A Simulation Approach to Estimate Inclusion Probabilities for PIAAC Germany

Siegfried Gabler, Sabine Häder & Jan-Philipp Kolb
GESIS – Leibniz Institute for the Social Sciences

Abstract
In PIAAC *(Programme for the International Assessment of Adult Competencies)* inclusion probabilities have to be known for every respondent at each sampling stage in all participating countries. However, in some cases it is not possible to calculate inclusion probabilities for a sample survey analytically – although the underlying design is probabilistic. In such cases, simulation studies can help to estimate inclusion probabilities and thus ensure that the necessary basis for the calculation of design weights is available. In this section, we present a Monte Carlo simulation using the German sample data. During the selection process for PIAAC Germany an error had occurred. Because of that, it was not possible to determine the inclusion probabilities analytically. Therefore a simulation study with 10,000 runs of the erroneous selection process was set up. As a result it was possible to compute the inclusion probabilities for the sample of PIAAC Germany.

Keywords: Monte Carlo simulation, inclusion probabilities, sampling for PIAAC Germany

1 Sampling for Comparative Surveys

Cross-national surveys have become very popular during the last decades. The reason for this is the multiplicity of questions that can be answered with the help of this kind of data. Lynn et al. (2006, p. 10) identify three main objectives for cross-national surveys, such as PIAAC (*Programme for the International Assessment of Adult Competencies)*:

a) Comparisons of estimates of parameters for different countries
b) Rankings of countries on different dimensions such as averages or totals
c) Estimates for a supra-national region such as the European Union aggregated from estimates of different countries.

Sampling strategies have to ensure the equivalence and/or combinability of these estimates. For this, both sample designs and estimation strategies have to be chosen carefully.

Kish (1994, p. 173) gives a theoretic basis for the application of sample designs in cross-cultural surveys:

"Sample designs may be chosen flexibly and there is no need for similarity of sample designs. Flexibility of choice is particularly advisable for multinational comparisons, because the sampling resources differ greatly between countries. All this flexibility assumes probability selection methods: known probabilities of selection for all population elements."

Following this idea, an optimal sample design for cross-national surveys should consist of the best random sampling practice used in each participating country. The choice of a special sample design depends on the availability of frames, experience, but also mainly on the costs in different countries. Once the survey has been conducted, and adequate estimators have been chosen, the resulting values become comparable. To ensure this comparability, design weights have to be computed for each country. For this, the inclusion probabilities of every respondent at each stage of selection must be known and recorded. Furthermore, the inclusion probabilities for non-respondents must also be recorded at every stage where the necessary information is available to have possibilities for the compensation of the nonresponse (see Helmschrott/Martin in this volume) by suitable weighting procedures.

Direct correspondence to
Siegfried Gabler, GESIS – Leibniz Institute for the Social Sciences,
PO Box 12 21 55, 68072 Mannheim, Germany
E-mail: siegfried.gabler@gesis.org

Acknowledgment: We thank Silke Martin, GESIS, for her comments to a draft of this paper.

In the following section basic requirements for the PIAAC sampling are explained. For the German survey the sample design is described in detail. Furthermore, the erroneous procedure applied by the survey institute during the selection of the PIAAC gross sample is presented. Then, the simulation setup is demonstrated. The simulation results are evaluated in section 3. Finally, conclusions are drawn in the last section.

2 Basic Requirements and Sample Design Features of PIAAC Germany

Derived from the principles of sampling for cross-cultural surveys mentioned above the international PIAAC-Consortium expressed the following basic requirements for sample designs in the participating countries (OECD 2009, p. 6):

- Clustered and stratified designs were advised since these design features ensure both cost efficiency and variation of socio-demographic variables.
- A variety of designs could be applied because different countries have different access to frames and varying experience with the application of sample designs. Self -weighting designs of dwelling units or individuals should be preferred.
- All countries had to use probability based sampling methods at each stage of selection.
- The target population was defined as non-institutionalized adults between the ages of 16 and 65 (inclusive).

2.1 Sample Design and Sample Selection in Germany[1]

The sample design can be described as stratified two-stage probability design.

Stage 1

The PSUs (municipalities = Primary Sampling Units) were explicitly stratified by the variables federal states (Bundesländer), administrative regions (Regierungsbezirke), districts (Kreise) and ten grades of urbanization.

The sample points within the PSUs consisted of a pre-specified number of individuals to be selected at the second sampling stage from the person register held by the municipalities. In the vast majority of cases, sample points corresponded to one municipality only, while very large municipalities were drawn more than once and therefore covered more than one sample point. The number of sample points was

1 For a more detailed description of the PIAAC sampling procedure, see Zabal et al. (2014) as well as Lynn et al. (2014).

