Modelling with Discretized Ordered Choice Covariates

Felix Chan*, Laszlo Matyas** and Agoston Reguly**

January 7, 2020

* Curtin University, Perth
** Central European University, Budapest

Abstract

This paper deals with econometric models where some (or all) explanatory variables (or covariates) are observed as discretized ordered choices. Such variables are in theory continuous, but in this form are not observed at all, their distribution is unknown, and instead, only a set of discrete choices are observed. We explore how such variables influence inference, more precisely, we show that this leads to a very special form of measurement error, and consequently to endogeneity bias. We then propose appropriate sub-sampling and instrumental variables (IV) estimation methods to deal with the problem.

JEL: C01, C13, C21, C25, C83

Keywords: Discretized variable, measurement error, sampling, survey methods.

1 Introduction

There is an increasing number of survey-based large data sets where many (sometimes all) variables are observed through the window of individual choices, i.e., by picking one option from a pre-set class list, while the original variables themselves are in fact continuous. For example, in transportation modelling, the US Federal Transportation Office creates surveys to measure different transportation behaviours. This practice is also common for major cities like London, Sydney and Hong Kong. Usually the reported values are a discretized version of variables, like average personal distance travelled, or use of public or private transportation (Santos et al., 2011). Also in transportation research, the use of Likert-scale type data on intentions or attitudes is quite common, such as data from question on the likelihood of utilizing a certain transportation mode (see Heath and Gifford, 2002). In happiness economics, variables are also often measured with Likert-scale data (see Frey and Stutzer (2002) and Stutzer (2004)). Such examples are also common in many other areas, like credit ratings in financial economics, corruption measures or institutional development in political economy. These are discrete variables which have the characteristics of ordered choices (see Mauro (1995), Méndez and Sepúlveda (2006), Knack and Keefer (1995) and Acemoglu et al. (2002)). Typically, such variables are related to income, expenditure on something over a period of time, willingness to take some action (e.g., how much would you be willing to pay for ... ?) or questions about likelihood(s) (e.g., how likely would you be to download this application ... ?) and questions related to time (e.g., how much time did you spend commuting last week ... ?). The main question we investigate in this paper is how this affects inference in an econometric model when such variables are used as explanatory variables or covariates.

Consider the random variable $x_i \sim f(0, 100)$, where $f(a_l, a_u)$ denotes[1] a distribution with support in $[a_l, a_u]$ with mean μ for $i = 1, \ldots, N$. Also, quite importantly, the distribution $f(\cdot)$ is unknown (and can be continuous or discrete). Furthermore, define

$$x_i^* = \begin{cases} z_1 & \text{if } c_0 \le x_i < c_1 \quad \text{or} \quad x_i \in C_1 = [c_0, c_1) \quad \text{1st choice,} \\ z_2 & \text{if } c_1 \le x_i < c_2 \quad \text{or} \quad x_i \in C_2 = [c_1, c_2), \\ \vdots & \vdots \\ z_m & \text{if } c_{m-1} \le x_i < c_m \quad \text{or} \quad x_i \in C_m = [c_{m-1}, c_m), \\ \vdots & \vdots \\ z_M & \text{if } c_{M-1} \le x_i \le c_M \quad \text{or} \quad x_i \in C_M = [c_{M-1}, c_M], \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{last choice.} \end{cases} \tag{1}$$

We refer to variable z_m as the choice values ($m = 1, \ldots, M$). It can be a measure of centrality of the given choice class, or can be a completely arbitrarily assigned value (say, for example, if we consider preferences). The class boundary c_m can be known, unknown or in some cases stochastic. The main difficulty is that instead of x_i we only observe x_i^*. Variable x is in fact observed through the discrete ordered window of x_i^*.

[1] A complete list of the notations used in the paper can be found in Appendix B.

2 Basic Setup

Let us assume that we have a simple linear econometric model of the form

$$y_i = w_i'\gamma + x_i^{*'}\beta + \varepsilon_i, \tag{2}$$

with the true Data Generating Process (DGP) being

$$y_i = w_i'\gamma + x_i'\beta + u_i, \tag{3}$$

where $i = 1, \ldots, N$, w is a set of 'usual' explanatory variables, x^* is a set of discretized choice variables as defined in (1), γ and β are unknown parameters and u_i is an idiosyncratic disturbance term for model (3) with ε_i being its perceived counterpart in model (2). We also maintain the independence of observations across individuals assumption. The main question is therefore how estimating model (2) differs from estimating model (3).

Remark: If $f(\cdot)$ is known, assuming known boundaries, the expected value of each variable in x^* is also known in each class and has an unbiased/consistent estimate, then the LS estimator of model (2) is unbiased/consistent. This is in fact the Berkson model (see Berkson (1980) and Wansbeek and Meijer (2000) pp. 29-30).

2.1 An Example

Let us assume that we would like to model in a given city the factors explaining individual transport expenditures (TE), in a given period of time with the simple model

$$TE_i = w_i'\gamma + \beta\, UPT_i + \varepsilon_i\,, \tag{4}$$

where TE_i is the transport expenditure for individual i, w_i are 'usual' controls and UPT_i is the use of public transport in commuting measured in percentage points. 100% if only PT was used and 0% if PT was not used at all for individual i ($i = 1, \ldots, N$). Now UPT is not observed; instead we observe only the individual's choice from a pre-set list UPT^*. We ask the use of public transport in the following way

$$\begin{array}{ll} 1 \rightarrow & \text{took almost only public transport,} \\ 2 \rightarrow & \text{took mostly public transport,} \\ 3 \rightarrow & \text{mostly did not take public transport,} \\ 4 \rightarrow & \text{almost did not take public transport,} \end{array} \tag{5}$$

which is referred to the following intervals

$$UPT_i^* = \begin{cases} 1, & \text{if} \quad 90\% \leq UPT_i \leq 100\%, \\ 2, & \text{if} \quad 50\% \leq UPT_i < 90\%, \\ 3, & \text{if} \quad 10\% \leq UPT_i < 50\%, \\ 4, & \text{if} \quad 0\% \leq UPT_i < 10\%. \end{cases} \tag{6}$$

We can code the responses as the mid-value of each class to UPT_i^* such that

$$UPT_i^* = \begin{cases} 0.95 \rightarrow \text{took almost only public transport,} & \text{if} \quad 90\% \leq UPT_i \leq 100\%, \\ 0.70 \rightarrow \text{took mostly public transport,} & \text{if} \quad 50\% \leq UPT_i < 90\%, \\ 0.30 \rightarrow \text{mostly did not take public transport,} & \text{if} \quad 10\% \leq UPT_i < 50\%, \\ 0.05 \rightarrow \text{almost did not take public transport,} & \text{if} \quad 0\% \leq UPT_i < 10\%. \end{cases} \tag{7}$$

Obviously, we can use many possible representations for the responses. Using the mid-values seems to be reasonable when the only available information is that an observation is in a given class.

2.2 Related Work

To the best of our knowledge, there has been no study investigating the estimation of discretized choice variable(s) when the categories/classes are not represented by the expected values of the underlying distribution(s). There has been though some work done on related issues. Taylor and Yu (2002) consider a regression model with three multivariate normal random variables. The first is linearly dependent on the second. Then they dichotomize this second one and include into the model another variable as well, and derive the asymptotic bias for its parameter. However, they do not connect this to the bias in the parameter of the other variable(s). Lagakos (1988) analyses the correct cut values for the grouping of continuous explanatory variables. He derives a test on deviating from the expected group mean and the categorized value if the group mean is known. He refers to this solution as the optimization criterion for discretizing an explanatory variable, using the argument in Connor (1972).
There are many papers considering the discretization of a continuous variable, but all assume that the choice values properly represent each class. In these papers, the main question is the effect of discretization in terms of efficiency loss (see, for example, Cox (1957), Cohen (1983), Johnson and Creech (1983)).
The measurement error literature has not considered the problem in details either, as it has been assumed that the class choice values are taking the expected values of the known underlying distribution (Wansbeek and Meijer (2001)), or the measurement error is on top of a categorized variable (Buonaccorsi (2010)).

3 Some Theory: Bias of the OLS Estimator

Let us assume for the sake of simplicity that there is only one explanatory variable in the model which is observed through discretized choices. It is also assumed, as said earlier, that it has a known support $[a_l, a_u]$ with known boundaries (C_m), and let z_m from Equation (1) be the class midpoint.[2]
The classes are now the following with their respective class values:

$$\begin{aligned}
&C_1 = \left[a_l, a_l + \frac{a_u - a_l}{M}\right) && z_1 = a_l + \frac{a_u - a_l}{2M},\\
&\vdots\\
&C_m = \left[a_l + (m-1)\frac{(a_u - a_l)}{M}, a_l + m\frac{a_u - a_l}{M}\right) && z_m = a_l + (2m-1)\frac{a_u - a_l}{2M},\\
&\vdots\\
&C_M = \left[a_l + (M-1)\frac{(a_u - a_l)}{M}, a_l + M\frac{a_u - a_l}{M}\right] && z_M = a_l + (2M-1)\frac{a_u - a_l}{2M}.
\end{aligned} \tag{8}$$

[2]In the special case of the uniform distribution, the midpoints coincide with the conditional expectation of the uniformly distributed explanatory variable x in that class.

Let N_m be the number of observations in each class C_m, that is $N_m = \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}$, where $\mathbf{1}_{\{x \in C\}}$ denotes the indicator function defined as

$$\mathbf{1}_{\{x \in C\}} := \begin{cases} 1, & \text{if } x \in C, \\ 0, & \text{if } x \notin C. \end{cases}$$

When x has a cummulative distribution cdf $F(\cdot)$,

$$\begin{aligned} \mathbb{E}(N_m) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}\right) \\ &= N \int_{C_m} f(x)\,\mathrm{d}x \\ &= N \Pr(c_{m-1} < x \leq c_m), \end{aligned}$$

using the independence assumption. When, for example, x has a uniform distribution, we have $\mathbb{E}(N_m) = N/M$ for all $m = 1, \ldots, M$.

The standard OLS estimation is given by

$$\begin{aligned} \hat{\beta}^*_{OLS} &= (x^{*'} x^*)^{-1} (x^{*'} y) \\ &= \frac{\sum_{m=1}^{M} z_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} (\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2} \\ &= \frac{\sum_{m=1}^{M} \left[a_l + (2m-1)\frac{a_u - a_l}{2M}\right] \left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} (\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m \left[a_l + (2m-1)\frac{a_u - a_l}{2M}\right]^2}. \end{aligned} \tag{9}$$

Using Equation (9), we can get the following general formula for the expected value of the OLS estimator

$$\mathbb{E}\left(\hat{\beta}^*_{OLS}\right) = \beta + \beta \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m N_m \upsilon^m}{\sum_{m=1}^{M} N_m z_m^2}\right\}, \tag{10}$$

where a respondent makes an error $\xi_i = x_i - x_i^*$ for each observation by setting the possible answer values at x_i^*. The last but one assertion in Equation (10) is based on the disturbance term u_i being independent of regressor x_i and $\mathbb{E}(u_i) = 0$ for all $i = 1, \ldots, N$. The last inference uses the fact that the errors ξ_i have the same conditional distribution over the class C_m, $\upsilon^m \overset{\mathrm{d}}{=} \xi_i | C_m$ for all $m = 1, \ldots, M$ and $i = 1, \ldots, N$. Importantly, the second term in expression (10) does not vanish in general, since $\upsilon^m | C_m$ is not independent of $N_m | C_m$, $\upsilon^m | C_m \not\perp\!\!\!\perp N_m | C_m$ nor $\mathbb{E}(\xi_i | C_m) = \mathbb{E}(\upsilon^m) = 0$ (see Figure 1, right panel). These would be sufficient assumptions for the OLS to be unbiased. The former issue can be eliminated by conditioning on the underlying distribution of x_i. Conditional on the distribution x_i and the class C_m, the number of observations in the class and assuming that the errors are independent of each other, $N_m | x_i, C_m \perp\!\!\!\perp \upsilon^m | x_i, C_m$, but knowing the underlying distribution makes the problem trivial. Nonetheless, because of both issues, the 'naive' OLS estimator is biased.

The uniform distribution, however, turns out to be a special case. Let us assume that $x_i \sim U(a_l, a_u)$ for all $i = 1, \ldots, N$, then both of the above disappear (see the left panel in Figure 1) if we are using the class mid points. The first problem is resolved, because in the case of the uniform distribution, both the number of observations N_m in each class C_m and the error

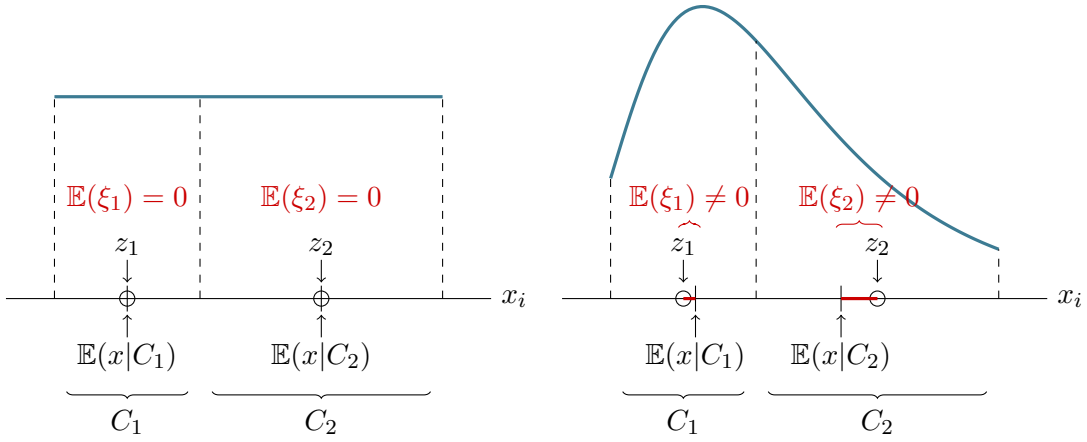

Figure 1: The difference between uniform (left panel) and general distributions (right panel)

term v^m are independent of the regressor's x_i distribution, while the second problem does not appear trivially, since now the class midpoints are proper estimates of the regressor's x_i expected value in the class C_m. From Equation (10), we obtain that

$$\mathbb{E}\left(\hat{\beta}^*_{OLS}\right) = \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m N_m v^m}{\sum_{m=1}^{M} N_m z_m^2}\right\} = \beta,$$

where v^m is a uniformly distributed random variable with zero expected value, $\mathbb{E}(v^m) = 0$ for all $m = 1, \ldots, M$. Hence, in the case of uniform distribution, unlike for other distributions, the OLS is unbiased.

Now, let us return to Equation (9). However, instead of taking the expectation, let us see what happens in the probability limit, when the sample size or the number of classes goes to infinity.

3.1 N (in)consistency

First, assume that $\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i = 0$, in other words that the choice set selection is independent of the disturbance terms, and also that with sample size N the number of classes M is fixed. Then

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}^*_{OLS} = \frac{\beta \sum_{m=1}^{M} z_m \left[\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m}. \tag{11}$$

Define $x^m = \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i$, then x^m sums the truncated version of the original random variables x_i on the class C_m, $x_m \overset{\mathrm{d}}{=} x_i|C_m$, for all $m = 1, \ldots, M$, therefore its asymptotic distribution can be calculated by applying the Lindeberg-Levy Central Limit Theorem,

$$x^m/N_m \overset{\mathrm{a}}{\sim} N\big(\mathbb{E}(x_m), \mathrm{V}(x_m)/N_m\big).$$

The $\hat{\beta}^*_{OLS}$ estimator is consistent if and only if the probability limit in Equation (11) equals β. To give a condition for consistency, first we rewrite the previous Equation (11) in terms

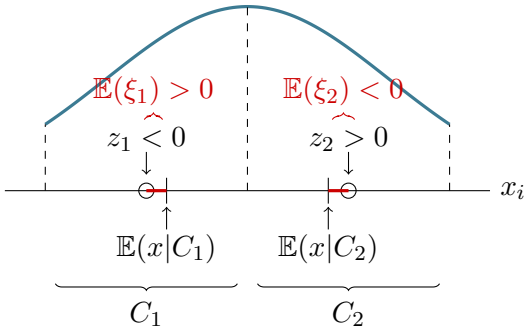

Figure 2: The estimator is inconsistent even in case of symmetric distributions (see Equation (13)).

of the error terms ξ_i,

$$\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^*_{OLS}-\beta\right)=\frac{\beta\sum_{m=1}^{M}z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}\xi_i\right]}{\sum_{m=1}^{M}z_m^2\operatorname{plim}_{N\to\infty}N_m}, \tag{12}$$

where the asymptotic distribution of the sum of errors in class C_m, $\xi^m=\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}\xi_i$, $m=1,\ldots,M$, can be given by

$$\xi^m/N_m \overset{\mathrm{d}}{=} x^m/N_m - z_m \overset{\mathrm{a}}{\sim} N\big(\mathbb{E}(x^m)-z_m, \mathrm{V}(x^m)/N_m\big).$$

After substituting back to the expression (12), we get

$$\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^*_{OLS}-\beta\right)=\frac{\beta\sum_{m=1}^{M}z_m\left\{\mathbb{E}(x_m)-z_m\right\}}{\sum_{m=1}^{M}z_m^2}. \tag{13}$$

Therefore, to obtain the (in)consistency of the OLS estimator $\hat{\beta}^*_{OLS}$ in the number of observations N, we only need to calculate the expected value of the truncated random variable x_m, $m=1,\ldots,M$ and check whether the expression (13) equals 0 to satisfy a sufficient condition. Let us apply these results to the uniform distribution. In this case, there is no consistency issue because the class midpoints coincide with the expected value of the truncated uniform random variable in each class, making the expression (13) zero, hence the OLS estimator is consistent.

Note that the consistency of the OLS estimator is not guaranteed even in the case of symmetric distributions and symmetric class boundaries. After appropriate transformations (e.g., demeaning), it can be seen that the sign of the differences between the expectation of the truncated random variables x_m and the class midpoints is opposite to the sign of the class midpoints on either side of the distribution, which implies negative overall asymptotic bias in N (see Figure 2).

In the case of a (truncated) normal variable, for example, we need to substitute the expected value of the truncated normal random variable x_m for each $m=1,\ldots,M$ in the consistency formula (13). As a result, the difference between the expectation and the class midpoints in general is not zero for all m, hence the formula cannot be made arbitrarily small. Therefore, the OLS estimator becomes inconsistent in N (see the size of the bias based on simulation results in Appendix A).

So far we have focused on the estimation of β in Equation (2). But how about γ? It can be shown that the bias and inconsistency presented above is contagious. Estimation of all parameters of a model is going to be biased and inconsistent unless the measurement error and x are orthogonal (independent), which is quite unlikely in practice. This is important to emphasize: a single choice type variable in a model is going to infect the estimation of all variables of the model.

