their career development. Hence, this article suggests that when researchers are going to start a panel survey, they should design the questionnaire cautiously, making sure there is no confounding question. When researchers are planning to use variables acquired from an early survey, they should check the questionnaire and understand whether there is any definition need to be modified due to change of time. On the other hand, if researchers need to use recall data from cross-sectional surveys, though the extent bias is acceptable, there is a systematical overestimation of their first jobs and should be considered when doing inference.

To sum up, this article concludes that there are proxy bias and recall bias in research on social mobility and the bias seems non-random. Nevertheless, the extent and pattern vary from different variables, measurements, or characters of subjects. These findings reveal that proxy bias and recall bias threaten the internal validity of research on social mobility. Some suggestions for data collecting and inference from results are discussed.

References

Bassett, S. S., Magaziner, J., & Hebel, J. R. (1990). Reliability of proxy response on mental health indices for aged, community-dwelling women. *Psychology and Aging*, *5*(1), 127-132.

Blau, P. M., & Duncan, O. D. (1967). The american occupational structure.

Breen, R., & Jonsson, J. O. (1997). How reliable are studies of social mobility: an investigation into the consequences of errors in measuring social class. *Research in social stratification and mobility*, *15*, 91-114.

Broom, L., Jones, F. L., McDonnell, P., & Duncan-Jones, P. (1978). Is it true what they say about daddy? *American Journal of Sociology*, *84*(2), 417-426.

Choi, B. C. K. (2000). Bias, overview. In B. J. Gail MH (Ed.), *Encyclopedia of epidemiologic methods* (p. 74–82). Wiley: Chichester.

Connelly, R., Gayle, V., & Lambert, P. S. (2016). A review of occupation-based social classifications for social survey research. *Methodological Innovations*, *9*, 1-14.

Dassonneville, R. (2013). Questioning generational replacement. an age, period and cohort analysis of electoral volatility in the netherlands, 1971–2010. *Electoral Studies*, *32*(1), 37–47.

Delgado-Rodriguez, M., & Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health*, *58*(8), 635-641.

Erikson, R., & Goldthorpe, J. H. (1992). *The constant flux: a study of class mobility in industrial societies*. Oxford University Press, USA.

Ganzeboom, H. B., & Treiman, D. J. (2010). Occupational status measures for the new international standard classification of occupations isco-08; with a discussion of the new classification. In *Annual conference of international social survey programme, lisbon.*

Hassan, E. (2006). Recall bias can be a threat to retrospective and prospective research designs. *The Internet Journal of Epidemiology*, *3*(2), 339-412.

Himmelweit, H. T., Biberian, M. J., & Stockdale, J. (1978). Memory for past vote: implications of a study of bias in recall. *British Journal of Political Science*, *8*(3), 365-375.

Krieger, N., Okamoto, A., & Selby, J. V. (1998). Adult female twins' recall of childhood social class and father's education: a validation study for public health research. *American Journal of Epidemiology*, *147*(7), 704-708.

Manzoni, A., Vermunt, J. K., Luijkx, R., & Muffels, R. (2010). Memory bias in retrospectively collected employment careers: a model-based approach to correct for measurement error. *Sociological methodology*, *40*(1), 39-73.

Massagli, M. P., & Hauser, R. M. (1983). Response variability in self-and

proxy reports of paternal and filial socioeconomic characteristics. *American Journal of Sociology*, *89*(2), 420-431.

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and reconstructive surgery*, *126*(2), 619.

Perry, J., & Felce, D. (2002). Subjective and objective quality of life assessment: Responsiveness, response bias, and resident: proxy concordance. *Mental retardation*, *40*(6), 445-456.

Pless, C. E., & Pless, I. B. (1995). How well they remember: the accuracy of parent reports. *Archives of pediatrics & adolescent medicine*, *149*(5), 553-558.

Plewis, I., & Bartley, M. (2014). Intra-generational social mobility and educational qualifications. *Research in Social Stratification and Mobility*, *36*, 1-11.

Raphael, K. (1987). Recall bias: a proposal for assessment and control. *International journal of epidemiology*, *16*(2), 167-170.

Sackett, D. L. (1979). Bias in analytic research. In *The case-control study consensus and controversy* (pp. 51–63). Elsevier.

Tamborini, C. R., & Kim, C. (2013). Are proxy interviews associated with biased earnings reports? marital status and gender effects of proxy. *Social science research*, *42*(2), 499-512.