Table 1 Allocation of the sample sizes to municipalities

Number of inhabitants	Sample size
- 99,999	60
100,000 - 499,999	120
500,000 and more	180

set to 320. This resulted in the selection of 277 municipalities. In every municipality the sample spread over the whole area, i.e. there was no local clustering. Some larger municipalities had more than one sample point. If there were *k* sample points in a municipality the number of persons selected was multiplied by *k* (see table 1).

The PSUs were allocated proportionally to the size of the target population within each stratum. As only whole numbers can be selected as PSUs, the exact number of sample points to be selected from each stratum was determined using the procedure for unbiased controlled rounding by Cox (1987). This so-called Cox-Algorithm assures that the cell totals as well as the marginal totals of the allocation table remain nearly unchanged by the rounding procedure so that the structural properties of the population are not lost due to rounding (see Lynn et al., 2014).

Stage 2

To ensure an equal selection process in each selected municipality the following instructions were sent to the registration offices:

A simple systematic random sample of individuals, with a random start number and a sampling interval had to be drawn. The sample size in each municipality depended on its population size according to table 1. Personal information such as name, address, age, gender, nationality had to be provided for each selected individual by the registration offices. Data delivered by them were checked for different aspects. For more details see Zabal et al. (2014, pp 51).

All individuals (= person addresses) per point were allocated to a matrix defined by the variables age (six groups) and gender (see Sample Frame in figure 1). With an Iterative Proportional Fitting procedure (IPF) 32 individuals per sample point were selected from the frame under the constraint to meet the age and gender distribution in the federal state (for the result of the selection process, see Allocation Matrix in figure 1).

The selection of the individuals from the pool of addresses per community was done systematically with a selection interval. Unfortunately, in this process a programming as well as a sorting error did occur. The length of the interval was computed by “number of cases on the sampling frame” divided by “number of

Official Statistics		
Sex	Age Group	Freq[1]
m	16 – 19	256511
m	20 – 29	660824
m	**30 – 39**	**672291**
m	40 – 49	939744
m	50 – 59	727218
m	60 – 65	323953
f	16 – 19	243785
f	20 – 29	648787
f	30 – 39	668194
f	40 – 49	899595
f	50 – 59	726175
f	60 – 65	330558
total		7097635

Sample Frame		
Sex	Age Group	Point 163
m	16 – 19	2
m	20 – 29	5
m	**30 – 39**	**6**
m	40 – 49	5
m	50 – 59	3
m	60 – 65	2
f	16 – 19	0
f	20 – 29	11
f	30 – 39	3
f	40 – 49	5
f	50 – 59	6
f	60 – 65	2
total		50

Allocation Matrix		
Sex	Age Group	Point 163
M	16 – 19	1
M	20 – 29	3
m	**30 – 39**	**4**
M	40 – 49	3
M	50 – 59	2
M	60 – 65	1
F	16 – 19	0
F	20 – 29	7
f	30 – 39	2
f	40 – 49	3
f	50 – 59	5
f	60 – 65	1
total		32

(1) Source: Statistisches Bundesamt Genesis Table 12411-0012 at 31.12.2009

Sex	**Age**	**selected**
m	31	1
m	32	0
m	32	1
m	34	1
m	35	0
m	38	1

Figure 1 Example for the functionality of the optimization algorithm (variables sex and age)

cases to be selected" (see Allocation Matrix) and was not rounded. For the start number, a random number between 0 and the length of the interval was generated. If the start number was between 0 and 1.5, the program rounded always to 1. If the start number was at least 1.5, the program rounded to the closest integer number (based on commercial rounding). From a statistical point of view it would have been correct to always round up to the closest integer number.

The example from figure 1 illustrates the optimization algorithm. The adjustment algorithm always results in the same solution of number of elements to be selected in each cell (unrounded, exact allocation). In the example this value is equal to 3.97. The rounding of this number to the closest integer number is done randomly. 3.97 is rounded to 4 in 97% of the cases and to 3 in 3% of the cases. Due to the random procedure in the example, the exact number of persons to be selected from males, age 30 to 39 in sample point 163 was set to 4. Thus, the interval length is 6/4. In a next step the algorithm computed a random start number between 0 and 1.5, which was here 1.1. If the algorithm had worked correctly, 1.1 would have been rounded in some occasions to 1 and in other occasions to 2. However, due to