3.2 M Consistency

Let us see next the case when N is fixed but $M \to \infty$. Now, we may have some classes that do not contain any observations, while others still do. Omitting, however, empty classes does not cause any bias because of our iid assumption. Furthermore, while we increase the number of classes, the size of the classes itself is likely to shrink and become so narrow that only one observation can fall into each. In the limit we are going to hit the observations with the class boundaries. To see that, we derive the consistency formula in the number of classes M assuming that $\operatorname{plim}_{M\to\infty} \sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m u_{i_m} = 0$, or with re-indexation $\operatorname{plim}_{M\to\infty} \sum_{i=1}^{N} z_{m_i} u_i = \sum_{i=1}^{N} x_i u_i = 0$, which should hold in the sample and is a stronger assumption than the usual $\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} x_i u_i = 0$:

$$
\begin{aligned}
\operatorname*{plim}_{M\to\infty} \left(\hat{\beta}^*_{OLS} - \beta\right) &= \operatorname*{plim}_{M\to\infty} \frac{\sum_{m=1}^{M} z_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i\in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2} - \beta \\
&= \operatorname*{plim}_{M\to\infty} \frac{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i\in C_m\}}(\beta x_i + u_i)\right]}{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} N_m z_m^2} - \beta \\
&= \operatorname*{plim}_{M\to\infty} \frac{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m (\beta x_{i_m} + u_{i_m})}{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m^2} - \beta \\
&= \operatorname*{plim}_{M\to\infty} \beta \left\{ \frac{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m x_{i_m}}{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m^2} - 1 \right\} \\
&= \operatorname*{plim}_{M\to\infty} \beta \left\{ \frac{\sum_{i=1}^{N} z_{m_i} x_i}{\sum_{i=1}^{N} z_{m_i}^2} - 1 \right\} \\
&= \beta \left\{ \frac{\sum_{i=1}^{N} \operatorname{plim}_{M\to\infty} z_{m_i} x_i}{\sum_{i=1}^{N} \operatorname{plim}_{M\to\infty} z_{m_i}^2} - 1 \right\} \\
&= \beta \left\{ \frac{\sum_{i=1}^{N} x_i x_i}{\sum_{i=1}^{N} {x_i}^2} - 1 \right\} \\
&= 0,
\end{aligned}
$$

where the index $i_m \in \{1,\ldots,N\}$ denotes observation i in class m (at the beginning there might be several observations that belong to the same class m), and index $m_i \in \{1,\ldots,M\}$ denotes the class m that contains observation i (at the and of the derivation one class m includes only one observation i). Note that the derivation does not depend on the distribution of the explanatory variable x, so consistency in the number of classes M holds in general. Let us also note, however, that this convergence in M is slow. Also, as $M \to \infty$, the class sizes go to zero, and the smaller the class sizes the smaller the bias. Of course, in practice, the number of classes M cannot be too large due to the limits of our cognitive capacities.

Typically, the optimal number of choices for a survey is relatively small, $M = 3, 5, 7$ or at most $M = 10$.[3] Section 5 proposes novel sub-sampling schemes that allow one to construct a *working sample* for estimation purposes with arbitrary large M, while keeping the number of classes in each survey small (and fixed). This will then allow us to invoke this result to obtain a consistent estimator.

3.3 Some Remarks

The above results hold for much simpler cases as well. If instead of model (2) we just take the simple sample average of x, $\bar{x} = \sum_i x_i/N$, then $\bar{x}^* = \sum_i x_i^*/N$ is going to be a biased and inconsistent estimator of $\bar{x}$.

The measurement error due to discretized choice variables, however, not only induces correlation between the error terms and the observed variables, but it also induces a non-zero expected value for the disturbance terms of the regression in (2). Consider a simple example where there is an unobserved variable x_i with an observed discretized choice version:

$$x_i^* = \begin{cases} z_1 & \text{if } c_0 \leq x_i < c_1, \\ z_2 & \text{if } c_1 \leq x_i < c_2, \end{cases} \tag{14}$$

and

$$y_i = x_i\beta + \varepsilon_i. \tag{15}$$

Using the discretized choice variable means:

$$y_i = x_i^*\beta + (x_i - x_i^*)\beta + u_i \tag{16}$$

and $\mathbb{E}\left[x_i - x_i^*\right]$ is

$$\begin{aligned} \mathbb{E}\left[x_i - x_i^*\right] =& \mathbb{E}(x_i) - \mathbb{E}(x_i^*) \\ =& \mathbb{E}(x_i) - \mathbb{E}\left[z_1\mathbf{1}(c_0 \leq x_i < c_1) + z_2\mathbf{1}(c_1 \leq x_i < c_2)\right] \\ =& \mathbb{E}(x_i) - z_1 \Pr(c_0 \leq x_i < c_1) - z_2 \Pr(c_1 \leq x_i < c_2). \end{aligned}$$

The last line above is not zero in general. Thus, it would induce a bias in the estimator if the regression did not include an intercept. This result generalizes naturally to variables with multiple choice values.

4 Estimation Reconsidered

Let us generalise the problem and re-write it in matrix form. Consider the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,, \tag{17}$$

where $\mathbf{X}$ and $\mathbf{W}$ are $N \times K$ and $N \times J$ data matrices of the explanatory variables, $\mathbf{y}$ is a $N \times 1$ vector containing the data of the dependent variable, $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of disturbance

[3] There is an abundant literature about the optimal number of choices (or 'scale points') in a survey, see e.g., Givon and Shapira (1984), Srinivasan and Basu (1989) or Alwin (1992).

terms, and finally $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $K \times 1$ and $J \times 1$ parameter vectors. $\mathbf{X}$ is not observed, only its discretized ordered choice version $\mathbf{X}^*$ is. Define the $MK \times K$ matrix as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{0} & \dots & \dots \\ \mathbf{0} & \mathbf{z}_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \dots & \ddots & \vdots \\ \dots & \dots & \mathbf{0} & \mathbf{z}_K \end{bmatrix},$$

where $\mathbf{z}_i = (z_{i1}, \dots, z_{iM})'$ contains the choice values for variable i. Let $\mathbf{E} = \{\mathbf{e}_{ki}\}$, where $k = 1, \dots K$ and $i = 1, \dots, N$ such that

$$\mathbf{e}_{ki} = \begin{bmatrix} \mathbf{1}(c_{k0} \leq x_{ki} < c_{k1}) \\ \mathbf{1}(c_{k1} \leq x_{ki} < c_{k2}) \\ \vdots \\ \mathbf{1}(c_{kM-1} \leq x_{ki} < c_{kM}) \end{bmatrix},$$

where x_{ki} denotes the value of the i^{th} observation from the explanatory variable x_k.
This implies $\mathbf{E}$ is a $MK \times N$ matrix since each entry $\mathbf{e}_{ki}$ is a $M \times 1$ vector. Following the definition of x_i^* in the paper, we can rewrite $\mathbf{X}^* = \mathbf{E}'\mathbf{Z}$.

4.1 The OLS Estimator

From Equation (17), consider the regression based on the observed data:

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + (\mathbf{X} - \mathbf{X}^*)\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{18}$$

then the OLS estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{*\prime}\mathbf{M}_{\mathbf{W}}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{M}_{\mathbf{W}}\mathbf{y},$$

where $\mathbf{M}_{\mathbf{W}} = \mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$ defines the usual residual maker. The standard derivation shows that

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\boldsymbol{\varepsilon}. \tag{19}$$

This implies OLS is unbiased if and only if $\left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X} = \mathbf{I}$. This allows us to investigate the bias analytically by examining the elements in $\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}$ and $\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X}$.

To simplify the analysis, we assume for the time being the following:

$$\mathbf{M}_{\mathbf{W}}\mathbf{X} = \mathbf{X} \tag{20}$$

$$\mathbf{M}_{\mathbf{W}}\mathbf{X}^* = \mathbf{X}^*. \tag{21}$$

In other words, we assume independence between $\mathbf{W}$ and $\mathbf{X}$, as well as its discretized choice version. This may appear to be a strong assumption but it does allow us to see what is happening somewhat better. We relax this at a latter stage.
The OLS estimator in this case becomes:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\boldsymbol{\varepsilon}.$$

The OLS is unbiased if $(\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}\mathbf{X} = \mathbf{I}$. Note that $\mathbf{Z}'$ and $\mathbf{E}$ are of size $K \times MK$ and $MK \times N$, respectively. This means $\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}$ are invertible as long as $N > K$, which is a standard assumption in classical regression analysis. Let us consider a typical element in $\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}$ first. Since $\mathbf{Z}$ is non-stochastic as it contains only all the pre-defined choice values, it is sufficient to examine $\mathbf{E}\mathbf{E}'$:

$$\mathbf{E}\mathbf{E}' = \begin{bmatrix} \mathbf{e}_{11} & \dots & \mathbf{e}_{1i} & \dots & \mathbf{e}_{1N} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}_{k1} & \dots & \mathbf{e}_{ki} & \dots & \mathbf{e}_{kN} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}_{K1} & \dots & \mathbf{e}_{Ki} & \dots & \mathbf{e}_{KN} \end{bmatrix} \begin{bmatrix} \mathbf{e}'_{11} & \dots & \mathbf{e}'_{k1} & \dots & \mathbf{e}'_{K1} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}'_{1i} & \dots & \mathbf{e}'_{ki} & \dots & \mathbf{e}'_{ki} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}'_{1N} & \dots & \mathbf{e}'_{kN} & \dots & \mathbf{e}'_{KN} \end{bmatrix}.$$

Note that each entry in $\mathbf{E}$ is a vector, so $\mathbf{E}\mathbf{E}'$ will result in a partition matrix whose elements are the sums of the outer products of $\mathbf{e}_{ki}$ and $\mathbf{e}_{lj}$ for $k, l = 1, \dots, K$ and $i, j = 1, \dots, N$. Specifically, let $\mathbf{q}_{kl}$ be a typical block element in $\mathbf{E}\mathbf{E}'$, then

$$\mathbf{q}_{kl} = \sum_{i=1}^{N} \mathbf{e}_{ki}\mathbf{e}'_{li}.$$

Let $\mathbf{1}_m^{ki} = \mathbf{1}\left(c_{km-1} \leq x_{ki} < c_{km}\right)$, then the (m, n) element in $\mathbf{q}_{kl}$, q_{mn} is $\sum_{i=1}^{N} \mathbf{1}_m^{ki}\mathbf{1}_n^{li}$ for $m, n = 1, \dots, M$. Thus, $\mathbb{E}\left(\mathbf{E}\mathbf{E}'\right)$ exists if $\mathbb{E}\left(\mathbf{1}_m^{ki}\mathbf{1}_n^{li}\right)$ exists,

$$\mathbb{E}\left(\mathbf{1}_m^{ki}\mathbf{1}_n^{li}\right) = \int_{\Omega} f(x_k, x_l) dx_k dx_l \,, \tag{22}$$

where $f(x_k, x_l)$ denotes the joint distribution of x_k and x_l and $\Omega = [c_{km-1}, c_{km}] \times [c_{ln-1}, c_{ln}]$ defines the region for integration. Thus, $N^{-1}b_{mn}$ should converge into Equation (22) under the usual WLLN.

Following a similar method, let a_{kl} be the (k, l) element in $\mathbf{Z}'\mathbf{E}\mathbf{X}$, then

$$a_{kl} = \sum_{i=1}^{N}\sum_{m=1}^{M} z_{km}\mathbf{1}_m^{ki}x_{li}.$$

Now,

$$\begin{aligned} \mathbb{E}\left[\sum_{m=1}^{M} z_{km}\mathbf{1}_m^{ki}x_{li}\right] &= \sum_{m=1}^{M} z_{km}\mathbb{E}\left[\mathbf{1}_m^{ki}x_{li}\right] \\ &= \sum_{m=1}^{M} z_{km}\int_{\Omega_1} x_l f(x_k, x_l) dx_k dx_l \,, \end{aligned} \tag{23}$$

where $\Omega_1 = [c_{km-1}, c_{km}] \times \Omega_{\mathbf{X}}$ with $\Omega_{\mathbf{X}}$ denotes the sample space of x_k and x_l. Thus, $N^{-1}a_{kl}$ should converge into Equation (23) under the usual WLLN.

In the case when Equations (20) and (21) do not hold, the analysis becomes more tedious algebraically, but it does not affect the result that OLS is biased. Recall Equation (19), and let ω_{ij} be the (i, j) element in $\mathbf{M}_{\mathbf{W}}$ for $i = 1, \dots, N$ and $j = 1, \dots, J$, then following the same

argument as above, $\mathbf{EM_W E'}$ can be expressed as a $M \times M$ block partition matrix with each entry a $K \times K$ matrix. The typical (m, n) element in the (k, l) block is

$$g_{kl} = \sum_{j=1}^{N} \sum_{i=1}^{N} \omega_{ij} \mathbf{1}_m^{ki} \mathbf{1}_n^{li} \tag{24}$$

with its expected value being

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \int_{\Omega} \omega_{ij} f\left(x_k, x_l, \mathbf{w}\right) dx_k dx_k d\mathbf{w}, \tag{25}$$

where $\mathbf{w} = (w_1, \ldots, w_J)$, $d\mathbf{w} = \prod_{i=1}^{J} dw_i$ and $\Omega = [c_{km-1}, c_{km}] \times [c_{ln-1}, c_{ln}] \times \Omega_{\mathbf{w}}$ where $\Omega_{\mathbf{w}}$ denotes the sample space of $\mathbf{w}$. Note that ω_{ij} is a nonlinear function of $\mathbf{w}$, and so the condition of existence for Equation (25) is complicated. However, under the assumption that the integral in Equation (25) exits, then $N^{-1} g_{kl}$ should converge to Equation (25) under the usual WLLN. It is also worth noting that $\mathbb{E}\left[\mathbf{M_W X}\right] = \mathbb{E}\left[\mathbf{M_W}\right] \mathbb{E}\left[\mathbf{X}\right] = \mathbb{E}\left[\mathbf{X}\right]$ and $\mathbb{E}\left[\mathbf{M_W X^*}\right] = \mathbb{E}\left[\mathbf{M_W}\right] \mathbb{E}\left[\mathbf{X^*}\right] = \mathbb{E}\left[\mathbf{X^*}\right]$ under the assumption of independence, which reduces Equation (25) to Equation (22).
Again, following the same derivation as above, a typical element in $\mathbf{Z'EM_W X}$ is

$$h_{kl} = \sum_{m=1}^{M} \sum_{i=1}^{N} z_{km} \mathbf{1}_m^{ki} u_{li}, \tag{26}$$

where $u_{li} = \sum_{\upsilon=1}^{N} \omega_{i\upsilon} X_{l\upsilon}$. Note that u_{li} is the i^{th} residual of the regression of X_l on $\mathbf{W}$. The expected value of h_{kl} can be expressed as

$$\sum_{m=1}^{M} z_{km} \int_{\Omega_m} u_l f(x_k, x_l, \mathbf{w}) dx_k dx_l d\mathbf{w}, \tag{27}$$

where u_l denotes the random variable corresponding to the i^{th} column of $\mathbf{M_W X}$ and $\Omega_m = [c_{km-1}, c_{km}] \times \Omega_{\mathbf{X}} \times \Omega_{\mathbf{w}}$ with $\Omega_{\mathbf{w}}$ denotes the sample space of $\mathbf{W}$. Note that $u_l = x_l$ under the assumption of independence, which reduces Equation (27) to Equation (23).

4.2 Extension to Panel Data

So far, we have dealt with cross-sectional data. Next, let us see what changes if we have panel data at hand, which is closer to the reality of data gathering through surveys. We can extend our basic model using Equation (2) to

$$y_{it} = w_{it}' \gamma + x_{it}^{*'} \beta + \varepsilon_{it}, \tag{28}$$

and adjust the DGP, based on Equation (3)

$$y_{it} = w_{it}' \gamma + x_{it}' \beta + u_{it}, \tag{29}$$

where $x_{it} \sim f_i(a_l, a_u)$ denotes an individual distribution with mean μ_i for $i = 1, \ldots, N$. Here we need to assume that $f_i(\cdot)$ is stationary, so the distribution may change over individual i

but not over time, t.

Now, the most important problem is identification. If the choice of an individual does not change over the time periods covered, the individual effects in the panel and the parameter associated with the choice variable cannot be identified separately. The Within transformation would wipe out the choice variable as well. When the choice does change over time, but not much, then we are facing weak identification, i.e., in fact very little information is available for identification, so the parameter estimates are going to be highly unreliable. This is a likely scenario when M is small, for example $M = 3$ or $M = 5$.
The bias of the panel data Within estimator can be easily shown. Let us re-write Equation (18) in a panel data context

$$\mathbf{y} = \mathbf{D}_N\boldsymbol{\alpha} + \mathbf{X}^*\boldsymbol{\beta} + [(\mathbf{X} - \mathbf{X}^*)\boldsymbol{\beta} + \boldsymbol{\varepsilon}]\,,$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$ and $\mathbf{D}_N$ is a $NT \times N$ zero-one matrix that appropriately selects the corresponding fixed effect elements form $\boldsymbol{\alpha}$. The Within estimator is

$$\hat{\boldsymbol{\beta}}^*_W = (\mathbf{X}^{*\prime}\mathbf{M}_{\mathbf{D}_N}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{M}_{\mathbf{D}_N}\mathbf{y}\,,$$

or equivalently

$$\hat{\boldsymbol{\beta}}^*_W = (\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{D}_N}\mathbf{E}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{D}_N}\mathbf{X}\boldsymbol{\beta} + (\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{D}_N}\mathbf{E}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{D}_N}\boldsymbol{\varepsilon}\,,$$

where

$$\mathbf{M}_{\mathbf{D}_N}\mathbf{y} = \mathbf{M}_{\mathbf{D}_N}\mathbf{X}^*\boldsymbol{\beta} + \mathbf{M}_{\mathbf{D}_N}[(\mathbf{X} - \mathbf{X}^*)\boldsymbol{\beta} + \boldsymbol{\varepsilon}].$$

The Within estimator is biased as $\mathbb{E}(\hat{\boldsymbol{\beta}}^*_W) \neq \boldsymbol{\beta}$, because $\mathbf{M}_{\mathbf{D}_N}\mathbf{E}'\mathbf{Z} = \mathbf{M}_{\mathbf{D}_N}\mathbf{X}^* \neq \mathbf{M}_{\mathbf{D}_N}\mathbf{X}$.

The solution is to have different choice classes (boundaries) for the different time periods as, for example, explained in the next section. After the appropriate Within transformation, the OLS can be applied with properties outlined in the previous sections and in the next.
In what follows, we remain with the cross-sectional data framework for simplicity, and when needed we refer to panel data solutions.

