Survey mode effects on income inequality measurement
Supplementary Appendix

Abstract

This appendix includes material supplementing the manuscript "Survey mode effects on income inequality measurement." It includes background information in Section 1, an extended description of the methods used in Section 2 as well as robustness checks and additional results in Section 3.

JEL Classification: C14, C21, C81, C83, D31

Key Words: Income Inequality, Survey Methodology, Survey Modes, Distributional Decompositions

1 Background and further literature

1.1 Interview modes

There is a sizable literature providing evidence on the effect of interview modes on unit non-response, item non-response, and mode dependent bias in responses. Two hypotheses regarding the effect of interview modes, in particular, are discussed in this literature. The first is that modes might affect the degree of "social desirability bias," that is the degree to which responses conform to perceived norms (for instance regarding dentist visits, drug use, etc.) rather than the truth. The second is that modes might affect the degree of "satisficing," that is the degree to which easy/simple answers are given, for instance by answering "yes" to all in a series of questions. It should be emphasized that, in contrast to our paper, no reference we could find discusses surveys of income or wealth; instead these papers mostly focus on attitudinal questions. This is important in particular since income and wealth are complicated variables calculated based on a number of survey items.

Evidence on the impact of the interview mode on response rates and bias in responses is mixed. In a meta-study de Leeuw (1992) compared face-to-face interviews, telephone surveys, and self-administered mail questionnaires. By employing 67 mode-comparison-papers she showed that the main difference lies between self-administrated and interviewer-based survey modes, the differences between interviewer-based modes, and therefore also between CATI and CAPI seem less pronounced. While there was no significant difference in response validity and social desirability bias, face-to-face interviews lead to slightly less item non-response. CATI is the oldest of the computer assisted interviewing methods and is heavily used especially in market research. CAPI is widely used especially for complex surveys of governmental statistic agencies and universities (de Leeuw (2008)).

> "Because of the increased availability of other survey modes, face-to-face interviews are typically reserved for the most difficult and longest surveys that place the greatest burden on respondents. These are all kinds of surveys for which the other modes are not so likely to perform well. Face-to-face surveys also tend to be reserved for surveys that are most important to society, for which sponsors are willing to pay the cost." (de Leeuw et al. (2008), p. 164)

The main advantage of face-to-face interviews is that they are more flexible than telephone interviews. The interviewer can use response cards, visual scales etc. but also explain things better by being physically present which allows for a broader range of communication and interaction between the interviewer and the respondent. Via telephone the respondent can only rely on his/her memory when answering questions with multiple answer possibilities. Also interview length is an important determinant of the quality of the interviews given a certain mode. While CAPI and face to face in general is usually used for longer interviews (>30min) telephone interviewing is less suitable for longer interviews. Major data collection organizations in the US are refusing to conduct telephone interviews which are expected to last longer than 18 minutes (de Leeuw et al. (2008)).

1.2 Prior studies of mode effects

A series of papers by Klausch, Hox and Schouten (Klausch et al., 2013a,b, 2015) investigates the effect of survey modes in the context of the Dutch Crime Victimisation Survey. In this survey, modes were experimentally (randomly) assigned. Four different modes are considered: CAPI, CATI, self-administered paper based and web based. Focusing on attitudinal questions, they find that item non-response is significantly higher for self-administered interview modes and lowest for CAPI. It also appears that the sample of respondents is most representative of the population for CAPI, but almost as good for the web-based mode. Their results suggest sequential (mixed mode) designs might improve response rates and the representativity of respondents.

Jäckle et al. (2010) similarly studies the effect of interview modes using experimental mode assignment, this time in a survey conducted in 2003/2005 in Hungary. They argue "[...] that in order to evaluate whether mode affects data comparability it is necessary to move away from an assessment of means and marginal distributions toward an assessment of the effect of mode on relevant estimates" (p. 12). This is one of the key motivations for the analysis in the present paper, where we study the effect of interview modes on estimated income inequality.

Martin (2011) provides a general discussion of potential mixed and multiple modes of interviewing and its implications especially for international comparisons. In line with our findings, this paper argues that cross-country comparisons might be rendered problematic by uncontrollable selectivity issues associated with various modes.