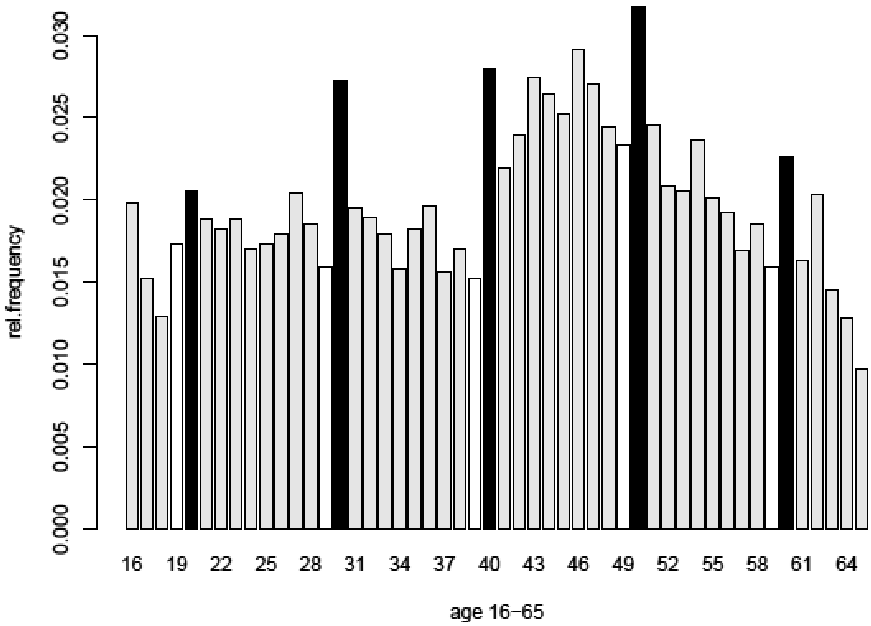

Figure 2 Age distribution (PIAAC sample unweighted) resulting from the erroneous algorithm

the error in the algorithm program, a random number of 1.1 would have always been rounded down to 1, and thus the chance for the first person on the frame to be selected was higher.

Summary of the selection process in the example
Number of cases due to IPF: 3.97 (rounded to 4)
Length of interval: 6/4 = 1.5
Start value: 1.1
Selected units unrounded: 1.1, 1.1+1.5, 1.1+1.5+1.5, 1.1+1.5+1.5+1.5
Selected units after commercial rounding: 1, 3, 4, 6

According to common practice of the survey institute, the pool of addresses on the sample frame is randomly ordered by the Fisher-Yates Shuffle before the sample is drawn. This procedure was done with the pool of addresses for the PIAAC sample as well. However, for some quality control checks the sample frame was sorted by age and this sorting order was unfortunately kept for the drawing. This mistake in accordance with the programming error (rounding error of the start number and thus higher chances of selection for the first person on the frame) both had a very negative impact (see Figure 2): Some age-groups (those ending with 0) are over-represented, others (in particular those ending with 9) are under-represented.

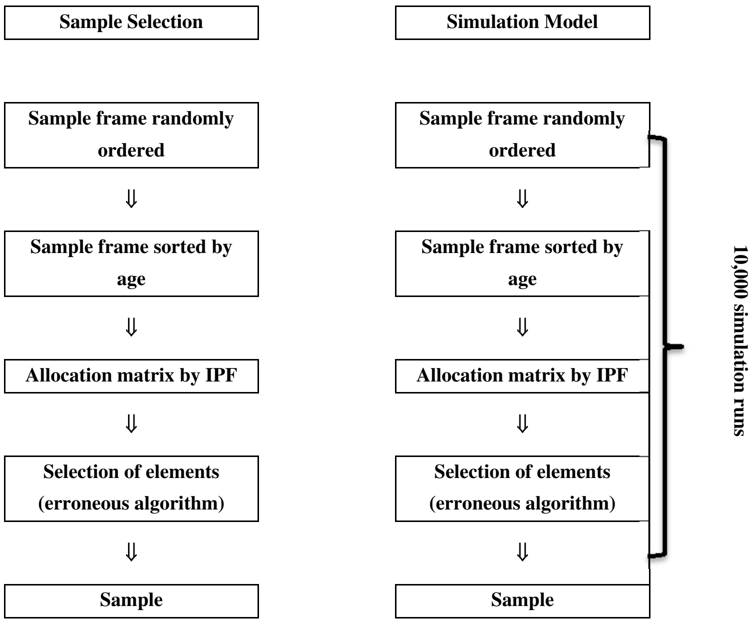

Figure 3 Comparison of the erroneous sample selection process and the simulation model

2.2 Simulation of the Selection Procedure

As a consequence, the gross sample has no longer the characteristic of equal selection probabilities for all elements. Instead, the selection probabilities for persons varied. Since it was too time consuming to model the incorrect selection probabilities, we decided to compute them through simulations, i.e. through a repetition of the erroneous optimization algorithm for 10,000 times. The idea was to rebuild the erroneous sampling procedure. Thus, the selection of the individuals from the pool of addresses was repeated 10,000 times. This was the basis of the simulation. The simulation model is described in figure 3.