5 Consistent Estimation: Sub-sampling and Instrumental Variables

This section carries the main contribution of this paper. Here we propose two possible approaches to ensure consistent estimation, which later is extended with the use of instrumental variables as well. Most importantly, we depart from the classical econometric approach to estimation: we do not start assuming that the sample is given, but our main aim is to design an environment and sampling that delivers estimation with good enough precision. In other words, we investigate what is a good method to gather the data (what is a good survey design) in order to reduce the estimation bias presented earlier.
The main approach of the proposed methods is to create a number of sub-samples (S), while fixing the number (M) of choices in each sub-sample, in order to reduce the bias. The reason for fixing M is the restricted human cognitive capacity as noted above. Despite the restricted human cognitive capacity, we can achieve an increase in M through changing the class boundaries between each sub-sample, which in practice means different survey questionnaires for

each sub-sample. In fact, this approach utilizes the M consistency result previously discussed in Section 3, and transforms it into N consistency through using several sub-samples. The intuition behind the method is that this leads to a better mapping of the unknown distribution of x and so reduces the estimation bias. By merging the different sub-samples into one data set (let us call this the '*working sample*'), we get $b = 1, \ldots, B$ overall number of choice classes across the merged sub-samples, where B is much larger than M. In a given sub-sample each respondent (individual i) is given one questionnaire only (in the case of cross sectional data). The set of respondents who fill in the questionnaire with the same class boundaries defines a sub-sample. Each sub-sample has $N^{(s)}$, $s = 1, \ldots, S$ number of observations ($\sum_s N^{(s)} = N$). In this setup, a sub-sample looks exactly as the problem introduced above in (1), with the only difference across sub-sample that the class boundaries are different.[4] Note that the number of observations across sub-samples can be the same or, more likely, different. Now a sub-sample looks like:

$$x_i^{(s)} = \begin{cases} z_1^{(s)} & \text{if } x_i \in C_1^{(s)} = [c_0^{(s)}, c_1^{(s)}), \\ & \qquad \text{1st choice for sub-sample } s, \\ z_2^{(s)} & \text{if } x_i \in C_2^{(s)} = [c_1^{(s)}, c_2^{(s)}), \\ \vdots & \vdots \\ z_m^{(s)} & \text{if } x_i \in C_m^{(s)} = [c_{m-1}^{(s)}, c_m^{(s)}), \\ \vdots & \vdots \\ z_M^{(s)} & \text{if } x_i \in C_M^{(s)} = [c_{M-1}^{(s)}, c_M^{(s)}], \\ & \qquad \text{last choice for sub-sample } s. \end{cases} \tag{30}$$

Let us see a very simple illustrative example of this. Assume that $M = 2$, $S = 2$ as well, $N = 60$, $N^{(1)} = 30$ and similarly $N^{(2)} = 30$. Let x be a continuously distributed variable in the $[0, 4]$ domain and let the class boundaries in the first sub-sample be $[0, 2)$ and $[2, 4]$, while in the second sub-sample $[0, 1)$ and $[1, 4]$, with $10, 20, 5$, and 25 observations respectively in each class. Next, we merge the information obtained in the two sub-samples in one working sample in such a way that we are not introducing any selection bias. This working sample now has $B = 3$ classes (or bins): $[0, 1)$, $[1, 2)$ and $[2, 4]$ and number of observations N^{WS} with the working sample's artificially created variable x_i^{WS}. Using the information from the 2nd sub-sample, we know that of 30 observations 5 are in the 1st bin. Similarly, we can deduce that in the 2nd bin there are 5 observations as well, while in the last 3rd bin 20 (see Figure 3 below). Piecing this information together, we can create x_i^{WS}. Clearly, this way the working sample maps better the unknown distribution of x than any of the two sub-samples.

[4] In order to simplify the notation, we use instead of $x^{*(s)}$ simply $x^{(s)}$. For each sub-sample there is a new random variable. Each of them is a discretized realization of the unknown random variable x.

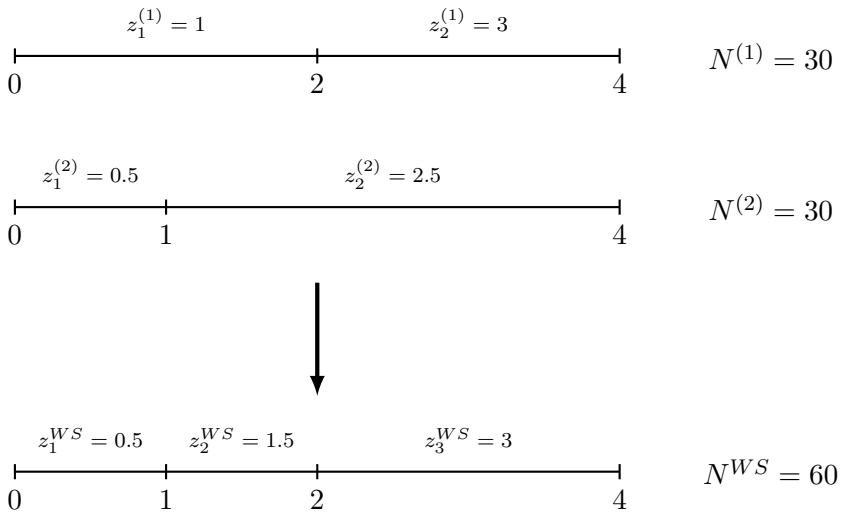

Figure 3: The basic idea of sub-sampling

5.1 Construction of the Working Sample

The construction of questionnaires for each sub-sample and the merger into the working sample can be done in many different ways, depending on boundary point setup ($c_m^{(s)}$) and on the choice values ($z_m^{(s)}$) for the sub-samples. We assume that the number of observations (N), their allocation among sub-samples ($N^{(s)}$) and the number of sub-samples (S) are given, and also that the number of choices (M) is fixed across sub-samples.

We assume now that the class boundaries in the working sample are the union of the sub-samples' class boundaries, that is

$$\bigcup_{b=0}^{B} c_b^{WS} = \bigcup_{s=1}^{S} \bigcup_{m=0}^{M} c_m^{(s)} \,.$$

This translates in our example to the following: $c_0^{WS} = c_0^{(1)} = c_0^{(2)} = 0$; $c_1^{WS} = c_1^{(2)} = 1$; $c_2^{WS} = c_1^{(2)} = 2$; $c_3^{WS} = c_2^{(1)} = c_2^{(2)} = 4$.
Also, we restrict the domain of the underlying distribution for each sub-sample: Let the two boundary points of the underlying distribution a_l and a_u, such that $a_l < a_u$. Then we construct the sub-sample questionnaires' and the working sample's boundary points such that: $a_l = c_0^{(s)} = c_0^{WS},\ a_u = c_M^{(s)} = c_B^{WS},\ \forall s$. A further important case is when we use infinite boundary points for a_l and/or a_u. Then all sub-samples have also infinite boundary points at the boundary.
With the creation of S sub-samples, we in fact introduce

$$x^{(1)}, \ldots, x^{(s)}, \ldots, x^{(S)}$$

new random variables ($x^{(s)} := \psi^{(s)}(x)$), where $\psi^{(s)}(\cdot)$ is the function that discretizes the continuous x into the choices of the sub-sample s. These then define a new random variable $x^{WS} = \Psi(x^{(1)}, \ldots, x^{(s)}, \ldots, x^{(S)})$ standing for the working sample, where $\Psi(\cdot)$ is the 'merging function'.

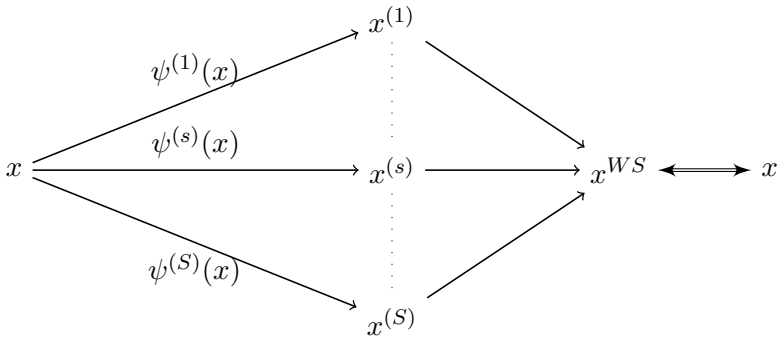

Figure 4: Creation of the working sample's random variable

Each of the methods to be discussed below specifies the functions $\psi^{(s)}$, the merging function $\Psi(\cdot)$ and defines the random variable of the working sample x^{WS}. These functions are different across the methods, but all of them reflect the unknown random variable x. To do so, we need the following property to hold

$$\lim_{s\to\infty} \mathbb{E}_S \left[x^{WS}|y\right] = \mathbb{E}\left[x|y\right] ,$$

which means that no selection bias is introduced through the mapping.

5.2 Probabilities in the Working Sample

First, when creating the working sample we must pay special attention to avoiding any kind of 'selection bias' through the merger of the sub-samples.
Then, we also need to address, for later use, the following question: What is the probability for a given observation to be in a given choice class in the working sample? To derive this, first we have to derive the probability of an observation falling into a given sub-sample's choice class. Then we can introduce an assigning mechanism for an observation in a sub-sample to a working sample class. Using this result, we can get the unconditional probability for an observation to be in a given class in the working sample.
At the start, all individuals are allocated into a sub-sample. This, of course, defines the number of observations in each sub-sample ($N^{(s)}$), which in turn translates into the probability of a given observation x, being in sub-sample s: $\Pr(x \in s)$. Uniformly assigning these individuals to sub-samples is the most straightforward procedure, thus $\Pr(x \in s) = 1/S$, however for the general case we are going to use the probabilistic notations.
Now, we can define the probability for an observation to be between given boundary points in a given sub-sample:

$$\Pr\left(x \in C_m^{(s)}\right) = \Pr(x \in s) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,.$$

In the next step, as we observe a response in a given sub-sample, we would like to derive the probability of an observation falling between given boundary points in the working sample. For this (with no additional information), we need to assign these uniformly into the working sample's classes.[5] This is a way to avoid introducing any kind of selection bias during the

[5]Here we assume that the boundary points in the working sample are the union of the sub-samples' boundary points.

merging process.

$$\Pr\left(x \in C_b^{WS} \mid x \in C_m^{(s)}\right) = \begin{cases} \frac{c_b^{WS} - c_{b-1}^{WS}}{c_m^{(s)} - c_{m-1}^{(s)}}, & \text{if} \quad c_b^{WS} \leq c_m^{(s)} \text{ and } c_{b-1}^{WS} \geq c_{m-1}^{(s)}\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

Using the above two equations, we need to assign each individual from all sub-samples into the working sample without any additional information. Thus, the unconditional probability of an individual falling in the working sample between given boundary points is

$$\Pr\left(x \in C_b^{WS}\right) = \sum_{s=1}^{S} \Pr(x \in s) \sum_{m=1}^{M} \Pr\left(x \in C_b^{WS} \mid x \in C_m^{(s)}\right) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,. \tag{31}$$

To simplify, we can assume uniform assignment of the observations to each sub-sample, and write

$$\Pr\left(x \in C_b^{WS}\right) = \frac{1}{S} \sum_{s=1}^{S} \sum_{\substack{m \\ \text{if } C_b^{WS} \in C_m^{(s)}}} \frac{c_b^{WS} - c_{b-1}^{WS}}{c_m^{(s)} - c_{m-1}^{(s)}} \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,.$$

Let us make a practical remark. In some cases x may have infinite support which implies classes not bounded from below and/or above. Usually, this is related to survey questions like *"... or less"* or *"... or more"*. Here we are facing censoring. As a consequence, the difference between the class's choice value (e.g., $z_1^{(s)}$ in Equation (1)) and the class's conditional mean for the underlying distribution can be potentially infinite, resulting in very large estimation biases. We are going to see how to deal with this problem later on.

5.3 The Magnifying Method

In the magnifying method, we magnify the domain of the answers within the original domain of the unknown distribution of x by one equally sized choice class. The sizes of the classes depend on the number of sub-samples (S) and the number of choice values (M). As the number of sub-samples increases the class sizes decrease, which is the main benefit helping us uncover the unknown underlying distribution.[6] Figure 5 shows the main idea of the magnifying method: The last line shows the working sample, while above, we can see the individual questionnaires for the case of $M = 3, S = 4$.

[6] We fix the number of choice values M by assumption.

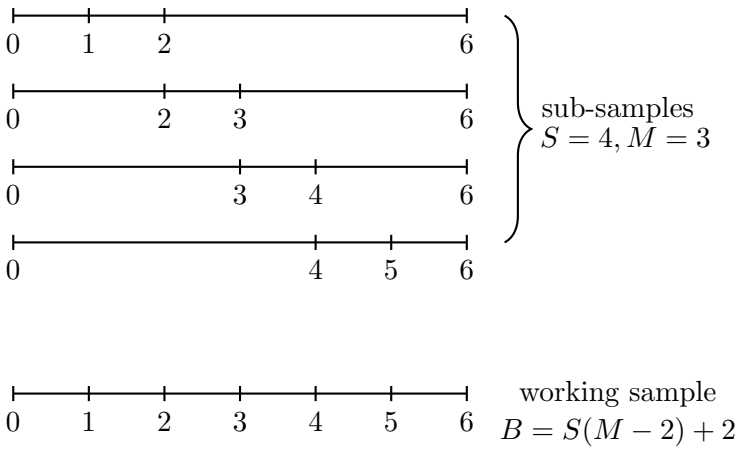

Figure 5: The magnifying method

The first and last sub-samples are slightly different from the sub-samples in between. They have one extra class with the same class width, while sub-samples in between have $M-2$ classes with the same class width. To further explore the properties of the magnifying method, let us establish the connection between the number of magnified classes in the working sample (B), and the number of sub-samples (S) and choices (M)

$$B = S(M-2)+2\,.$$

As mentioned above, we have 2 sub-samples, which lie in the boundary of the domain and capture $M-1$ classes of equal size; and there are $S-2$ sub-samples in between which are capturing $M-2$ classes. After some manipulations, we get the number of the classes in the working sample.
Given the fact that there are B classes in the working sample, we get the widths of these classes, let us call it h such

$$h = \frac{a_u - a_l}{S(M-2)+2}\,.$$

Fixing the upper and lower bounds on the support for the sub-samples ($a_l = c_0^{WS} = c_0^{(s)};\ a_u = c_B^{WS} = c_M^{(s)},\ \forall s$), one can reduce the class size $h \to 0$ as $S \to \infty$. This can also be seen through the working sample's boundary points, which have the following simple form

$$c_b^{WS} = a_l + bh = a_l + b\frac{a_u - a_l}{S(M-2)+2}\,.$$

To show how the number of sub-samples affects the bias, we need the boundary points for each sub-sample:

$$c_m^{(s)} = \begin{cases} a_l \text{ or } -\infty & \text{if } m = 0, \\ a_l + mh & \text{if } 0 < m < M \text{ and } s = 1, \\ a_l + h\left[(s-2)(M-2) + M + m - 2\right] & \text{if } 0 < m < M \text{ and } s > 1, \\ a_u \text{ or } \infty & \text{if } m = M. \end{cases} \tag{32}$$

The intuition behind this is the following: on the boundaries of the support, the sub-samples take the value of the lower and upper bound. For the first sub-sample, one needs to shift the

boundary points m times. However, for the other sub-samples, one needs to push by $h(M-1)$ to shift through the first questionnaire and then $h(M-2)$ to shift through each in between sub-samples $s-2$ times, and then go to the $m-1$ part. Doing the algebra will result in the above equation.[7] Algorithm 1 below summarizes how to create the sub-samples in practice.

Algorithm 1 Magnifying method – creation of the sub-samples ($\psi^{(s)}(\cdot)$)

1: For any given S and M. Set

$$\begin{aligned} B =& S(M-2)+2 \\ h =& \frac{a_u - a_l}{B} \\ s =& 1. \end{aligned}$$

2: Set $c_0^{(s)} = a_l$ and $c_M^{(s)} = a_u$.
3: If $s = 1$, then set

$$c_1^{(s)} = c_0^{(s)} + h,$$

else set

$$c_1^{(s)} = c_{M-1}^{(s-1)}.$$

4: Set $c_m^{(s)} = c_{m-1}^{(s)} + h$ for $m = 2, \ldots, M-1$.
5: If $s < S$ then $s := s+1$ and goto Step 2.

From Equation (32), it is clear that the class widths differ from each other within a sub-sample. Let $||C_m^{(s)}|| = c_m^{(s)} - c_{m-1}^{(s)}$ be the m-th class width, then for the sub-samples which are in-between the boundaries ($1 < s < S$) and substituting for h, we can write

$$||C_m^{(s)}|| = \begin{cases} (a_u - a_l)\left(\frac{s(M-2)+2}{S(M-2)+2} + \frac{1-M}{S(M-2)+2}\right) & \text{if } m = 1,\ 1 < s < S, \\ \frac{a_u - a_l}{S(M-2)+2} & \text{if } 1 < m < M,\ 1 < s < S, \\ (a_u - a_l)\left(1 - \frac{s(M-2)+1}{S(M-2)+2}\right) & \text{if } m = M,\ 1 < s < S\,. \end{cases}$$

We can also define the class widths for the first and last sub-samples such as

$$||C_m^{(1)}|| = \begin{cases} \frac{a_u - a_l}{S(M-2)+2} & \text{if } 1 \le m < M, \\ (a_u - a_l)\left(1 - \frac{M-1}{S(M-2)+2}\right) & \text{if } m = M, \end{cases}$$

$$||C_m^{(S)}|| = \begin{cases} (a_u - a_l)\left(1 - \frac{M-1}{S(M-2)+2}\right) & \text{if } m = 1, \\ \frac{a_u - a_l}{S(M-2)+2} & \text{if } 1 < m \le M. \end{cases}$$

Note that $||C_m^{(s)}|| \le ||C_1^{(s)}||$ and $||C_m^{(s)}|| \le ||C_M^{(s)}||$. Formally, let us define $\zeta := \{C_m^{(s)} \mid 1 < m < M, 1 < s < S, C_m^{(1)} \mid 1 \le m < M, C_m^{(S)} \mid 1 < m \le M\}$ as the set of classes which have the class width $\frac{a_u - a_l}{S(M-2)+2}$. Then we can write $\Pr\left((x - x^{(s)})^2 | x \in \zeta \le (x - x^{(s)})^2 \mid x \notin \zeta\right) = 1$, which is true if and only if, $\mathbb{E}[x] = \mathbb{E}\left[x^{(s)}\right], \forall x$. One example is when x is uniformly distributed.

[7] There is an alternative way to formalize the boundary points, when one starts from a_u. The formalism will then be symmetric and results in the same conclusions.