An older contribution by Schwarz et al. (1991) provides some theoretical discussions comparing interview modes. They argue, in particular, that social desirability bias might be more pronounced in interviewer-based rather than self-administered modes. Holbrook et al. (2011) compare the effect of CATI and CAPI modes on the responses to attitudinal questions in several older surveys. They conclude that both satisficing and social desirability bias are significantly more pronounced in telephone (CATI) interviews. Savage and Waldman (2008) investigate the effect of survey mode on respondent learning and fatigue during repeated choice experiments. They find that lower cost survey modes are associated with larger measurement errors. Couper (2011) discuss a wide range of survey modes in the context of smaller surveys. In line with our conclusions, they emphasize the value of interviewer administered modes, which will in their opinion not become obsolete despite the emergence of new technologies. Vannieuwenhuyze et al. (2014) statistically model both selection and measurement effects in the context of mixed mode surveys. Their application focuses on the comparison between self-administered postal surveys and CAPI. They find only small effects of interview modes, with qualitative conclusions changing depending on modeling assumptions. Atkeson et al. (2014) compare CATI and self-administered (web and paper based) interviews in a survey on voting in New Mexico. They use matching on observable socio-demographic variables to compare modes. They find that both modes predict the election outcome equally well, while self-administered surveys are cheaper by 30% relative to CATI. A new textbook by Dillman et al. (2014), finally, provides a comprehensive introduction to the design of surveys using various modes.

2 Methods

This section includes details on section 3 on identification and estimation of the manuscript.

Influence functions of measures of inequality

We estimate the effect on several popular measures of income inequality, in particular the Gini-Coefficient and the poverty rate (the share of the household population with less than 60% of the median income), as well as the $P90/P10$-percentile ratio. For the RIF regression approach, we need to derive the influence functions of these measures.

The Recentered Influence Function (RIF) for the Gini-Coefficient (see e.g. Firpo et al. (2007) page 24f.) is given by

$$RIF(y, \nu^{GI}, F^0) = 1 + \frac{2}{\mu^2} R(F_y) y - \frac{2}{\mu} \left[y(1 - p(y)) + GL(p(y), F_y) \right] \tag{1}$$

where $R(F_y)$ is the integral of the Generalized Lorenz curve with the ordinates $GL(\bullet)$, and $p(y)$ is the density.

The poverty rate is defined as $\nu^{PR}(F) = F\left(0.6 \cdot F^{-1}(0.5)\right)$, where $med := F^{-1}(0.5)$ is the median for the distribution given by F and $0.6 \cdot F^{-1}(0.5) = 0.6 \cdot med$ is the poverty ceiling. The RIF for the poverty rate is equal to

$$RIF(y; \nu^{PR}, F^0) = \mathbf{1}(y \leq 0.6 \cdot med) - \frac{0.6 \cdot f(0.6 \cdot med)}{f(med)} \cdot \mathbf{1}(y \leq med). \tag{2}$$

The $P90/P10$-percentile ratio is defined as $\nu^{RA} = \frac{Q^{90}}{Q^{10}}$ where Q^i is the i's percentile of the income distribution. The RIF of the $P90/P10$-percentile ratio equals

$$RIF(y; \nu^{RA}, F^0) = \nu + \nu * \left[\frac{\mathbb{1}(y \leq Q^{10})}{Q^{10} * f(Q^{10})} - \frac{\mathbb{1}(y \leq Q^{90})}{Q^{90} * f(Q^{90})} \right]. \tag{3}$$

2.1 Estimation methods

There are numerous different ways proposed in the literature to estimate parameters identified under the conditional independence assumption $(Y^1, Y^0) \perp M|X$, good reviews are given by Imbens (2004) and Little and Rubin (2014). A classic reference for estimation using the propensity score is Rosenbaum and Rubin (1983).

We use the following alternative estimation approaches as robustness checks, all building on the same identifying assumption of conditional independence, to estimate the causal effect of interview modes on various outcomes. While we report coarsened exact matching, as our preferred method, in the manuscript we provide estimates using a fully interacted linear model as well as propensity score matching as robustness checks in this supplementary appendix.

- **Fully interacted linear model**:
  We first estimate a fully interacted linear model (FILM), which allows the effect of the mode to vary over all controls. To do so, we use *film*, a STATA program provided by Edwin Leuven and Barbara Sianesi, see `http://www.ifs.org.uk/publications/2712` [accessed on 19th of November 2012].