In our model the random shuffle was repeated each time before a new iteration occurred – as it was done in the original optimization process. Thus, the following steps were repeated 10,000 times:

- The sample frame was randomly ordered according to the Fisher-Yates Shuffle.
- The sample frame was sorted by age.
- The sample was drawn.

Again, in order to estimate the selection probability of each element on the sample frame, a count was made of how many times an element was selected in each of the 10,000 samples.

The results of the simulation study are presented in the next section.

3 Evaluation of the PIAAC Sampling Procedure

To evaluate the results of the simulation – it was a Monte Carlo simulation – some theoretical considerations have to be explained first. The sample selection of the PIAAC sample is the result of a random experiment. For the PIAAC sample the random experiment to generate the PIAAC sample consists of several random experiments. If the random experiment would have been conducted as planned by the survey institute, the result would be that each person of the population would have the same selection probability. Due to the described error in the course of the random experiment, the equal probability is interfered, but not the general character of a probability sample as a result of a random experiment, i.e.

- that the random experiment could be repeated unlimited times, and
- that the results of the random experiment, i.e. possible samples, may be different, meaning that the result of the random experiment cannot be predicted with certainty for each iteration.

For the $r = 10{,}000$ simulation runs it was never the case that an element was not at all selected. Thus, it can be concluded that the selection probabilities are all positive. The error in the course of the random experiment affected only a part of the whole random experiment.

Ideally, as mentioned above, the selection of the PIAAC sample should have led to equal selection probabilities. This condition is no longer given due to the error in the course of the random experiment. The question is which selection probabilities have been generated by the selection process. Due to the error and the fact that the whole sampling procedure is built on random processes and sort sequences, it is very difficult and time-consuming for either GESIS or the survey institute to reflect this error in formulas in order to exactly calculate the selection probabilities.

The delivery deadline for the sample to the Consortium was dated shortly after the problem was noticed. The calculations needed to be carried out within a short time span. Furthermore, the amount of time that is necessary for one simulation run is not negligible, but cannot be determined exactly. It was thus necessary to find a trade-off between calculation time and an adequate number of simulation runs. The number of 10,000 simulation runs was the highest number which could be achieved under the prevailing circumstances. It was important that in every single simulation run the erroneous algorithm performed like in reality. Therefore, it was neither possible nor justified to program the algorithm more effectively. If one regards the high sampling fraction, it is clear that a higher number was not necessary in this coherence.

However, the r = 10,000 samples generated by the simulation, provide an excellent basis for a sufficiently precise estimation of the true selection probabilities. This will be justified as follows:

We observe the event of selecting a person into a sample. The true selection probability given a single iteration of the random experiment is P. In r independent repetitions of the random experiment the person is selected in, say $p \cdot r$ samples. Thus, according to statistical rules for large r and not too small p (since p is expected to differ not too much from the theoretical inclusion probability)

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{r}};\ p + 1.96\sqrt{\frac{p(1-p)}{r}} \right]$$

includes the true value P with a probability of 95%. Due to $1.96\sqrt{\frac{p(1-p)}{r}} < \sqrt{\frac{1}{r}} = 0.01$ with r = 10,000, the value p computed by simulation only deviates *at maximum* in the third decimal place from the true value P, most likely even later. This error seems to be negligible in practice.

The experiment is repeated very often and following the law of large numbers the averaged inclusion probability for one element gets asymptotically closer to the true inclusion probability. This principle is commonly used in Monte Carlo simulations. For the statistical properties of the Monte Carlo Estimator, see for example Robert et al. (1999, pp. 20), Rizzo (2008, pp. 153) or Hammersley (1964, pp. 51). Theoretical inclusion probabilities are the result of

$$\pi_{gi} = \begin{cases} \pi_g \cdot \pi_{gi}^b \cdot \pi_{gi|b}^{PIAAC} = m\frac{N_g}{N} \cdot \frac{M_g}{N_g} \cdot \frac{32}{M_g} = 32 \cdot \frac{m}{N} \\ \qquad \text{for community } g \text{ with MOS } m\frac{N_g}{N} \le 1 \\ \pi_g \cdot \pi_{gi}^b \cdot \pi_{gi|b}^{PIAAC} = \frac{M_g}{N_g} \cdot \frac{32 \cdot smp_g}{M_g} = 32 \cdot \frac{m}{N} \\ \qquad \text{for community } g \text{ with MOS } smp_g = m\frac{N_g}{N} > 1 \end{cases}$$