Now, let us check the limit in the number of sub-samples. We end up with the following limiting cases

$$\lim_{S\to\infty}\left(||C_m^{(s)}||\right)=\begin{cases}0 & \text{if } m=1,\ 1<s<S,\\ 0 & \text{if } 1<m<M,\ 1<s<S,\\ a_u-a_l & \text{if } m=M,\ 1<s<S.\end{cases}$$

And for the first and last sub-sample

$$\lim_{S\to\infty}\left(||C_m^{(1)}||\right)=\begin{cases}0 & \text{if } 1\leq m<M,\\ a_u-a_l & \text{if } m=M,\end{cases}$$

$$\lim_{S\to\infty}\left(||C_m^{(S)}||\right)=\begin{cases}a_u-a_l & \text{if } m=M,\\ 0 & \text{if } 1<m\leq M.\end{cases}$$

This formulation takes a_l as the starting point and expresses the boundary points given a_l. However, we can use a_u as the starting point as well to shift the boundary point. This implies that the convergences on the bounds $(||C_1^{(s)}||, ||C_M^{(s)}||)$ will change, resulting in those parts not converging to 0 in general.

Based on the different magnitude of the measurement error, depending on class widths, now it is clear that there are two types of observations: The first type is $x_i^{(s)} \in \zeta$. Here, the error is the smallest and has the feature of $\lim_{S\to\infty}||C_m^{(s)}|| = 0$. Moreover, these observations have the same class width as the working sample's classes and each of them can be directly linked to a certain working sample class by design. Formally, $\exists\, C_m^{(s)} \cong C_b^{WS}$ such that $c_m^{(s)} = c_b^{WS}$, $c_{m-1}^{(s)} = c_{b-1}^{WS}$. We call these values '*directly transferable observations*', as we can directly transfer and use them in the working sample. These observations are denoted by $x_{i,DTO}^{WS} := x_i^{(s)} \in \zeta,\ \forall s$, and the related random variable by x_{DTO}^{WS}.[8]

The second type of observations are all the others for which none of the above is true. We call them '*non–directly transferable observations*'. Algorithm 2 describes how to construct in practice the working sample, using the directly transferable observations.

[8]Notation: for the estimation we are using the superscript 'WS' and defining the construction method in the subscript – here 'DTO'.

Algorithm 2 Magnifying method - creation of the 'DTO' working sample ($\Psi_{DTO}(\cdot)$)

1: Set $m = 1, s = 1$ and $x^{WS}_{i,DTO}, y^{WS}_{i,DTO}, w^{WS}_{i,DTO} = \emptyset$.

2: If $C^{(s)}_m \in \zeta$, add observations from class $C^{(s)}_m$ to the working sample:

$$x^{WS}_{i.DTO} := \left\{ x^{WS}_{i,DTO}, \bigcup_{j=1}^{N} \left(x^{(s)}_j \in C^{(s)}_m \right) \right\},$$

$$y^{WS}_{i.DTO} := \left\{ y^{WS}_{i,DTO}, \bigcup_{j=1}^{N} y^{(s)}_j \mid \left(x^{(s)}_i \in C^{(s)}_m \right) \right\},$$

$$w^{WS}_{i.DTO} := \left\{ w^{WS}_{i,DTO}, \bigcup_{j=1}^{N} w^{(s)}_j \mid \left(x^{(s)}_i \in C^{(s)}_m \right) \right\},$$

3: If $s < S$, then $s := s + 1$ and go to Step 2.

4: If $s = S$, then $s := 1$ and set $m = m + 1$ and go to Step 2.

Before proving the consistency of $\hat{\beta}$, using only $x^{WS}_{i,DTO}$ — the *directly transferable observations* in the working-sample — we need to make some assumptions on these observations.
The probability that a *directly transferable observation* lies in a given class of the working sample can be written based on Equation (31) as follows

$$\Pr\left(x \in C^{WS}_b\right) = \Pr(x \in s) \int_{c^{WS}_{b-1}}^{c^{WS}_b} f(x)\mathrm{d}x\,.$$

Here we used the fact that individual i being assigned to a sub-sample s is independent from i choosing the class with choice value $z^{(s)}_m$.
We want to ensure that in each class in the working sample, there are directly transferred observations. This will ensure that estimation is feasible. Thus, for each sub-sample the expected number of directly transferable observations is

$$\begin{aligned} \mathbb{E}(N^{WS}_b) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C^{WS}_b\}} \right) \\ &= N \Pr(x \in s) \int_{c^{WS}_{b-1}}^{c^{WS}_b} f(x)\mathrm{d}x. \end{aligned} \tag{33}$$

Looking at Equation (33), we need some basic assumptions:

- $\Pr(x \in s) > 0$, which means that there are individuals assigned to each sub-sample s.
- $\int_{c^{WS}_{b-1}}^{c^{WS}_b} f(x)\mathrm{d}x > 0$, thus the underlying distribution has positive values between the boundary points c^{WS}_{b-1}, c^{WS}_b. This means there is a positive probability that the assigned individuals take choices between c^{WS}_{b-1}, c^{WS}_b.

To achieve consistency, we would like to establish that we have observations in each working sample class

$$\Pr\left(\mathbb{E}(N^{WS}_b) > 0\right) \to 1\,. \tag{34}$$

For this, we can reformulate the Equation (33) by considering the number of observations up to a certain boundary point, rather than the number of observations in a particular class. That is checking for

$$\Pr\left(\mathbb{E}\left[\sum_{i=1}^{b} N_i^{WS}\right] > 0\right) \to 1\,.$$

This gives the possibility to replace $\int_{c_{b-1}^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x$ with $\int_{c_0^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x$. Since this is a CDF, and hence a non-decreasing function, which is effectively showing that each class has non-empty observations, we can write the following:

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{b} N_i^{WS}\right) =& \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i < c_b^{WS}\}}\right)\\ =& N\Pr(x \in s)\int_{c_0^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x.\end{aligned}$$

Next, we need to show that this is an increasing function in C_b^{WS}. Now as $N \to \infty$, under the assumption that $\Pr(x \in s) = 1/S$ and $S/N \to \mathrm{c}$ with $\mathrm{c} \in (0,1)$ (this is satisfied when $S = cN$)

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left(\sum_{i=1}^{b} N_i^{WS}\right) =& N\Pr(x_i < C_b^{WS})\\ =& \frac{1}{\mathrm{c}}\int_{C_0^{WS}}^{C_b^{WS}} f(x)dx.\end{aligned}$$

Note that the derivative with respect to C_b^{WS} is $\frac{1}{\mathrm{c}} f\left(C_b^{WS}\right) > 0$, so the expected number of observations in each class is not 0. This completes our proof.

Some remarks: We can decrease c as close to 0 as we would like to. This means that there is an equal or higher number of sub-samples than observations. On the other hand, we exclude by assumption the case when $\mathrm{c} \geq 1$, which means that there are equal or more number of sub-samples than observations. Then we most certainly would not observe values for each working sample class.

This leads us to convergence in distribution. Using x_{DTO}^{WS}, there is a direct mapping between the classes of the sub-samples directly transferable observations and the working sample classes (thus, not inducing any distortion to the working sample's distribution). As we have proven Equation (34), we can say $x_{DTO}^{WS} \xrightarrow{d} x$ under the above assumptions. This way we have converted S convergence into an M convergence, restoring the underlying continuous distribution, which implies that the classical econometric results stand and the OLS estimator is going to be consistent for β (and γ as well).

Next, let us consider the placement of the *non-directly transferable observations*. We have seen that these observations do not reduce the measurement error in a systematic way. One way to proceed is to drop them completely so that they do not appear in the working sample (thus only using $x_{i,DTO}^{WS}$). However, it seems that too many could fall into this category, inducing a large efficiency loss during the estimation.

Another approach would be to use the information available for these observations: the known boundary points for these values. Then we could use all the *directly transferable observations* from the working sample to calculate the conditional averages for all *non-directly transferable observations* and replace them with those values. This way one could create another variable

to work with, which has the same number of observations as the number of respondents. Let us denote this new variable $x^{WS}_{i,ALL}$. This represents all the directly transferable observations and the replaced values for non-directly transferable observations as well.
Let us formalize the non-directly transferable observations such as $x^{(s)}_i \in C_\chi$, where

$$C_\chi := \bigcup_{s,m} C^{(s)}_m \bigcap_b C^{WS}_b = \zeta^{\complement}$$

is the set for non-directly transferable observations from all sub-samples, with $\chi = 1, \ldots, 2(S-1)$. We can then replace $x^{(s)}_i \in C_\chi$ with $\hat{\pi}_\chi$, which denotes the sample conditional averages

$$\hat{\pi}_\chi = \left(\sum_{i=1}^{N} \mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}} \right)^{-1} \sum_{i=1}^{N} \mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}} x^{WS}_{i,DTO}.$$

Let us introduce $x^{WS}_{i,NDTO}$ as the variable which contains all the replaced values with $\hat{\pi}_\chi$, $\forall x^{(s)}_i \in C_\chi$. This way we can create a new working sample as $x^{WS}_{i,ALL} := \{x^{WS}_{i,DTO}, x^{WS}_{i,NDTO}\}$, which contains information from both types of observations.

Let us call $\hat{\pi}_\chi$ the 'replacement estimator' of the conditional expectation of the given class. Under the WLLN, it is straightforward to show that the 'replacement estimator' for the sample conditional averages converges to the conditional expectations, thus $\hat{\pi}_\chi \to \mathbb{E}(x|x \in C_\chi)$ as $N, S \to \infty$ under the same assumptions as before. This also implies $x^{WS}_{i,NDTO} \to \mathbb{E}(x|x \in C_\chi)$, which means we are not introducing any errors during the estimation when working sample $x^{WS}_{i,ALL}$. Algorithm 3 describes how to create in practice the working sample using all observations. We also need to say something about the asymptotic standard errors

Algorithm 3 The magnifying method - creation of 'ALL' working sample ($\Psi_{ALL}(\cdot)$)

1: Let, $x^{WS}_{i,ALL} := \{x^{WS}_{i,DTO}\}$, $y^{WS}_{i,ALL} := \{y^{WS}_{i,DTO}\}$, $w^{WS}_{i,ALL} := \{w^{WS}_{i,DTO}\}$
2: Set, $m = 1, s = 1$
3: If $C^{(s)}_m \in C_\chi$, then calculate $\hat{\pi}_\chi$ and expand the working sample as,

$$x^{WS}_{i.ALL} := \left\{ x^{WS}_{i,ALL}, \bigcup_{j=1}^{N} \hat{\pi}_\chi \mid \left(x^{(s)}_j \in C^{(s)}_m \right) \right\},$$

$$y^{WS}_{i.ALL} := \left\{ y^{WS}_{i,ALL}, \bigcup_{j=1}^{N} y^{(s)}_j \mid \left(x^{(s)}_i \in C^{(s)}_m \right) \right\},$$

$$w^{WS}_{i.ALL} := \left\{ x^{WS}_{i,ALL}, \bigcup_{j=1}^{N} w^{(s)}_j \mid \left(x^{(s)}_i \in C^{(s)}_m \right) \right\},$$

4: If $s < S$, then $s := s + 1$ and go to Step 3.
5: If $s = S$, then $s := 1$ and set $m = m + 1$ and go to Step 3.

of this estimator, because if these are large, the replacement might not be favorable, as it may induce some uncertainty. To do so, one can think of $\hat{\pi}_\chi$ as an OLS estimator, regressing $\mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}}$ on $x^{WS}_{i,DTO}$. Here $\mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}}$ is a vector of indicator variables, created by $2(S-1)$

indicator functions: It takes a value of one for the directly transferable observations, which are within C_χ.[9] We can now write the following:

$$x_{i,DTO}^{WS} = \boldsymbol{\pi}_\chi \mathbf{1}_{\{x_{i,DTO}^{WS} \in C_\chi\}} + \eta_i,$$

where $\boldsymbol{\pi}_\chi$ stands for the vector of $\pi_\chi, \forall\chi$. The OLS estimator of $\boldsymbol{\pi}_\chi$ is

$$\hat{\boldsymbol{\pi}}_\chi = \left(\mathbf{1}'_{x_{i,DTO}^{WS} \in C_\chi} \mathbf{1}_{\{x_{i,DTO}^{WS} \in C_\chi\}}\right)^{-1} \mathbf{1}'_{\{x_{i,DTO}^{WS} \in C_\chi\}} x_{i,DTO}^{WS},$$

and under the standard OLS assumptions, we can write

$$\sqrt{N_{DTO}^{WS}} \left(\hat{\boldsymbol{\pi}}_\chi - \boldsymbol{\pi}_\chi\right) \overset{\text{a}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Omega}_\chi\right),$$

where $\boldsymbol{\pi}_\chi = \mathbb{E}(x|x \in C_\chi), \forall\chi$.
The variance of the OLS estimator is

$$\boldsymbol{\Omega}_\chi = \mathrm{V}\left(\eta_i\right) \left(\mathbf{1}'_{x_{i,DTO}^{WS} \in C_\chi} \mathbf{1}_{\{x_{i,DTO}^{WS} \in C_\chi\}}\right)^{-1}.$$

Using this results we may decide whether to replace or not to replace.

We need to consider the censoring case for the magnifying method. One easy solution is to drop those observations which have infinite class boundary. In the magnifying method, this means not having observations in the class(es) C_1^{WS} if we have $a_l = -\infty$ and/or C_B^{WS} if $a_u = \infty$. This is having an effect on the underlying distributions as well. We artificially truncate both $y \to y^{tr}$ and $x \to x^{tr}$. For the truncated distribution, we can use all the derivations presented above, and we end up with convergence in distribution as well: $F_{n,s}(x^{WS} \in \zeta^{tr}) \xrightarrow{d} F(x^{tr})$.[10] Furthermore, the parameter estimates $\beta^{tr} = \beta$ (under some reasonable assumptions), which implies that the OLS estimator is consistent for the truncated observations as well. Note that truncation implies that we cannot replace the observations from the sub-samples with infinite boundaries, and also that the replacement estimator does not converge to the conditional expectation due to the truncation.

5.4 The Shifting Method

The shifting method approaches the problem in a different way. It takes the original questionnaire as given, with fixed class widths, and shifts the boundaries of each choice with a given fixed value. This results in fixed class widths for the different sub-samples, except in the boundary classes where the widths are changing. Increasing the sub-sample size does not affect the boundary widths in between the support, only the size of the shift. We can approach this method in two ways: logically we could consider the original questionnaire, and take the number of choices as fixed here, then as we shift the boundaries, add an extra class for each sub-sample at the boundary where, due to the shift, the class width has increased. For the mathematical derivations, however, it is better to look at each sub-sample separately and take the number of classes in each sub-sample as given, with the exception of the first sub-sample, regarded as the starting benchmark. This way, in the first sub-sample there is

[9] The indicator variables are not independent of each other, while the non-transferable observation classes (C_χ) are overlapping each other.

[10] $\zeta^{tr} := \zeta \cap \{C_1^{WS}, C_B^{WS}\}$.

one class less. We rely on this approach and Figure 6 shows the sub-samples in this logic with $S = 4$ and with $M = 4$ classes.

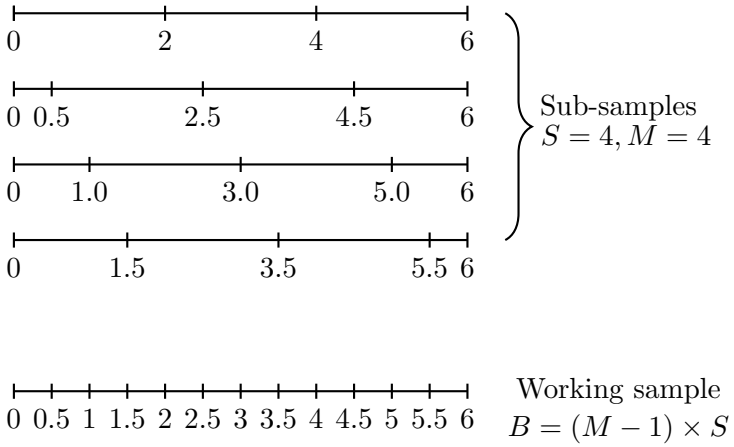

Figure 6: The shifting method

As Figure 6 shows there is one sub-sample (the benchmark $s = 1$) where there is one class less $(M - 1)$. Or, if we prefer, we can look at the benchmark as where we shifted the boundaries with zero. To get the properties of the working sample, let us define the class widths for the first sub-sample as $\frac{a_u - a_l}{M-1}$. We want to split this into S part in order to be able to shift the boundaries S times in order to have S sub-samples. Thus, the size of the shift is $\frac{a_u - a_l}{S(M-1)}$. This way we can define the number of classes in the working sample as

$$B = S(M - 1).$$

Now, the boundary points for each sub-sample are

$$c_m^{(s)} = \begin{cases} a_l \text{ or } -\infty, & \text{if } m = 0, \\ a_l + (s-1)\frac{a_u - a_l}{S(M-1)} + (m-1)\frac{a_u - a_l}{M-1} & \text{if } 0 < m < M, \\ a_u \text{ or } \infty, & \text{if } m = M. \end{cases}$$

For the working sample, we get $c_b^{WS} = a_l + b\frac{a_u - a_l}{S(M-1)}$. The class widths are

$$||C_m^{(s)}|| = \begin{cases} 0, & \text{if } s = 1 \text{ and } m = 1, \\ \frac{a_u - a_l}{M-1}, & \text{if } 1 < m < M, \\ (s-1)\frac{a_u - a_l}{S(M-1)}, & \text{otherwise.} \end{cases}$$

and for the class size in the working sample: $||C_b^{WS}|| = \frac{a_u - a_l}{S(M-1)}$.
Some additional remarks on the boundary points:

- $C_1^{(1)}$ is a non-existent class defined by the formalism, meaning it does not contain any observations, while it has class a size of 0.
- There are only two classes in the sub-samples which are congruent (with the same boundary points) with the classes in the working sample: $C_1^{(2)} \cong C_1^{WS}, C_M^{(S)} \cong C_B^{WS}$. This means that directly transferable observations will not help us here.

- One can not decrease the class widths between $C_2^{(s)}$ and $C_{M-1}^{(s)}$ in the sub-samples, with increasing the number of sub-samples.
- However, the class widths in the working sample are indeed decreasing with increase the number of sub-samples.

Algorithm 4 describes how to create in practice sub-samples using the shifting method.

Algorithm 4 The shifting method - creation of sub-samples ($\psi^{(s)}(\cdot)$)

1: For any given S and M, set

$$\begin{aligned} B =& S(M-1) \\ h =& \frac{a_u - a_l}{B} \\ \Delta =& \frac{a_u - a_l}{M-1} \\ s =& 1. \end{aligned}$$

2: Set $c_0^{(s)} = a_l$ and $c_M^{(s)} = a_u$.
3: If $s = 1$, set

$$c_m^{(s)} = c_{m-1}^{(s)} + \Delta, \qquad m = 2, \ldots, M-1$$

else

$$c_m^{(s)} = c_m^{(s-1)} + h, \qquad m = 1, \ldots, M-1.$$

Note: $c_1^{(1)}$ does not exist.
4: If $s < S$ then $s := s + 1$ and goto Step 2.