- **Propensity score matching**:
  We next estimate treatment effects based on propensity score matching (PSM).To do so, we use *psmatch2*, a STATA program provided by Edwin Leuven and Barbara Sianesi, see `http://www.ifs.org.uk/publications/2684` [accessed on 19th of November 2012].

While FILM allows the mode effects to vary over all controls, propensity score matching additionally allows (i) to impose common support based on the overlapping regions of the propensity scores (see Figure 1 in the manuscript) and (ii) does - due to its semi-parametric nature - not impose as strong linearity assumptions as logit and FILM are imposing. On the imposed common support we match the nearest - in terms of the propensity score - CATI-neighbour to every CAPI-household (1 to 1 matching) in order to balance the joint distribution of the covariates (controls).

- **Coarsened exact matching**:
  We finally use coarsened exact matching (CEM) as a further robustness check. Iacus et al. (2008) developed a method to temporarily coarse data based on ex-ante user choice and then run the analysis on the common support of the uncoarsened data. We use *cem*, a STATA program provided by Matthew Blackwell, Stefano Iacus, Gary King and Giuseppe Porro, see `http://ideas.repec.org/c/boc/bocode/s457127.html` [accessed on 19th of November 2012].

The support restrictions on covariates imposed by these alternative estimation methods differ. Since propensity score matching is not exact matching, it imposes weaker restrictions for covariate values to be included in the support than does exact matching, or coarsened exact matching.

Reweighting procedure for CEM

As matching variables we use a subset of our household and personal level covariates. For categorical variables (household size and being female as the household head) we impose an exact matching strategy and for the two continuous variables we use a coarsening strategy for matching on certain parts of the distributions by imposing cut-points (household level: household disposable income 20000, 30000, 40000, 50000, 60000; and personal level: age 30, 40, 50, 60, 70). This matching strategy leads to a perfectly balanced dataset in terms of the joint distribution of the categorical variables. As household disposable income is allowed to be matched in approximately 10,000 Euro brackets (and below 20,000 Euro as well as over 60,000 Euro) and age is allowed to be matched in about 10 year age brackets (and below 30 as well as older than 70) some imbalances remain with respect to those two variables. Out of the 343 covariate combinations defined we find 224 which define the common support, i.e. where at least one CAPI and one CATI observation can be found. In terms of the sample the total of 3,377 observations collapses to 3,190 observations which lie inside the common support. 63 CATI- and 124 CAPI-observations lie outside the common support. The weights for the matched observations are chosen in a way that CAPI and CATI observations are balanced for each covariate combination. To control for the remaining imbalances in the matched dataset we still use age and household disposable income as controls when we calculate the treatment effects.

3 Results

3.1 Preliminaries: (Self-) Selection into interview modes, and non-response

In order to model selection into interview modes we run a logit regression of interview mode (CATI=0, CAPI=1) on the set of available controls. We also control for regional dummies; the coefficients on these are not significantly different from 0. Table 1 shows the average marginal effects calculated from this model.

We estimate logit regressions of HINR (household item non-response) on interview mode and household-level as well as person-level controls, and a similar regression using household income as dependent variable (see tables 2 and 3). As the results in table 2 show, the fully interacted linear model as well as propensity score matching lead to qualitatively and quantitatively similar results as the logit model using all controls as well as reweighting based on CEM.

Table 1: Logit-regression of the (Self-)Selection of the Mode on Control Variables

Selection towards CAPI

	Mode Selection
household characteristics	
Household size	0.031
	(0.012)
Household with kids	-0.041
	(0.028)
Single family home	0.023
	(0.022)
Owner occupier	-0.028
	(0.021)
Living space in sqm	-0.001
	(0.000)
Land line	-0.178
	(0.020)
Mobile phone	-0.097
	(0.029)
Log-disposable income	-0.052
	(0.018)
Personal characteristics of household head	
Female	-0.072
	(0.018)
Self-employed	0.009
	(0.037)
Jobless	0.079
	(0.037)
Weekly working hours	0.002
	(0.001)
Married and living together	-0.068
	(0.022)
Education: apprenticeship	-0.090
	(0.023)
Education: higher sec. school	-0.191
	(0.030)
Education: university	-0.131
	(0.036)
Age	-0.001
	(0.001)
N	3376

Notes:
(i)This table shows average marginal effects (AME) of household characteristics of the 2007 EU-SILC wave on being interviewed by CAPI in the 2008 EU-SILC wave. All average marginal effects are calculated from a logistic regression using an CATI(0)/CAPI(1) as dependent variable. Furthermore we controlled for federal states and regional population density. The coefficents of both are not significant.
(ii) Standard errors calculated by the delta method are given in parentheses.
(iii) *Source:* EU-SILC 07/08.