The idea is to reconstruct the underlying distribution $f(x)$, with creating a new random variable, which incorporates the information content of the boundary points.
The observations of a class in the sub-sample s, can end up in several classes in the working sample so the union of these classes gives one of the classes from the sub-samples

$$C_m^{(s)} = \begin{cases} \emptyset, & \text{if, } s = 1 \text{ and } m = 1, \\ \bigcup_{b=1}^{s-1} C_b^{WS}, & \text{if, } s \neq 1 \text{ and } m = 1, \\ \bigcup_{b=s-1+(m-1)(M-1)}^{s-1+(m-1)(M-1)} C_b^{WS}, & \text{if, } 1 < m < M, \\ \bigcup_{b=B-S+s-1}^{B} C_b^{WS}, & \text{if } m = M. \end{cases} \tag{35}$$

Now, define $Z(s, m)$, which creates sets for the scalar values of the working sample's choice values (z_b^{WS}) for each sub-sample class $C_m^{(s)}$,

$$Z(s,m) = \begin{cases} \{\emptyset\}, & \text{if, } s = 1 \text{ and } m = 1, \\ \bigcup_{b=1}^{s-1} \{z_b^{WS}\}, & \text{if, } s \neq 1 \text{ and } m = 1, \\ \bigcup_{b=s-1+(m-1)(M-1)}^{s-1+(m-1)(M-1)} \{z_b^{WS}\}, & \text{if, } 1 < m < M, \\ \bigcup_{b=B-S+s-1}^{B} \{z_b^{WS}\}, & \text{if } m = M. \end{cases} \tag{36}$$

The number of elements that $Z(s, m)$ contains, depends on the sub-sample and its class. We use these sets to create a new artificial variable $x_i^{\dagger}$ as follows.

Let us start with an example. *Let an observation $x_i^{(s)} \in C_m^{(s)}$. From Equation (35) we know which working sample classes are included in $C_m^{(s)}$. Furthermore, we also have a set of possible working sample choice values $Z(s,m)$. Now $x_i^\dagger$ will be a randomly chosen element of $Z(s,m)$, using uniform probabilities.*
The assignment mechanism can be written as

$$x_i^\dagger | x_i^{(s)} \in C_m^{(s)} = z \in Z(s,m), \text{ with } \begin{cases} \Pr(1), & \text{if } s=1 \text{ and } m=1, \\ \Pr(1/(s-1)), & \text{if } s \neq 1 \text{ and } m=1, \\ \Pr(1/S), & \text{if } 1 < m < M, \text{ or} \\ \Pr(1/(S-s+1)), & \text{if } m = M. \end{cases} \tag{37}$$

While by the definition, there is a direct mapping between $z \in Z(s,m)$ and C_b^{WS}, we can write the probability of $x_i^\dagger \in C_b^{WS}$, using Equation (31) and assuming $\Pr(x \in s) = 1/S$,

$$\Pr\left(x_i^\dagger \in C_b^{WS}\right) = \begin{cases} 0, & \text{if } s=1 \text{ and } m=1, \\ \frac{1}{S}\sum_{s=2}^{S} \frac{1}{s-1} \int_{C_1^{(s)} | C_b^{WS} \in C_1^{(s)}} f(x)\mathrm{d}x, & \text{if } s \neq 1 \text{ and } m=1, \\ \frac{1}{S^2}\sum_{s=1}^{S} \int_{C_m^{(s)} | C_b^{WS} \in C_m^{(s)}} f(x)\mathrm{d}x, & \text{if } 1 < m < M, \\ \frac{1}{S}\sum_{s=1}^{S} \frac{1}{S-s+1} \int_{C_M^{(s)} | C_b^{WS} \in C_M^{(s)}} f(x)\mathrm{d}x, & \text{if } m = M\,. \end{cases}$$

Algorithm 5 describes how to create an artificial variable which approximates the underlying distribution of x.

Algorithm 5 The shifting method – creation of artificial variable ($x_i^\dagger$)

1: Set $s := 1, m := 1, x_i^\dagger = \emptyset$.
2: Create the set of observations from the defined sub-sample class:

$$\mathcal{A}_m^{(s)} := \{x_i^{(s)} \in C_m^{(s)}\}\, \forall i,$$

where $\mathcal{A}_m^{(s)}$ has $N_m^{(s)}$ number of observations.
3: Create $Z(s,m)$, the set of possible working sample choice values, defined by Equation (36).
4: Draw $\mathcal{Z}_j \in Z(s,m), j = 1, \ldots, N_m^{(s)}$, with uniform probabilities given by Equation (37). Example: Let $C_3^{(2)} = [2.5, 4.5]$, $\mathcal{A}_m^{(s)} = \{3.5, 3.5, 3.5\}, N_m^{(s)} = 3$, $Z(s,m) = \{2.75, 3.25, 3.75, 4.25\}$, the uniform probabilities are $1/4$ for each choice value. Then we pick values with the defined probability from the set of $Z(s,m)$, 3 times with repetition, resulting in $\bigcup_{j=1}^{N_m^{(s)}} \mathcal{Z}_j = \{2.75, 3.25, 3.25\}$
5: Add these new values to $x_i^\dagger$,

$$x_i^\dagger := \left\{ x_i^\dagger, \bigcup_{j=1}^{N_m^{(s)}} \mathcal{Z}_j \right\}$$

6: If $s < S$, then $s := s + 1$ and go to Step 2.
7: If $s = S$, then $s := 1$ and set $m = m + 1$ and go to Step 2.

Now we show that the distribution of this new variable converges to the distribution of the

true underlying random variable (x) as we increase the number of sub-samples. That is

$$\lim_{S\to\infty} \Pr(x^\dagger < c) = \Pr(x < c) \qquad \forall c \in (a_l, a_u)$$

or

$$\lim_{S\to\infty} F_S(c) = F(c) \qquad \forall c \in (a_l, a_u),$$

where $F_S(c) = \Pr(x^\dagger < c)$ and $F(c) = \Pr(x < c)$. As $S \to \infty$, $\exists c_b^{WS} = c$ for any $c \in (a_l, a_u)$, by construction. Furthermore, for any c_b^{WS}, $\exists l \in [1, S]$, $m \in [1, M]$ such that $c_b^{WS} = c_m^{(l)}$. Also note that as $S \to \infty$, we need $N \to \infty$ as well. Now consider $\Pr(x^\dagger < c_b^{WS}) = \Pr(x^\dagger < c_m^{(l)})$, given $\Pr(x \in S) = 1/S$ and using equation (31) gives

$$\Pr(x^\dagger < c_m^{(l)}) = \frac{1}{S}\sum_{s=1}^{S} \Pr(x < c_m^{(l)} | x < c_m^{(s)}) \Pr(x < c_m^{(s)}).$$

Note that the summation over the different classes in Equation (31) is being replaced as we are considering the cumulative probability and that no value greater than $c_m^{(l)}$ will be used as a candidate in the working sample for c_b^{WS}. Under the shifting method, $c_m^{(s)} \leq c_m^{(l)}$ for $s < l$ and using the definition of conditional probability gives

$$\begin{aligned}
\Pr(x^\dagger < c_m^{(l)}) =& \frac{1}{S}\sum_{s=1}^{S} \Pr(x < c_m^{(l)}, x < c_m^{(s)}) \\
=& \frac{1}{S}\sum_{s=1}^{l} \Pr(x < c_m^{(l)}, x < c_m^{(s)}) + \frac{1}{S}\sum_{s=l+1}^{S} \Pr(x < c_m^{(l)}, x < c_m^{(s)}) \\
=& \frac{1}{S}\sum_{s=1}^{l} \Pr(x < c_m^{(s)}) + \frac{1}{S}\sum_{s=l+1}^{S} \Pr(x < c_m^{(l)}).
\end{aligned}$$

The last line follows from the fact that $\Pr(x < a, x < b) = \Pr(x < a)$ if $a < b$, and the construction of the shifting method allows us to always disentangle the two cases. Since l is fixed

$$\begin{aligned}
\Pr(x^\dagger < c_m^{(l)}) =& \frac{S-l-1}{S} \Pr(x < c_m^{(l)}) + \frac{1}{S}\sum_{s=1}^{l} \Pr(x < c_m^{(s)}) \\
\lim_{S\to\infty} \Pr(x^\dagger < c_m^{(l)}) =& \Pr(x < c_m^{(l)}).
\end{aligned}$$

This completes the proof.
In addition, to show that $x^\dagger$ shares the same distribution as x in the limit, we are able say something about the speed of convergence as well, as we increase the number of sub-samples (S). For each of the conditions in Equation (5.4), the corresponding expression is $o(1)$. To see this, note that $f(x)$ is a density, so the integral is less than 1. First, consider the case of

$s \neq 1$ and $m = 1$,

$$\begin{aligned}\frac{1}{S}\sum_{s=2}^{S}\frac{1}{s-1}\int_{C_1^{(s)}|C_b^{WS}\in C_1^{(s)}} f(x)\mathrm{d}x, &\leq \frac{1}{S}\sum_{s=2}^{S}\frac{1}{s-1} \\ &= \frac{1}{S}\sum_{s=1}^{S}\frac{1}{s} \\ &= \frac{1}{S}\int_1^S \frac{1}{s}ds \\ &= \frac{\log S}{S}.\end{aligned}$$

As $S \to \infty$, the ratio in the last line goes to 0. This is expected if the widths of the classes in the working sample go to zero. This is straightforward, while the probability that an observation belongs to a point is 0. The same derivations applies to the case when $m = M$. Now, consider the case of $1 < m < M$,

$$\begin{aligned}\frac{1}{S^2}\sum_{s=1}^{S}\int_{C_m^{(s)}|C_b^{WS}\in C_m^{(s)}} f(x)\mathrm{d}x &\leq \frac{1}{S^2}\sum_{s=1}^{S} 1 \\ &= \frac{1}{S},\end{aligned}$$

which also converges to 0 as $S \to \infty$, but at a faster rate than in the previous cases.
Note we cannot directly use $x_i^\dagger$ for estimation, while by design each individual observation only represents the conditional mean for the given sub-sample's class, and not the underlying variable's conditional expectation

$$\mathbb{E}\left(x_i^\dagger \in C_m^{(s)}\right) = \mathbb{E}\left(x_i^{(s)} \in C_m^{(s)}\right) \neq \mathbb{E}\left(x_i \in C_m^{(s)}\right).$$

However, while $F_S(x^\dagger)$ approximates the underlying distribution, we can use these values to calculate the sample conditional means for a given sub-sample class. Thus, the idea is to use this artificial distribution to calculate the conditional means and replace the class observations with these values.
Let $\hat{\pi}_\tau$ be the replacement estimator for the shifting method, where $\tau = 1, \ldots, S \times M$. Let us define

$$\hat{\pi}_\tau := \left(\sum_{i=1}^{N} \mathbf{1}'_{x_i^{(s)}\in C_m^{(s)}}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}'_{x_i^{(s)}\in C_m^{(s)}} x_i^\dagger. \tag{38}$$

Using the WLLN, it can be shown that the $\hat{\pi}_\tau$ for the sample conditional averages are in fact converging to the true underlying distribution's conditional expectations, thus

$$\hat{\pi}_\tau \to \mathbb{E}(x|x \in C_m^{(s)})$$

as $N, S \to \infty$ under the same assumptions as before.
Using this fact, we can replace $x_i^{(s)} \in C_m^{(s)}$ with $\hat{\pi}_\tau$ for each value, thus the working sample becomes the set of replacement estimators for each observation

$$x_{i,Shifting}^{WS} := \{\hat{\pi}_\tau\}.$$

We can also check the standard errors of the replacement estimator to have an idea how precise our results are:

$$x_i^\dagger = \boldsymbol{\pi}_\tau \mathbf{1}_{\{x_i^\dagger \in C_m^{(s)}\}} + \eta_i,$$

where $\boldsymbol{\pi}_\tau$ stands for the vector of $\pi_\tau, \forall\tau$. Using the standard OLS technique we can derive

$$\hat{\boldsymbol{\pi}}_\tau = \left(\mathbf{1}'_{x_i^\dagger \in C_m^{(s)}} \mathbf{1}_{\{x_i^\dagger \in C_m^{(s)}\}}\right)^{-1} \mathbf{1}'_{x_i^\dagger \in C_m^{(s)}} x_i^\dagger.$$

Under the standard OLS assumption, we can write

$$\sqrt{N^{WS}} \left(\hat{\boldsymbol{\pi}}_\tau - \mathbb{E}\left[\boldsymbol{\pi}_\tau\right]\right) \overset{\text{a}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Omega}_\tau\right),$$

where $\mathbb{E}\left(\boldsymbol{\pi}_\tau\right) = \mathbb{E}(x|x \in C_m^{(s)}), \forall\tau$. Furthermore, the variance of the OLS estimator is given by

$$\boldsymbol{\Omega}_\tau = V\left(\eta_i\right)\left(\mathbf{1}'_{x_i^\dagger \in C_m^{(s)}} \mathbf{1}_{\{x_i^\dagger \in C_m^{(s)}\}}\right)^{-1},$$

where $\hat{\pi}_\tau$ represents the first moments of the underlying random variable, thus using $x_{i,Shifting}^{WS}$ for estimation will result in a consistent estimator for β. Algorithm 6 describes how to create in practice a working sample with the shifting method.

Algorithm 6 Th shifting method – creation of working sample ($\Psi_{Shifting}(\cdot)$)

1: Set $s := 1, m := 1, \{x_i^{WS}, y_i^{WS}, w_i^{WS}\} = \emptyset$.
2: Calculate the sample conditional mean $\hat{\pi}_\tau$, for the given $C_m^{(s)}$ class, using Equation (38).
3: Add the conditional mean $\hat{\pi}_\tau$ and the observed values $y_j^{(s)}, w_j^{(s)}$ to the working sample,

$$x_i^{WS} := \left\{ x_i^{WS}, \bigcup_{j=1}^{N} \hat{\pi}_\tau \mid \left(x_j \in C_m^{(s)}\right) \right\}$$
$$y_i^{WS} := \left\{ y_i^{WS}, \bigcup_{j=1}^{N} y_j^{(s)} \mid \left(x_j \in C_m^{(s)}\right) \right\}$$
$$w_i^{WS} := \left\{ w_i^{WS}, \bigcup_{j=1}^{N} w_j^{(s)} \mid \left(x_j \in C_m^{(s)}\right) \right\}.$$

4: If $s < S$, then $s := s + 1$ and go to Step 2.
5: If $s = S$, then $s := 1$ and set $m = m + 1$ and go to Step 2.

5.5 Instrumental Variables Estimation

We can use the sub-sampling methods to create instrumental variables instead of replacing the x_i^* observations. In this case we need two survey questions in the case of cross-sectional data: one is the original question, which gives the choice variable (x_i^*) and another, which will be providing the IV. Usually it is not practical to ask the same questions with different possible choices, but we may refer to different time periods/locations/taste/etc., where the underlying distribution is the same. For example, in the case of the shifting method, we can ask 'How much have you used public transport *this* week?' '0-20%, 20-40%, 40-60%, 60-80%,

or 80-100%' as the original choices, with a second question: 'How much did you use public transport *last* week?' '0-10%, 10-30%, 30-50%, 50-70%, 70-90%, or 90-100%' for the IV.
There are some possible alternative specifications, which are out of the scope of this paper, but worth noting. One is, when we ask a more realistic question for the IV, which depends on the question of the original choice response, like: 'How much more or less have you used public transportation last week?', with answers such as '20% less, 10% less or equal, 10% more or equal, 20% more'. This is more realistic, however, it may induce an autoregressive process, which must be modeled for proper inference.

In the case of panel data, similar methods can be used as those outlined above for cross-sections. This, however, may give us some additional flexibility. Sub-sampling now can be used as follows: for the magnifying method, we can randomize the surveys assigned to each individual, this way ensuring variation in the response.[11] For the shifting method, we can ask each individual with randomly changing shifts.
With respect to the use of instrumental variables, we can ask the choice question at some t points and the IV question at some other t time points (the same question, same M, but with different class limits). The assumption needed in this case is that, the questions are paired in such a way that they belong to the same underlying distribution and, of course, an even number of type periods are needed.
Finally, for repeated cross-sections the same procedures can be applied as for panel data.

5.6 Monte Carlo Simulation Results

Next, we show the performance of our sub-sampling methods through some Monte Carlo simulations. Overall, the results are aligned with the theoretical ones (see Tables 2, 3, 4 and 5). The estimation biases are in general decreasing as we increase the number of observations and the number of sub-samples. The relative performance of the methods, however, essentially depends on two characteristics of the underlying distribution: Curvature (or the classes' conditional expectations relation to the choice values, $\mathbb{E}[x \mid x \in C_m]$ and z_m), and the fraction of the probability mass covered by the surveys (or what is the probability that a certain part of the distribution is neglected by the surveys: $\Pr(x < a_l)$ or $\Pr(x > a_u)$).
In order to disentangle these two effects (as can be seen in Table 1), we have used an exponential distribution with parameter 0.5, which provides a distribution with flat curvature (thus $\mathbb{E}[x \mid x \in C_m]$ and z_m are close to each other) and a normal distribution with $\mu_x = 0, \sigma_x^2 = 0.2$, where the curvature is quite steep (thus $\mathbb{E}[x \mid x \in C_m]$ and z_m are far from each other). Furthermore, we have checked the truncated case, where the probability mass is completely covered by the surveys and the censored case, where there is a non-negligible part of the probability mass which cannot be utilized for the estimation.

$f(\cdot; a_l, a_u)$	$\mathbb{E}[x \mid x \in C_m]$ and z_m	$\int_{a_l}^{a_u} f(\cdot)$
$Exp(0.5; 0, 1)$	close to each other	complete mapping (100%)
$Exp(0.5; 0, \infty)$		weak mapping (39%)
$\mathcal{N}(0, 0.2; -1, 1)$	far from each other	complete mapping (100%)
$\mathcal{N}(0, 0.2; -\infty, \infty)$		good mapping (99%)

Table 1: Distributions used for the underlying random variable x.

[11] With the magnifying method dropping observations may generate a missing data problem.

As theory suggests, in general, the bias decreases as we increase the number of observations (N) and the number of sub-samples (S). Next, let us go through the main simulation results method by method.

- **Magnifying method – Truncated case**
  - $Exp(0.5; 0, 1)$: The bias decreases in S and N. The increase of M has no significant effect, because the conditional expected values and choice values are close to each other. The standard errors are decreasing in N, but slightly increasing in S. This is due to the fact that the share of directly transferable observations is decreasing in S. This implies more replacement estimators, which increases the standard errors of the estimated coefficient. The absolute bias therefore first decreases, then starts to increase as the effect of standard errors starts to dominate. *Overall, with flat curvature and complete mapping of the probability mass, S/N should be above 0.01%, and M can be small.*
  - $\mathcal{N}(0, 0.2; -1, 1)$: The bias decreases in S and N. There is a significant decrease in the bias if we increase M, because the conditional expected values and choice values are not close to each other. All other results are the same as in the exponential case above. *Overall, with steep curvature and complete mapping of probability mass, S/N should be above 0.01%, and increasing M can significantly reduce the bias.*
- **Magnifying method – Censored case**
  - $Exp(0.5; 0, \infty)$ and $\mathcal{N}(0, 0.2; -\infty, \infty)$: The bias first decreases, but then it starts to increase again. This is due to the fact there are only a few observations to calculate the replacement estimator values for non-directly transferable observations. This lack of precision introduces bias during the estimation of β. The number of observations is radically decreasing as S increases and the standard errors are increasing in S. The absolute bias is mainly driven by the standard errors. *Overall, without complete mapping of the probability mass, the main driver of the bias is the number of observations in the working sample. With fewer sub-samples, we can decrease the absolute bias, but using too many sub-samples is counter-productive. $S/N < 0.01\%$ is a good rule of thumb here as well.*
- **Shifting method – Truncated case**
  - $Exp(0.5; 0, 1)$: The bias decreases in S and N. Using larger S will not help reduce the bias on the same scale as in the magnifying method due to the boundary classes' slow convergence. On the other hand, using more choices (M) will reduce the bias. It is interesting to note that the standard errors remain unchanged as S increases. The absolute bias decreases and gets smaller than in the benchmark case (with no sub-sampling) if we have a large amount of observations. *Overall, with complete mapping of the probability mass and flat curvature distribution, increasing M helps to reduce the bias, and increasing S also decreases it, but at a much slower rate. We need a large amount of observations in order to reduce the standard errors as well. As a rule of thumb we may use a smaller number of sub-samples.*
  - $\mathcal{N}(0, 0.2; -1, 1)$: The bias decreases in S and N. Using larger S helps to significantly reduce the bias similarly to using larger M. This makes the approximation much better at the boundaries. Standard errors are the same as in the benchmark

case, and does not change as S or M increases. The absolute bias is decreasing in N and S. *Overall, with complete mapping of the probability mass and steep curvature distribution, increasing M and S helps to reduce the bias more effectively. The absolute bias is also decreasing in N, M and S.*

- **Shifting method – Censored case**
  - $Exp(0.5; 0, \infty)$: The bias is decreasing in N and S, but it decreases more slowly in S, because the main drivers of the bias are the boundary classes. Increasing M will help to significantly reduce the bias. The standard errors and the absolute bias behave similarly as in the truncated case. Note that the number of observations used for the estimation is much larger than in the magnifying case! *Overall, without complete mapping of the probability mass, with flat curvature distribution, using few sub-samples will eliminate the main bias, and increasing M can help to reduce it even more.*
  - $\mathcal{N}(0, 0.2; -\infty, \infty)$: The bias is decreasing in N and S. Now, the boundary classes only take up a small fraction of the probability mass of the distribution, so these classes have a much smaller role in driving the bias, resulting in a much faster bias reduction. Furthermore, increasing the number of choices decreases the bias further. The standard errors, however, are slightly larger than in the benchmark case. The absolute bias is decreasing in N, M and S as well. *Overall, without complete mapping of the probability mass, with steep curvature distribution, increasing both S and M will significantly reduce the bias.*

- **Comparison of the Magnifying and Shifting methods**
  - $Exp(0.5; \cdot)$: In the truncated case the performances are very similar. In the censored case, the *bias* is smaller for the magnifying method when $S/N < 0.01\%$. In all other cases, the shifting method outperforms the magnifying one. This is due to the fact that the magnifying method drops many more observations by construction.
  - $\mathcal{N}(0, 0.2; \cdot)$: In the truncated case, the magnifying method decreases the bias much more efficiently than the shifting method. For the censored case, the results are very similar to the exponential distribution if M is small. However, the shifting method becomes better if we use larger M.

- **Survey design implications**
  - When some features of the underlying distribution are known or some assumptions about them can be made (about the curvature and the probability mass's distribution), then the most suitable method, sub-sample size, etc. can be picked for a given application:
    - With steep curvature you should use larger M.
    - When only a small fraction of probability mass is covered by the surveys, you must choose your main aim. If you intend to minimize the absolute bias, use shifting; if you prefer a small bias but are not worried about a more noisy estimator, then use the magnifying method.
  - In the case of shifting and/or censoring, extra choices on the boundaries can help to improve the performance of the methods:

* In the case of shifting, you may add an extra small class in the boundaries, which will result in a faster bias reduction.
* In the case of censoring, there is a clear cut from where to drop the observations, which enables us to control the censoring and thus reduce the number of dropped observations.

		Magnifying method - used as $x_{i,All}^{WS}$							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0182	0.0032	0.0020	0.0015	0.1341	-0.0026	-0.1147	-0.2728
	N=100,000	-0.0185	0.0020	0.0012	0.0016	0.1342	-0.0029	-0.0151	-0.0473
	N=500,000	-0.0190	0.0004	0.0004	0.0008	0.1339	-0.0008	-0.0013	-0.0182
absbias	**N=10,000**	0.0415	0.0807	0.0902	0.0929	0.1342	0.3105	0.4537	0.5312
	N=100,000	0.0208	0.0284	0.0312	0.0320	0.1339	0.0971	0.1676	0.2264
	N=500,000	0.0191	0.0121	0.0138	0.0140	0.1342	0.0438	0.0760	0.1049
se	**N=10,000**	0.0489	0.1024	0.1147	0.0445	0.0785	0.3872	0.5653	0.6019
	N=100,000	0.0163	0.0355	0.0392	0.0401	0.0137	0.1218	0.2108	0.2794
	N=500,000	0.0073	0.0152	0.0172	0.0175	0.0061	0.0554	0.0961	0.1301
numObs	**N=10,000**	10,000	10,000	10,000	10,000	10,000	696	212	181
	N=100,000	100,000	100,000	100,000	100,000	100,000	6,874	1,693	941
	N=500,000	500,000	500,000	500,000	500,000	500,000	34,348	8,267	4,310
		Shifting method - used as x_i^{WS}							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0182	0.0023	0.0025	0.0027	0.1341	0.0864	0.0843	0.0861
	N=100,000	-0.0185	0.0019	0.0021	0.0021	0.1342	0.0859	0.0809	0.0801
	N=500,000	-0.0190	0.0018	0.0017	0.0016	0.1339	0.0865	0.0815	0.0805
absbias	**N=10,000**	0.0415	0.0703	0.0701	0.0701	0.1342	0.1811	0.1642	0.1630
	N=100,000	0.0208	0.0238	0.0236	0.0235	0.1339	0.0926	0.0873	0.0869
	N=500,000	0.0191	0.0103	0.0103	0.0103	0.1342	0.0866	0.0816	0.0806
se	**N=10,000**	0.0489	0.0879	0.0878	0.0879	0.0785	0.2078	0.1891	0.1864
	N=100,000	0.0163	0.0297	0.0294	0.0293	0.0137	0.0683	0.0633	0.0632
	N=500,000	0.0073	0.0130	0.0130	0.0130	0.0061	0.0308	0.0283	0.0280
numObs	**N=10,000**	10,000	10,000	10,000	10,000	10,000	5,071	5,334	5,387
	N=100,000	100,000	100,000	100,000	100,000	100,000	50,704	53,162	53,491
	N=500,000	500,000	500,000	500,000	500,000	500,000	253,492	265,711	267,270

Table 2: $Exp\,(0.5)\,, Supp = [0, 1]$, **M=3**; BM: Benchmark (see Table 9)

		Magnifying method - used as $x_{i,ALL}^{WS}$							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0798	-0.0051	-0.0015	-0.0005	-0.0552	-0.0053	-0.1419	-0.3182
	N=100,000	-0.0800	-0.0055	-0.0002	-0.0003	-0.0552	-0.0053	-0.0188	-0.0751
	N=500,000	-0.0803	-0.0057	-0.0002	0.0000	-0.0554	-0.0054	-0.0037	-0.0160
absbias	**N=10,000**	0.0798	0.0264	0.0318	0.0356	0.0553	0.0669	0.1699	0.3198
	N=100,000	0.0800	0.0100	0.0109	0.0120	0.0552	0.0226	0.0461	0.0863
	N=500,000	0.0803	0.0063	0.0050	0.0054	0.0554	0.0104	0.0194	0.0301
se	**N=10,000**	0.0224	0.0329	0.0401	0.0447	0.0220	0.0842	0.1534	0.1485
	N=100,000	0.0074	0.0111	0.0136	0.0151	0.0074	0.0282	0.0540	0.0721
	N=500,000	0.0033	0.0051	0.0063	0.0068	0.0031	0.0117	0.0241	0.0349
numObs	**N=10,000**	10,000	10,000	10,000	10,000	10,000	946	241	195
	N=100,000	100,000	100,000	100,000	100,000	100,000	9,381	1,983	1,069
	N=500,000	500,000	500,000	500,000	500,000	500,000	46,891	9,730	4,953
		Shifting method							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0811	-0.0244	-0.0240	-0.0242	-0.0552	0.0106	0.0067	0.0049
	N=100,000	-0.0810	-0.0246	-0.0241	-0.0241	-0.0552	0.0103	0.0069	0.0062
	N=500,000	-0.0811	-0.0246	-0.0242	-0.0242	-0.0554	0.0102	0.0071	0.0065
absbias	**N=10,000**	0.0811	0.0288	0.0285	0.0286	0.0553	0.0346	0.0323	0.0316
	N=100,000	0.0810	0.0246	0.0241	0.0241	0.0552	0.0137	0.0115	0.0112
	N=500,000	0.0811	0.0246	0.0242	0.0242	0.0554	0.0104	0.0076	0.0072
se	**N=10,000**	0.0224	0.0251	0.0253	0.0253	0.0220	0.0421	0.0401	0.0395
	N=100,000	0.0071	0.0083	0.0083	0.0082	0.0074	0.0134	0.0127	0.0126
	N=500,000	0.0033	0.0036	0.0037	0.0037	0.0031	0.0059	0.0056	0.0055
numObs	**N=10,000**	10,000	10,000	10,000	10,000	10,000	8,064	8,250	8,280
	N=100,000	100,000	100,000	100,000	100,000	100,000	80,631	82,428	82,654
	N=500,000	500,000	500,000	500,000	500,000	500,000	403,167	412,108	413,203

Table 3: $\mathcal{N}\,(0, 0.2)\,, Supp = [-1, 1]$, **M=3**; BM: Benchmark (see Table 10)

		Magnifying method - used as $x_{i,All}^{WS}$							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0074	0.0037	0.0004	-0.0002	0.1304	-0.0063	-0.1343	-0.2709
	N=100,000	-0.0072	0.0014	0.0013	0.0012	0.1307	-0.0038	-0.0156	-0.0494
	N=500,000	-0.0078	0.0005	0.0006	0.0007	0.1303	-0.0011	-0.0033	-0.0099
absbias	**N=10,000**	0.0394	0.0841	0.0908	0.0919	0.1305	0.2472	0.4654	0.5010
	N=100,000	0.0145	0.0291	0.0314	0.0325	0.1307	0.0809	0.1599	0.2147
	N=500,000	0.0090	0.0124	0.0138	0.0140	0.1303	0.0365	0.0746	0.1027
se	**N=10,000**	0.0489	0.1068	0.1155	0.1164	0.0437	0.3081	0.5710	0.5767
	N=100,000	0.0165	0.0364	0.0395	0.0405	0.0135	0.1025	0.2048	0.2647
	N=500,000	0.0073	0.0155	0.0173	0.0177	0.0059	0.0458	0.0938	0.1278
numObs	**N=10,000**	10,000	10,000	10,000	10,000	10,000	804	218	183
	N=100,000	100,000	100,000	100,000	100,000	100,000	7,962	1,750	955
	N=500,000	500,000	500,000	500,000	500,000	500,000	39,775	8,547	4,383
		Shifting method - used as x_i^{WS}							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0074	0.0020	0.0018	0.0016	0.1304	0.0269	0.0292	0.0311
	N=100,000	-0.0072	0.0016	0.0015	0.0015	0.1307	0.0218	0.0209	0.0212
	N=500,000	-0.0078	0.0007	0.0006	0.0006	0.1303	0.0221	0.0218	0.0218
absbias	**N=10,000**	0.0489	0.0674	0.0678	0.0680	0.1305	0.1112	0.1057	0.1053
	N=100,000	0.0165	0.0230	0.0230	0.0230	0.1307	0.0394	0.0380	0.0381
	N=500,000	0.0073	0.0099	0.0099	0.0099	0.1303	0.0247	0.0243	0.0243
se	**N=10,000**	0.0843	0.0837	0.0844	0.0846	0.0437	0.1359	0.1294	0.1280
	N=100,000	0.0279	0.0285	0.0286	0.0285	0.0135	0.0444	0.0425	0.0425
	N=500,000	0.0131	0.0124	0.0124	0.0124	0.0059	0.0200	0.0195	0.0193
numObs	**N=10,000**	10,000	10,000	10,000	10,000	10,000	6,379	6,552	6,589
	N=100,000	100,000	100,000	100,000	100,000	100,000	63,814	65,404	65,619
	N=500,000	500,000	500,000	500,000	500,000	500,000	319,061	326,956	327,963

Table 4: $Exp\,(0.5)\,, Supp = [0, 1]$, **M=5**; BM: Benchmark (see Table 9)

		Magnifying method - used as $x_{i,All}^{WS}$							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0311	-0.0005	-0.0002	-0.0004	-0.0097	-0.0044	-0.1536	-0.3267
	N=100,000	-0.0313	-0.0004	0.0001	0.0000	-0.0099	-0.0010	-0.0178	-0.0794
	N=500,000	-0.0315	-0.0008	-0.0000	0.0000	-0.0100	-0.0010	-0.0039	-0.0164
absbias	**N=10,000**	0.0328	0.0274	0.0324	0.0368	0.0195	0.0649	0.1780	0.3282
	N=100,000	0.0313	0.0093	0.0111	0.0121	0.0106	0.0204	0.0460	0.0901
	N=500,000	0.0315	0.0042	0.0050	0.0054	0.0100	0.0095	0.0200	0.0306
se	**N=10,000**	0.0226	0.0343	0.0404	0.0461	0.0234	0.0815	0.1523	0.1495
	N=100,000	0.0078	0.0116	0.0139	0.0152	0.0074	0.0257	0.0544	0.0747
	N=500,000	0.0033	0.0052	0.0063	0.0068	0.0033	0.0118	0.0243	0.0346
numObs	**N=10,000**	10,000	10,000	10,000	10,000	10,000	973	243	196
	N=100,000	100,000	100,000	100,000	100,000	100,000	9,643	1,994	1,072
	N=500,000	500,000	500,000	500,000	500,000	500,000	48,204	9,771	4,965
		Shifting method - used as x_i^{WS}							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0311	-0.0052	-0.0050	-0.0052	-0.0097	0.0049	0.0038	0.0034
	N=100,000	-0.0313	-0.0049	-0.0047	-0.0047	-0.0099	0.0054	0.0038	0.0035
	N=500,000	-0.0315	-0.0052	-0.0049	-0.0049	-0.0100	0.0053	0.0037	0.0035
absbias	**N=10,000**	0.0328	0.0198	0.0198	0.0197	0.0195	0.0248	0.0241	0.0240
	N=100,000	0.0313	0.0077	0.0076	0.0076	0.0106	0.0089	0.0082	0.0081
	N=500,000	0.0315	0.0054	0.0052	0.0052	0.0100	0.0058	0.0046	0.0045
se	**N=10,000**	0.0226	0.0242	0.0243	0.0243	0.0234	0.0307	0.0300	0.0299
	N=100,000	0.0078	0.0082	0.0083	0.0083	0.0074	0.0098	0.0095	0.0095
	N=500,000	0.0033	0.0036	0.0036	0.0036	0.0033	0.0044	0.0043	0.0043
numObs	**N=10,000**	10,000	10,000	10,000	10,000	10,000	9,089	9,156	9,168
	N=100,000	100,000	100,000	100,000	100,000	100,000	90,884	91,525	91,606
	N=500,000	500,000	500,000	500,000	500,000	500,000	454,421	457,602	457,994

Table 5: $\mathcal{N}\,(0, 0.2)\,, Supp = [-1, 1]$, **M=5**; BM: Benchmark (see Table 10)

Next, let us expand our approacha little toward the use of instrumental variables. We can in fact construct an instrument, using the sub-sampling methods. In practice, there might be several ways to carry this out. One possibility is to split the sample into two parts: The first keeps the 'original' surveys without any sub-sampling; in the second, sub-sampling is performed by any of the methods outlined above. Another way is to ask the same question twice: one refers to the 'original' question, and the second is coming from a sub-sampling method.

The success of the instrumental estimation depends on the correlation between the 'original' variable and the instrumental variable. We used the same simulation setup as outlined above,

but now carried out IV estimation. Our results suggest that using a larger number of choice classes (M) significantly increases the correlation between the variables, resulting in better estimates in general. For the truncated case, the IV estimation resulted in the same or worse results in terms of bias and absolute bias, except when the shifting method is employed on the normal distribution. Here we get higher correlation values and smaller bias. When we used the shifting method in the censored case, the bias became significantly smaller with larger M and S. *Overall, with the magnifying method or when the support of the underlying distribution is controlled (truncated case), using IV estimation will not result in (much) better results. However, when we use the shifting method and there is censoring, the IV estimation becomes much better, especially with larger M, independently of the underlying distribution.*

6 Extensions

At the end of the paper it is worth talking briefly about some extensions.
First, the lines between discretized ordered choice type observations and continuous ones can be quite blurred. Let us take again our example in model 4. But ask the question directly, say by moving an indicator on the screen between 0 and 100. The 'usual' way to approach these types of observations is to consider them with a measurement error, i.e., by adding a white noise random term. We argue here that in some cases another approach may be more realistic. Say, that in our example one observation is 63%. This means that it can be considered as a choice type observation that falls into the [65%–70%] class (if we assume 20 classes, i.e., 5% 'precision' for the observations), with random class boundaries and random class midpoints. That is the 'real' answer to the question in this case is 'around 65%–70%'.
Going back to Equation (9), but instead of assuming that each choice class is represented by its midpoint z_m, $m = 1, \ldots, M$, we can assume that the responses are randomly distributed on the domain of each class C_m, $\zeta_m \sim f_{\zeta_m}$. Even if assume that this distribution is known, which is in fact the expectation of the kind of bias the respondents might make when answering, it can be shown that the OLS will only be consistent in some very special cases. Namely, only when the random variable ζ_m has the same expected value as the underlying random variable x, conditional on each class C_m, which in general is quite an unlikely scenario.
Let us now return to the definition of the class boundaries in Equation (8) and consider the case when they all are random variables on disjoint sub-intervals of $[a, b]$ rather than constants. Let $\delta_m \sim f_{\Delta_m}$ for $m = 0, \ldots, M$ be the independent random boundaries and the expected value of the intermediate boundaries be $\mathbb{E}(\Delta_m) = c_m$ for $m = 1, \ldots, M-1$. Therefore, now we have random classes of the following form $C_1 = [\Delta_0, \Delta_1), C_2 = [\Delta_1, \Delta_2), \ldots, C_M = [\Delta_{M-1}, \Delta_M]$. Note that if the distribution of Δ_0 and Δ_M is not trivial (Dirac–delta), then their expected value does not match the lower and upper bound of the whole domain a and b. Now, very similarly to the case of the stochastic class midpoint, $\hat{\beta}^*_{OLS}$ is only consistent in the unlikely case when the expected value of the class value Z_m matches the expected value of the underlying random variable x conditional on each class C_m, $m = 1, \ldots, M$.
Let us realize here that this case covers the 'rounding up' problem as well, when an answer, say 65%, is in fact the respondent's rounded up value. This 65 can be considered as a stochastic class midpoint, with random class boundaries, where the width of a class is dependent on the researcher's confidence in the answer.