Table 2: Interview Mode Effect on Income Item Non-Response

	I	*II*	*III*
Average effect of CAPI on item non-response			
Logit Model	-0.088	-0.075	-0.071
	(0.014)	(0.014)	(0.015)
Fully Interacted Model			-0.076
			(0.016)
Propensity Score Matching			-0.072
			(0.022)
Coarsened Exact Matching			-0.068
			(0.014)
N	3291	3290	
Household Controls		yes	yes
Personal Controls			yes

Notes:
(i)This table shows average partial effects (APE) of being interviewed by CAPI on household income item non-response. Results are reported from a logistic (using an item non-response dummy for household income [at least one item non-response in an income question] as dependent variable) as well as a fully interacted model and various matching techniques.
(ii) Standard errors are given in parentheses (for the standard errors of the marginal effects that delta method is applied).
(iii) *Source:* EU-SILC 07/08.

As the results in table 3 show, the fully interacted linear model as well as propensity score matching lead to qualitatively similar results as CEM and the OLS specification using all controls. While the Fully interacted model is also quantitatively similar to OLS, CEM leads to a somewhat larger negative effect, which is likely due to the fact that CEM does not extrapolate outside the common support of the cells used. Propensity score matching leads to somewhat smaller negative effect than the other methods. However, they are not statistically significant from each other. This points towards the use of examining the effect across the full distribution (see below and chapter 4.3 of the manuscript) instead of only the average.

Table 3: Interview Mode Effect on Household Income

	I	*II*	*III*
Average effect of CAPI on log household income			
OLS-Regression	-0.205	-0.052	-0.040
	(0.022)	(0.016)	(0.015)
Fully Interacted Model			-0.043
			(0.016)
Propensity Score Matching			-0.017
			(0.032)
Coarsened Exact Matching			-0.060
			(0.017)
N	3377	3376	
Household Controls		yes	yes
Personal Controls			yes

Notes:
(i) This table shows the effect (regression coefficient, as well as matching estimators) of being interviewed by CAPI on the logarithm of household disposable income.
(ii) Standard are given in parentheses.
(iii) *Source:* EU-SILC 07/08.

Differences

While the manuscript deals with the estimation of the total (causal) effect purged of selection bias, we compare inequality measures across the samples of households interviewed with CAPI and CATI here, i.e. observed difference, including selection bias. Note, however, that this will still underestimate the true total effect as we cannot include unit-non-response bias (see section 2.3 of the manuscript). We provide standard errors of this difference using a bootstrap with 1000 replicates. Appropriate sampling weights are used in the calculation of these statistics. For the estimations the whole procedure is bootstrapped in the sense that a bootstrap sub-sample is drawn, and then the difference of the statistic calculated. With these 1000 estimates of the differences we are able to estimate standard errors. We apply this procedure to the Gini-coefficient; the poverty rate, i.e. the proportion of population with lower income than 60% of the median; and the 90/10 percentile ratio. Additionally, we provide the estimates of the difference between the CAPI- and the CATI sample statistics (i) without any adjustments and (ii) on the common support resulting from the CEM-Procedure, which controls partly for selection bias but at the cost of a reduction of the sample.

Table 4 reports first the measures of the Gini-Coefficient, the Poverty Rate, and the 90/10 percentile ratio for both the CAPI and CATI sub-samples in EU-SILC 2008. We find that the difference of the Gini-Coefficient is 0.026 for the full sample and 0.024 using only the common support sample implying a 8.1% and 7.2% lower Gini-coefficient using CATI instead of CAPI. While the estimate of the whole sample is significant at the 5% level, the difference for the balanced sample is only significant at the 10% level. To check for the differences at the top we also used a General Entropy Class index with $\alpha = 2$ which shows huge differences (0.18 for the CATI versus 0.49 for the CAPI common support samples). This is due to the fact that it is very sensitive to top income and as we saw the highest income observations are all from CAPI interviews. This sensitivity, however, also renders a high variability of replicates in this bootstrapping procedure and thus generates high standard errors yielding insignificant results. Table 4 additionally reports a significantly higher poverty rate for the CAPI sample (i.e. 3.9 percentage points in the whole and 3.1 percentage points in the balanced sample) and an (statistically) insignificant difference of the 90/10-percentile ratio between the two subsamples. In the manuscript, we employ RIF-regressions controlling for covariates to estimate the causal (or measurement) effect of interview mode on the distributional statistics of income.