There is another issue as well. There is some evidence in the behavioural literature, that the answers to a question may depend on the way the question is asked (see, e.g., Diamond and Hausman (1994), Haisley et al. (2008) and Fox and Rottenstreich (2003)). Let us call

this *perception effect*. This is present regardless whether sub-sampling has been performed or not. However, with sub-sampling, there is a way to tackle this issue, much akin to the way a similar problem has been dealt with in the panel data literature.
More specifically, in our case the definition of the classes may affect the way the participants respond to the survey question. A way to formalize such effects is by redefining the discretization of x_i as follows

$$x_i^{**} = \begin{cases} z_1 & \text{if} \quad c_0 < x_i + B_s < c_1 \\ & \vdots \\ z_m & \text{if} \quad c_{m-1} < x_i + B_s < c_M, \end{cases} \tag{39}$$

where B_s denotes the perception effect for sub-sample s, $s = 1, \ldots, S$. Let $\tilde{x}_i^*$ and $\tilde{x}_i^{**}$ denote the observations in the working sample that derived from x_i^* and x_i^{**}, respectively. Following the derivation of the working sample from the methods above, all observations in the working samples can be expressed as

$$\tilde{x}_i^{**} = \tilde{x}_i^* + B_s \tag{40}$$

given the corresponding x_i^* and x_i^{**} came from the sub-sample s. Thus, the regression

$$y_i = \beta \tilde{x}_i^{**} + u_i \tag{41}$$

is equivalent to

$$y_i = \beta \tilde{x}_i^* + B_s \beta + u_i. \tag{42}$$

Rewrite the above in matrix form using standard definitions gives

$$\mathbf{y} = \tilde{\mathbf{x}}^* \beta + \mathbf{D}\mathbf{B}\beta + \mathbf{u}, \tag{43}$$

where $\mathbf{B} = (B_1, \ldots, B_S)'$ and $\mathbf{D}$ is a $N \times S$ zero-one matrix that extracts the appropriate elements from $\mathbf{B}$. So the estimation of β can be done in the spirit of a fixed effect estimator. Define the usual residual maker, $\mathbf{M_D} = \mathbf{I}_N - \mathbf{D}\left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'$, then

$$\hat{\beta} = \left(\tilde{\mathbf{x}}^{*'}\mathbf{M_D}\tilde{\mathbf{x}}^*\right)^{-1}\tilde{\mathbf{x}}^{*'}\mathbf{M_D}\mathbf{y} \tag{44}$$

is a consistent estimator of β following the similar argument of the standard fixed effect estimator in the panel data literature.
We also need to slightly modify the replacement estimator in order for the above to hold. The main problem is to keep track of the perception effects. This means we need to keep track of which sub-sample each observation is coming from when estimating the conditional averages. This means

$$\hat{\pi}_{\chi,s} = \left(\sum_{i=1}^{N} \mathbf{1}_{\{\tilde{x}_i^{**} \in C_\chi, \tilde{x}_i^{**} \in s\}}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}_{\{\tilde{x}_i^{**} \in C_\chi, \tilde{x}_i^{**} \in s\}} \tilde{x}_i^{**} \tag{45}$$

and as $N \to \infty$

$$\hat{\pi}_{\chi,s} = \mathbb{E}(x_i | x_i \in C_\chi) + B_s + o_p(1).$$

This shows that equation (44) provides a valid replacement estimator in the presence of perception effects.
While the discussion above focus on the case with one regressor, the generalisation to K regressors is straightforward. Perhaps a more interesting question is the presence of perception

effects over different m as well. This can also be incorporated in principle, by replacing B_s with B_{sm} for $s = 1, \ldots, S$ and $m = 1, \ldots, M$. Therefore, this particular setup does not just allow for perception effects due to different sub-samples, but rather, it provides a framework to investigate different types of perception effects. This would be an interesting avenue of future research in this area.

7 Conclusion

This paper has investigated the effects of using discretized ordered choice variables in a linear regression model when the underlying variable is not observed. This situation often arises in survey data when continuous variables, such as income for example, are not captured directly, but rather, are replaced by a set of M choices. Unlike other studies in the literature, our approach has considered the more realistic case when the underlying distribution of the unobserved explanatory variables is unknown and the values of each choice can be arbitrarily assigned. With fixed M, the results show that using the discretized ordered choices as explanatory variables in a linear regression will lead to biased and inconsistent parameter estimates. The well-known techniques to create consistent estimators require information from the distributions of the underlying explanatory variables, which are presumed to be unknown, and therefore cannot be applied here.
This paper proposes a novel survey construction by sub-sampling. Using the fact that the discretized variables approaches their unobserved continuous counterparts when M grows, the proposed approach essentially replaces the requirement of M being sufficiently large with the more standard scenario where the number of individuals, N is very large, utilizing different questionnaires for each sub-sample. Theoretical results show that these techniques will lead to a proper mapping of the true underlying distribution. Monte Carlo simulations show that the methods put forward work reasonably well, and may have significant implications on the future of survey design.

Appendix A: Some Monte Carlo Simulation Results on the Bias

Let us use the same very simple model as in Section 3.
The basic setup of the Monte Carlo experiment is, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\beta = 0.5$, x is generated as Uniform, Normal, Exponential, and Weibull distributions with several different parameter setups. One thousand Monte Carlo experiments ($mc = 1, \ldots, 1000$) were run for each setup, for sample sizes ($N =$) 10,000; 100,000 and 500,000 and different σ_ε^2 variances. When generating x^*, observation outside the support, whenever relevant, would be discarded (truncated approach), or assigned to the limit of the class (censored approach). We report the *average bias* ($\bar{\beta}_{mc} = \sum_{mc}(\hat{\beta}_{mc} - \beta)/1000$), the *average absolute bias* ($\sum_{mc} |\hat{\beta}_{mc} - \beta|/1000$), and the *standard error* of the $\hat{\beta}$ estimated parameter ($\sqrt{\sum_{mc}(\hat{\beta}_{mc} - \bar{\beta}_{mc})^2/1000}$). The Kullback–Leibler proximity/discrepancy index (Kullback and Leibler (1951), Kullback (1959), Kullback (1987)) has also been calculated to appreciate how different a given distribution is from the uniform:

$$KL = \int p(x) \log \frac{p(x)}{f(x)} dx,$$

where $p(x)$ is the uniform distribution and $f(x)$ is the relevant truncated or censored normal distribution.

Uniform Distribution

		Uniform[-1,1]				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**
bias	**N=10,000**	-0.0005	-0.0005	-0.0005	-0.0005	-0.0006
	N=100,000	-0.0008	-0.0010	-0.0008	-0.0008	-0.0008
	N=500,000	-0.0008	-0.0010	-0.0010	-0.0010	-0.0010
absbias	**N=10,000**	0.0322	0.0307	0.0303	0.0302	0.0300
	N=100,000	0.0103	0.0100	0.0098	0.0097	0.0097
	N=500,000	0.0049	0.0049	0.0049	0.0048	0.0048
se	**N=10,000**	0.0406	0.0390	0.0384	0.0382	0.0380
	N=100,000	0.0129	0.0124	0.0123	0.0122	0.0122
	N=500,000	0.0060	0.0059	0.0058	0.0058	0.0058
		Uniform[0,1]				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**
bias	**N=10,000**	-0.0008	-0.0008	-0.0008	-0.0008	-0.0008
	N=100,000	-0.0006	-0.0007	-0.0006	-0.0006	-0.0006
	N=500,000	-0.0010	-0.0012	-0.0012	-0.0011	-0.0012
absbias	**N=10,000**	0.0298	0.0295	0.0293	0.0292	0.0292
	N=100,000	0.0100	0.0098	0.0098	0.0098	0.0098
	N=500,000	0.0044	0.0044	0.0044	0.0044	0.0044
se	**N=10,000**	0.0375	0.0372	0.0369	0.0369	0.0369
	N=100,000	0.0126	0.0123	0.0123	0.0123	0.0123
	N=500,000	0.0054	0.0054	0.0054	0.0054	0.0054
		Uniform[0,10]				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**
bias	**N=10,000**	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
	N=100,000	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
	N=500,000	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
absbias	**N=10,000**	0.0031	0.0030	0.0029	0.0029	0.0029
	N=100,000	0.0010	0.0010	0.0010	0.0010	0.0010
	N=500,000	0.0005	0.0004	0.0004	0.0004	0.0004
se	**N=10,000**	0.0038	0.0037	0.0037	0.0037	0.0037
	N=100,000	0.0013	0.0012	0.0012	0.0012	0.0012
	N=500,000	0.0006	0.0005	0.0005	0.0005	0.0005

Table 6: **Uniform distribution:** $\beta = 0.5, \sigma_\varepsilon^2 = 5$

From Table 6 the unbiasedness and consistency (in sample size) of the OLS estimator can clearly be seen in the case of the uniform distribution, similarly to the, somewhat slower, convergence in M. We have also done simulations with different σ_ε^2 and β, where the same results hold. For smaller σ_ε^2, the bias is smaller, for different β the results are almost exactly the same.
Next, let us turn our attention to some other distributions.

Normal Distribution

From Table 7 it is clear that the OLS estimator is biased and inconsistent, with a negative bias, as predicted by the theory, both in the case of truncation and censoring. Although the theory suggests that intercept picks up some of the bias, in practice the difference between with and without intercept – in this case – is small, approximately 3-5%. It also interesting to note that the Kullback-Liebler index gives a good indication of the bias (see Table 8). The bias tends to be smaller where this index is small, and vice versa.

	Bias					
	Truncated			Censored		
	N=10,000	**N=100,000**	**N=500,000**	**N=10,000**	**N=100,000**	**N=500,000**
$\sigma_x^2 = 0.1$	-0.0593	-0.0603	-0.0607	-0.0582	-0.0567	-0.0575
$\sigma_x^2 = 0.2$	-0.0320	-0.0323	-0.0329	-0.0110	-0.0101	-0.0103
$\sigma_x^2 = 0.3$	-0.0224	-0.0223	-0.0226	0.0272	0.0283	0.0280
$\sigma_x^2 = 0.4$	-0.0176	-0.0171	-0.0173	0.0619	0.0630	0.0628
$\sigma_x^2 = 0.5$	-0.0142	-0.0139	-0.0141	0.0938	0.0950	0.0948
$\sigma_x^2 = 0.6$	-0.0118	-0.0118	-0.0120	0.1239	0.1248	0.1245
$\sigma_x^2 = 0.7$	-0.0102	-0.0103	-0.0105	0.1517	0.1527	0.1524
$\sigma_x^2 = 0.8$	-0.0092	-0.0091	-0.0093	0.1783	0.1791	0.1788
$\sigma_x^2 = 0.9$	-0.0082	-0.0082	-0.0084	0.2032	0.2042	0.2039
$\sigma_x^2 = 1$	-0.0074	-0.0075	-0.0077	0.2271	0.2280	0.2278
	Abs. Bias					
	Truncated			Censored		
	N=10,000	**N=100,000**	**N=500,000**	**N=10,000**	**N=100,000**	**N=500,000**
$\sigma_x^2 = 0.1$	0.0730	0.0603	0.0607	0.0710	0.0568	0.0575
$\sigma_x^2 = 0.2$	0.0485	0.0326	0.0329	0.0417	0.0151	0.0106
$\sigma_x^2 = 0.3$	0.0416	0.0233	0.0226	0.0435	0.0285	0.0280
$\sigma_x^2 = 0.4$	0.0382	0.0188	0.0173	0.0651	0.0630	0.0628
$\sigma_x^2 = 0.5$	0.0363	0.0162	0.0141	0.0941	0.0950	0.0948
$\sigma_x^2 = 0.6$	0.0350	0.0147	0.0121	0.1239	0.1248	0.1245
$\sigma_x^2 = 0.7$	0.0339	0.0136	0.0107	0.1517	0.1527	0.1524
$\sigma_x^2 = 0.8$	0.0335	0.0129	0.0097	0.1783	0.1791	0.1788
$\sigma_x^2 = 0.9$	0.0331	0.0125	0.0089	0.2032	0.2042	0.2039
$\sigma_x^2 = 1$	0.0326	0.0121	0.0084	0.2271	0.2280	0.2278
	SE					
	Truncated			Censored		
	N=10,000	**N=100,000**	**N=500,000**	**N=10,000**	**N=100,000**	**N=500,000**
$\sigma_x^2 = 0.1$	0.0661	0.0212	0.0098	0.0662	0.0210	0.0088
$\sigma_x^2 = 0.2$	0.0520	0.0165	0.0079	0.0518	0.0156	0.0068
$\sigma_x^2 = 0.3$	0.0473	0.0150	0.0072	0.0457	0.0137	0.0059
$\sigma_x^2 = 0.4$	0.0451	0.0144	0.0068	0.0421	0.0128	0.0055
$\sigma_x^2 = 0.5$	0.0436	0.0139	0.0067	0.0403	0.0124	0.0053
$\sigma_x^2 = 0.6$	0.0428	0.0136	0.0065	0.0387	0.0120	0.0051
$\sigma_x^2 = 0.7$	0.0419	0.0134	0.0064	0.0379	0.0117	0.0050
$\sigma_x^2 = 0.8$	0.0415	0.0132	0.0064	0.0368	0.0115	0.0049
$\sigma_x^2 = 0.9$	0.0412	0.0132	0.0063	0.0360	0.0114	0.0047
$\sigma_x^2 = 1$	0.0408	0.0131	0.0063	0.0356	0.0113	0.0047

Table 7: **Truncated and Censored Normal Distributions, estimated without intercept,** $M = 5, \beta = 0.5, \sigma_\varepsilon^2 = 5, Supp = [-1, 1]$

	Truncated	**Censored**
$\sigma_x^2 = 0.1$	0.7396	0.7407
$\sigma_x^2 = 0.2$	0.2287	0.2536
$\sigma_x^2 = 0.3$	0.1091	0.1783
$\sigma_x^2 = 0.4$	0.0634	0.1829
$\sigma_x^2 = 0.5$	0.0414	0.2109
$\sigma_x^2 = 0.6$	0.0291	0.2463
$\sigma_x^2 = 0.7$	0.0216	0.2835
$\sigma_x^2 = 0.8$	0.0167	0.3203
$\sigma_x^2 = 0.9$	0.0132	0.3558
$\sigma_x^2 = 1$	0.0197	0.3899

Table 8: **Kullback-Leibler ratio: Uniform vs. Truncated/Censored Normal with different σ_x^2 values,** $a = -1, b = 1$

Exponential Distribution and Weibull Distributions

We carried out a large number of simulations with different parametrisations for both distributions. In Table 9 we report the bias from the exponential distribution, which highlights the effect of censoring. Although we do no observe large bias with truncation, when the choices are censored the bias increases dramatically.
From Table 11, the main takeaway is that, as expected, there is no convergence in the sample size, while the convergence speed in M is 'slow' and depends heavily on the shape of the distribution. Also, the results about the Kullback-Liebler index (not reported here) are very similar to those obtained for the normal distribution, i.e., a larger index implies systematically a larger bias.
We have also tried several different distributions and parameterisation, and the main take away is very similar.

		$Exp[\lambda], Supp = [0,1]$									
		Truncated					Censored				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**	**M=3**	**M=5**	**M=10**	**M=20**	**M=50**
bias	**N=10,000**	-0.0182	-0.0074	-0.0027	-0.0015	-0.0011	0.1341	0.1304	0.1235	0.1190	0.1160
	N=100,000	-0.0185	-0.0072	-0.0025	-0.0014	-0.0011	0.1342	0.1307	0.1239	0.1193	0.1163
	N=500,000	-0.0190	-0.0078	-0.0032	-0.0020	-0.0017	0.1339	0.1303	0.1235	0.1190	0.1160
absbias	**N=10,000**	0.0415	0.0394	0.0388	0.0388	0.0388	0.1342	0.1305	0.1237	0.1191	0.1162
	N=100,000	0.0208	0.0145	0.0133	0.0131	0.0131	0.1342	0.1307	0.1239	0.1193	0.1163
	N=500,000	0.0191	0.0090	0.0064	0.0060	0.0059	0.1339	0.1303	0.1235	0.1190	0.1160
se	**N=10,000**	0.0489	0.0489	0.0489	0.0490	0.0490	0.0445	0.0437	0.0427	0.0422	0.0419
	N=100,000	0.0163	0.0165	0.0164	0.0164	0.0164	0.0137	0.0135	0.0131	0.0130	0.0129
	N=500,000	0.0073	0.0073	0.0073	0.0073	0.0073	0.0061	0.0059	0.0058	0.0057	0.0057

Table 9: **Exponential distribution:** $\beta = 0.5, \sigma_\varepsilon^2 = 5, \lambda = 0.5$

		$\mathcal{N}\left(\mu_x, \sigma_x^2\right), Supp = [-1, 1]$									
		Truncated					Censored				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**	**M=3**	**M=5**	**M=10**	**M=20**	**M=50**
bias	**N=10,000**	-0.0798	-0.0311	-0.0078	-0.0017	0.0000	-0.0552	-0.0097	0.0088	0.0120	0.0120
	N=100,000	-0.0800	-0.0313	-0.0079	-0.0017	0.0000	-0.0552	-0.0099	0.0084	0.0115	0.0114
	N=500,000	-0.0803	-0.0315	-0.0081	-0.0020	-0.0003	-0.0554	-0.0100	0.0082	0.0113	0.0112
absbias	**N=10,000**	0.0798	0.0328	0.0198	0.0188	0.0187	0.0553	0.0195	0.0198	0.0209	0.0209
	N=100,000	0.0800	0.0313	0.0092	0.0066	0.0064	0.0552	0.0106	0.0092	0.0117	0.0117
	N=500,000	0.0803	0.0315	0.0081	0.0032	0.0028	0.0554	0.0100	0.0082	0.0113	0.0112
se	**N=10,000**	0.0224	0.0226	0.0234	0.0234	0.0234	0.0220	0.0228	0.0230	0.0229	0.0228
	N=100,000	0.0074	0.0078	0.0080	0.0080	0.0080	0.0074	0.0074	0.0074	0.0074	0.0074
	N=500,000	0.0033	0.0033	0.0034	0.0034	0.0034	0.0031	0.0033	0.0033	0.0033	0.0032

Table 10: **Normal distribution:** $\beta = 0.5, \sigma_\varepsilon^2 = 1, \mu_x = 0, \sigma_x^2 = 0.2$

		$Weibull\,[b, c]\,, Supp = [0, 1]$									
		Truncated					Censored				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**	**M=3**	**M=5**	**M=10**	**M=20**	**M=50**
bias	**N=10,000**	-0.0369	-0.0128	-0.0031	-0.0010	-0.0004	1.8197	1.7475	1.6828	1.6486	1.6278
	N=100,000	-0.0369	-0.0130	-0.0033	-0.0011	-0.0005	1.8209	1.7487	1.6840	1.6498	1.6289
	N=500,000	-0.0371	-0.0131	-0.0035	-0.0013	-0.0007	1.8197	1.7475	1.6828	1.6486	1.6278
absbias	**N=10,000**	0.0371	0.0178	0.0144	0.0142	0.0141	1.8197	1.7475	1.6828	1.6486	1.6278
	N=100,000	0.0369	0.0131	0.0056	0.0049	0.0048	1.8209	1.7487	1.6840	1.6498	1.6289
	N=500,000	0.0371	0.0131	0.0038	0.0024	0.0022	1.8197	1.7475	1.6828	1.6486	1.6278
se	**N=10,000**	0.0174	0.0179	0.0179	0.0179	0.0179	0.0492	0.0474	0.0458	0.0450	0.0445
	N=100,000	0.0058	0.0060	0.0060	0.0060	0.0060	0.0154	0.0148	0.0144	0.0141	0.0140
	N=500,000	0.0026	0.0027	0.0027	0.0027	0.0027	0.0071	0.0069	0.0066	0.0065	0.0064

Table 11: **Weibull distribution:** $\beta = 0.5, \sigma_\varepsilon^2 = 0.5, b = 1, c = 0.5$

Appendix B: Summary of the Notation Used in the Paper

Scalars:
N – number of individuals in the sample
T – number of time period in the sample (panel case)
a_l – lower boundary point for distribution's ($f(\cdot)$) support
a_u – upper boundary point for distribution's ($f(\cdot)$) support
μ or μ_i – first moment for distribution $f(\cdot)$ or $f_i(\cdot)$
M – number of possible choice values for a questionnaire
z_m – choice value of class m
c_m – m'th class's lower boundary point
β – parameter for DOC variable
γ – parameter for control variables
K – number of DOC variables (matrix notations)
J – number of control variables (matrix notations)
B – number of working sample classes
S – number of sub-samples
$N^{(s)}$ – number of observations in sub-sample s
$z_m^{(s)}$ – choice value of class m in sub-sample s
$c_m^{(s)}$ – s'th sub-sample, m'th class's lower boundary point
c_b^{WS} – working sample b'th class's lower boundary point
h – working sample's class widths
Δ – size of shift for the shfting method

Running indexes
i – refers to individual $i = 1, \ldots, N$, and in some places it is a running index.
t – refers to time $t = 1, \ldots, T$
m – refers to class $m = 1, \ldots, M$

k – refers to a DOC variables in matrix formulation, $k = 1, \ldots, K$
j – refers to a control variables in matrix notation, $j = 1, \ldots, J$, and in some places it is a running index.
b – working sample classes, $b = 1, \ldots, B$
s – sub-sample index
i_m – running index, where m is the indication in which class that observation is (M consistency)
m_i – i-th observation in the m-th class (M consistency)

Random variables
X or x – true choices with distribution $X \sim f(\cdot)$ (unknown)
X^* – discretized choice (DOC), with distribution $\psi(X)$ (observed)
$\hat{\beta}$ – parameter estimate for β with OLS (estimate)
$\hat{\gamma}$ – parameter estimate for γ with OLS (estimate)
$\bar{x}$ – sample average of the underlying variable x (not observed)
$\bar{x}^*$ – sample average of the observed discretized variable x^* (estimate)
x^{WS} – working sample (concept)
$\hat{\pi}_\chi$ – replacement estimator for non-directly transferable observations (estimate)
y^{tr}, x^{tr} – artificially truncated variables of the original r.v. (concept)
$\hat{\pi}_\tau$ – replacement estimator for shifting method (estimate)

Individual observations of random variables
x_i – true choice values for individual i (not observed)
x_i^* – discretized choice values (DOC) for individual i (observed)
y_i – outcome variable's values for individual i (observed)
w_i – control variable's values for individual i (observed)
ϵ_i – model disturbance term
u_i – idiosyncratic disturbance term for DGP (not observed)
N_m – number of observations in class m (observed)
ξ_i – error due to discretization $\xi_i = x_i - x_i^*$ (not observed)
υ^m – conditional distribution for errors of class m, formally: $\upsilon^m \overset{d}{=} \xi_i | C_m$ (not observed)
x_m – conditional distribution for x_i within class m, formally: $x_m \overset{d}{=} x_i | C_m$ (not observed)
x^m – sum of the true observed values in class m, formally: $x^m = \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\} x_i}$ (not observed)
ξ^m – sum of the errors in class m, formally: $\xi^m = \sum_{i=1}^{N} \mathbf{1}_{\{\xi_i \in C_m\} \xi_i}$ (not observed)
$x_i^{(s)}$ – discretized choice values (DOC) for individual i in sub-sample s (observed)
N^{WS} – number of observations in the working-sample (observed)
$N_m^{(s)}$ – number of observations in sub-sample s in class $C_m^{(s)}$ (observed)
x_i^{WS} – working-samples DOC observations (observed)
$x_{i,DTO}^{WS}$ – magnifying method's working sample, constructed by only the directly transferable observations (observed)
N_{DTO}^{WS} – number of observations in the magnifying method's 'DTO' working sample. (observed)
$x_{i,NDTO}^{WS}$ – magnifying method's working sample, constructed by only the directly transferable observations (observed)
η_i – error component from models to get $\hat{\pi}_\chi$ or $\hat{\pi}_\tau$ (observed)
$x_i^\dagger$ – artificial variable created during the shifting method (constructed)

$x^{WS}_{i,Shifting}$ – shifting method's working sample (constructed)

Functions
$f(\cdot)$ – probability distribution function
$\psi(\cdot)$ – discretization function $\psi(x_i) = x_i^*$
$\mathbf{1}_{\{\cdot\}}$ – indicator function, which takes 1 if the condition in the subscript is true, otherwise 0
$F(\cdot)$ – cdf of x
$U(\cdot)$ – Uniform distribution
$\psi^{(s)}(\cdot)$ – discretization function for sub-sample s
$\Psi(\cdot)$ – merging function
$||\cdot||$ – width of a class (or euclidean distance)
$Z(s,m)$ – set 'creator' function: given a sub-sample class, creates a set of choice values, which lies in the interval of the working-sample
$\mathcal{F}^{\dagger}$ – assign choice values from $Z(s,m)$ to each observation $x_i^{(s)} \in C_m^{(s)}$, with a given (uniform) probability
$\mathcal{F}^{WS}$ – assign estimated values $\hat{\pi}_\tau$ to each observation $x_i^{(s)} \in C_m^{(s)}$

Intervals
C_m – m'th class
$C_m^{(s)}$ – s sub-sample's, m'th class
C_b^{WS} – working sample's, b'th class

Sets
ζ – set of classes, which contains the directly transferable observations
C_χ – set of classes, which contains the non-directly transferable observations
ζ^{tr} – ζ without the first and last class
$\mathcal{A}_m^{(s)}$ – set for observations $x_i^{(s)}$ which are in class $C_m^{(s)}$

Matrix notations
$\mathbf{y}$ – $y_i, N \times 1$
$\mathbf{X}$ – $(x_{1,i}, \ldots, x_{k,i}, \ldots, x_{K,i}), N \times K$
$\mathbf{W}$ – $(w_{1,i}, \ldots, w_{j,i}, \ldots, w_{K,i}), N \times J$
$\boldsymbol{\varepsilon}$ – $\epsilon_i, N \times 1$
$\boldsymbol{\beta}$ – $\beta_k, K \times 1$
$\boldsymbol{\gamma}$ – $\gamma_j, J \times 1$
$\mathbf{z}_k$ – $(z_{1,i}, \ldots, z_{m,i}, \ldots, z_{M,i}), 1 \times M$
$\mathbf{Z}$ – $\text{diag}(\mathbf{z}_{1,i}, \ldots, \mathbf{z}_{k,i}, \ldots, \mathbf{z}_{K,i}), MK \times K)$
$\mathbf{e}_{ki}$ – is the indicator vector for k'th DOC variable
$\mathbf{E}$ – matrices for the indicator vectors, $MK \times N$
$\mathbf{X}^* = \mathbf{E}'\mathbf{Z}$
$\mathbf{M}_W$ – residual maker
$\mathbf{q}_{kl}$ – typical block element in $\mathbf{E}\mathbf{E}'$
Ω – region for integration $[c_{km-1}, c_{km}] \times [c_{ln-1}, c_{ln}]$
a_{kl} – auxiliary variable for $\mathbf{Z}'\mathbf{E}\mathbf{X}$
$\Omega_{\mathbf{X}}$ – sample space of x_k and x_l
ω_{ij} – (i,j) element in $\mathbf{M}_W$
g_{kl} – auxiliary variable for proof Eq. 26

h_{kl} – auxiliary variable for proof Eq. 28
u_{li} – auxiliary variable for proof Eq. 28
Panel
$\boldsymbol{\beta}_W$ – within estimator for panel
$\mathbf{D}_N$ – individual fixed effect
$\mathbf{M}_{D_N}$ – panel projection matrix
Sub-sampling
$\hat{\boldsymbol{\pi}}_\chi$ – vector of replacement estimator for magnifying method
$\boldsymbol{\Omega}\chi$ – asymptotic standard errors for $\hat{\boldsymbol{\pi}}_\chi$
$\hat{\boldsymbol{\pi}}_\tau$ – vector of replacement estimator for shifting method
$\boldsymbol{\Omega}_\tau$ – asymptotic standard errors for $\hat{\boldsymbol{\pi}}_\tau$

Appendix C: Technical Details

Derivation of equation (9)

$$
\begin{aligned}
\hat{\beta}^*_{OLS} &= \left(x^{*'}x^*\right)^{-1}(x^{*'}y) \\
&= \frac{z_1\left(\sum_{i=1}^{N_1} y_i\right) + z_2\left(\sum_{i=N_1+1}^{N_1+N_2} y_i\right) + \cdots + z_M\left(\sum_{i=N-N_M+1}^{N_M} y_i\right)}{N_1 z_1^2 + N_2 z_2^2 + \cdots + N_M z_M^2} \\
&= \frac{z_1\left(\sum_{i=1}^{N_1} \beta x_i + u_i\right) + \cdots + z_M\left(\sum_{i=N-N_M+1}^{N_M} \beta x_i + u_i\right)}{N_1 z_1^2 + \cdots + N_M z_M^2} \\
&= \frac{z_1\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_1\}}(\beta x_i + u_i)\right] + \cdots + z_M\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_M\}}(\beta x_i + u_i)\right]}{N_1 z_1^2 + \cdots + N_M z_1^2} \\
&= \frac{\sum_{m=1}^{M} z_m\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2} \\
&= \frac{\sum_{m=1}^{M}\left[a_l + (2m-1)\frac{a_u - a_l}{2M}\right]\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m\left[a_l + (2m-1)\frac{a_u - a_l}{2M}\right]^2}.
\end{aligned}
$$

Derivation of Equation (10)

$$
\begin{aligned}
\mathbb{E}\left(\hat{\beta}^*_{OLS}\right) &= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \left(\beta(x_i^* + \xi_i) + u_i\right)\right]}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \left[\beta\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i^* + \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i\right) + \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i\right]}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i^*}{\sum_{m=1}^{M} N_m z_m^2}\right\} + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&\qquad + \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m N_m \upsilon^m}{\sum_{m=1}^{M} N_m z_m^2}\right\}.
\end{aligned}
$$

Derivation of Equation (11)

$$
\begin{aligned}
\operatorname*{plim}_{N\to\infty} \hat{\beta}^*_{OLS} &= \operatorname*{plim}_{N\to\infty} \frac{\sum_{m=1}^{M} z_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} (\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2} \\
&= \frac{\sum_{m=1}^{M} z_m \left[\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} (\beta x_i + u_i)\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\sum_{m=1}^{M} z_m \left[\operatorname{plim}_{N\to\infty} \beta \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\beta \sum_{m=1}^{M} z_m \left[\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m},
\end{aligned}
$$

Derivation of Equation (12)

$$
\begin{aligned}
\operatorname*{plim}_{N\to\infty} \left(\hat{\beta}^*_{OLS} - \beta\right) &= \frac{\beta\left(\sum_{m=1}^{M} z_m \left[\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i\right] - \sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m\right)}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\beta \sum_{m=1}^{M} z_m \left[\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} (x_i - x_i^*)\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\beta \sum_{m=1}^{M} z_m \left[\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m},
\end{aligned}
$$

Derivation of Equation (13)

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^{*}_{OLS}-\beta\right) &= \frac{\operatorname{plim}_{N\to\infty}\beta\sum_{m=1}^{M} z_m\xi^m}{\operatorname{plim}_{N\to\infty}\sum_{m=1}^{M} z_m^2 N_m} \\
&= \frac{\operatorname{plim}_{N\to\infty} O(N)\beta\sum_{m=1}^{M} z_m\xi^m/N_m}{\operatorname{plim}_{N\to\infty} O(N)\sum_{m=1}^{M} z_m^2} \\
&= \frac{\beta\sum_{m=1}^{M} z_m \operatorname{plim}_{N\to\infty}\xi^m/N_m}{\sum_{m=1}^{M} z_m^2}O(N) \\
&= \frac{\beta\sum_{m=1}^{M} z_m\left\{\mathbb{E}(x_m)-z_m\right\}}{\sum_{m=1}^{M} z_m^2}O(N),
\end{aligned}$$

the last step in the above derivation can simply be obtained from the definition of the plim operator, i.e., for any $\varepsilon > 0$ given

$$\begin{aligned}
&\operatorname*{plim}_{N\to\infty}\xi^m = \mathbb{E}(X_m)-z_m \\
&\quad\Longleftrightarrow \lim_{N\to\infty}\Pr\left(\left|\xi^m-\left\{\mathbb{E}(X_m)-z_m\right\}\right|>\varepsilon\right) \\
&\qquad= \lim_{N\to\infty} F_{\xi^m}\left(-\varepsilon+\mathbb{E}(X_m)-z_m\right)\left[1-F_{\xi^m}\left(\varepsilon+\mathbb{E}(X_m)-z_m\right)\right]=0.
\end{aligned}$$

The convergence holds, because for any given $\delta > 0$, there is a threshold N_0 for which the term in the limit becomes less than δ. This can be seen from $F_{\xi^m}(\cdot)$ being close to a degenerate distribution above a threshold number of observations N_0, or intuitively, since the variance of the sequence of random variables ξ^m collapses in N, its probability limit equals its expected value.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2002). Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *The Quarterly journal of economics*, 117(4):1231–1294.

Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological methodology*, pages 83–118.

Berkson, J. (1980). Minimum chi-square, not maximum mikelihood! *The Annals of Statistics*, 8:457–487.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. CRC Press.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7(3):249–253.

Connor, R. J. (1972). Grouping for testing trends in categorical data. *Journal of the American Statistical Association*, 67(339):601–604.

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, 52(280):543–547.

Diamond, P. A. and Hausman, J. A. (1994). Contingent valuation: Is some number better than no number? *American Economic Review*, 8(4):45–64.

Fox, C. R. and Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, 14(3):195–200.

Frey, B. S. and Stutzer, A. (2002). What can economists learn from happiness research? *Journal of Economic Literature*, 40(2):402–435.

Givon, M. M. and Shapira, Z. (1984). Response to rating scales: a theoretical model and its application to the number of categories problem. *Journal of Marketing Research*, 21(4):410–419.

Haisley, E., Mostafa, R., and Loewenstein, G. (2008). Subjective relative income and lottery ticket purchases. *Journal of Behavioral Decision Making*, 21:283–295.

Heath, Y. and Gifford, R. (2002). Extending the theory of planned behavior: Predicting the use of public transportation. *Journal of Applied Social Psychology*, 32(10):2154–2189.

Johnson, D. R. and Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, pages 398–407.

Knack, S. and Keefer, P. (1995). Institutions and economic performance: cross-country tests using alternative institutional measures. *Economics & Politics*, 7(3):207–227.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley & Sons; Republished by Dover Publications in 1968; reprinted in 1978.

Kullback, S. (1987). Letter to the Editor: The Kullback-Liebler distance. *The American Statistician*, 41:340–341.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Lagakos, S. (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine*, 7(1-2):257–274.

Mauro, P. (1995). Corruption and growth. *The Quarterly Journal of Economics*, 110(3):681–712.

Méndez, F. and Sepúlveda, F. (2006). Corruption, growth and political regimes: Cross country evidence. *European Journal of political economy*, 22(1):82–98.

Santos, A., McGuckin, N., Nakamoto, H. Y., Gray, D., and Liss, S. (2011). Summary of travel trends: 2009 national household travel survey. Technical report.

Srinivasan, V. and Basu, A. K. (1989). The metric quality of ordered categorical data. *Marketing Science*, 8(3):205–230.

Stutzer, A. (2004). The role of income aspirations in individual happiness. *Journal of Economic Behavior & Organization*, 54(1):89–109.

Taylor, J. M. and Yu, M. (2002). Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis*, 83(1):248–263.

Wansbeek, T. and Meijer, E. (2000). *Measurement Error and Latent Variables in Econometrics*. North-Holland Elsevier.

Wansbeek, T. and Meijer, E. (2001). Measurement error and latent variables. In Baltagi, B. H., editor, *A Companion to Theoretical Econometrics*, chapter 8, pages 162–179. John Wiley & Sons.