Table 4: Differences of Inequality Measures between Interview Modes

	Gini Coeff.		Poverty rate		90/10 Percentile Ratio	
	Ia	*IIa*	*Ib*	*IIb*	*Ic*	*IIc*
Inequality Measure CAPI	0.3341	0.3332	0.2280	0.2226	4.801	4.791
Inequality Measure CATI	0.3076	0.3096	0.1885	0.1920	4.523	4.527
Difference (Bootstrap)	0.026	0.024	0.039	0.031	0.278	0.263
Bootstrapped Std. Err.	(0.012)	(0.013)	(0.014)	(0.013)	(0.218)	(0.267)

Notes:
(i) This table shows the effect of the interview mode on aggregate measures of inequality. We report the inequality statistic (Gini Coefficient, the Poverty Rate, and the Percentile Ratio), and the difference between the sub-samples.
(ii) Standard errors are reported using bootstrapping method.
(iii) Columns *Ia-c* show the results using the full (panel) sample, i.e. 3.377 observation; and columns *IIa-c* only use the matched observation from the above coarsened exact matching procedure.
(iv) *Source:* EU-SILC 07/08.

References

Atkeson, L. R., Adams, A. N., and Alvarez, R. M. (2014). Nonresponse and mode effects in self- and interviewer-administered surveys. *Political Analysis*.

Couper, M. P. (2011). The Future of Modes of Data Collection. *Public Opinion Quarterly*, 75(5):889–908.

de Leeuw, E. D. (1992). *Data Quality in Mail, Telephone and Face to Face Surveys*. TT-Publikaties, Amsterdam, Amsterdam.

de Leeuw, E. D. (2008). The effect of computer-assisted interviewing on data quality: A review of the evidence. Methodika/department of methodology and statistics, utrecht university.

de Leeuw, E. D., Hox, J. J., and Dillman, D. A. (2008). *International Handbook of Survey Methodology*. Psychology Press, Taylor & Francis, New York, European Association of Methodology.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons.

DiNardo, J., Fortin, N., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64:1001–1044.

Firpo, S., Fortin, N., and Lemieux, T. (2007). Decomposing wage distributions using recentered influence function regressions. Unpublished working paper, University of British Columbia.

Holbrook, A. L., Green, M. C., and Krosnick, J. A. (2011). Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias. *Public Opinion Quarterly*, 67(1):79–125.

Iacus, S. M., King, G., and Porro, G. (2008). Matching for causal inference without balance checking. Working paper series, Harvard.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29.

Jäckle, A., Roberts, C., and Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, 78(1):3–20.

Klausch, T., Hox, J. J., and Schouten, B. (2013a). Assessing the mode-dependency of sample selectivity across the survey response process. Technical Report 2013-03, Statistics Netherlands.

Klausch, T., Hox, J. J., and Schouten, B. (2013b). Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions. *Sociological Methods & Research*, 42(3):227–263.

Klausch, T., Hox, J. J., and Schouten, B. (2015). Selection error in single- and mixed mode surveys of the dutch general population. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data.* John Wiley & Sons.

Martin, P. (2011). A Good Mix? Mixed Mode Data Collection and Cross-national Surveys. *Ask: Research & Methods*, 20(1):5–26.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62:1349–1382.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Savage, S. J. and Waldman, D. M. (2008). Learning and fatigue during choice experiments: a comparison of online and mail survey modes. *Journal of Applied Econometrics*, 23(3):351–371.

Schwarz, N., Strack, F., Hippler, H.-J., and Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, 5:193–212.

Vannieuwenhuyze, J. T., Loosveldt, G., and Molenberghs, G. (2014). Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models. *Journal of Official Statistics*, 30(1):1–21.