where MOS is the measure of size and

$$\pi_g = \begin{cases} m\dfrac{N_g}{N} & \text{for community } g \text{ with MOS } \; m\dfrac{N_g}{N} \le 1 \\ 1 & \text{for community } g \text{ with MOS } \; smp_g = m\dfrac{N_g}{N} > 1 \end{cases}$$

is the probability for selecting community g. smp_g is the number of sample points in community g, which were selected using the Cox (1987) algorithm, i.e. $smp_g \in \left\{ \left[m\frac{N_g}{N} \right], \left[m\frac{N_g}{N} \right] + 1 \right\}$ with $[x]$ the largest integer $\le x$. The $m = 277$ communities are selected proportionally to the number N_g of their 16-65 year-old inhabitants. Overall in Germany there are $N = 53{,}989{,}232$ 16-65 year-olds. If $m\frac{N_g}{N} > 1$ then π_g is set to 1.

$\pi_{gi}^b = \frac{M_g}{N_g}$ is the probability that unit i is part of the gross sample of persons of size M_g which was provided from the registry of community g.

Under equal probability sampling, $\frac{32 \cdot smp_g}{M_g}$ is the probability that unit i is selected from the PIAAC gross sample of size $32 \cdot smp_g$. For community g both the gross sample size M_g of persons and the number of inhabitants that are 16-65 years-old N_g are known; thus $\pi_{gi}^b = \frac{M_g}{N_g}$.

Due to the error in the optimization algorithm used by our survey organization for the sample selection, an equal probability sample was not realized. Thus, inclusion probabilities $\pi_{gi|b}^{PIAAC}$ could only be determined by approximation through simulations. For $r = 10{,}000$ simulated samples, the following inclusion probabilities $\pi_{gi}^b \cdot \pi_{gi|b}^{PIAAC}$ for the units i of the PIAAC sample are computed given the erroneous and correct algorithm (see histograms in Figure 4).

The design effect due to unequal selection probabilities is

$$Deff_p = n \frac{\sum_{i=1}^{I} n_i w_i^2}{\left(\sum_{i=1}^{I} n_i w_i\right)^2} = 1.22,$$

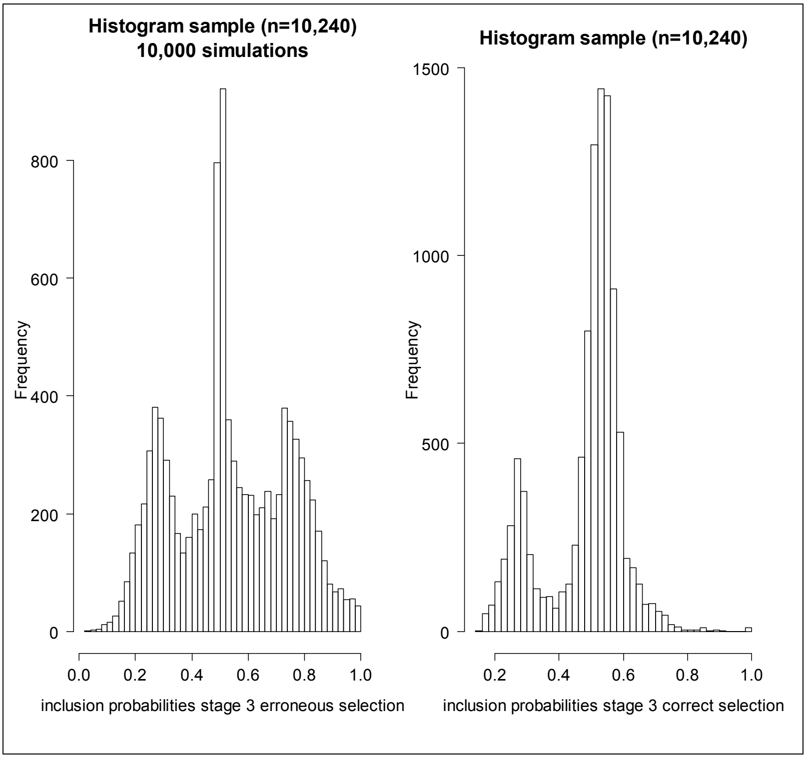

Figure 4: Histograms of $\pi_{gi|b}^{PIAAC}$ for units i of the PIAAC sample given the erroneous and correct algorithm.

where n_i is the number of observations in weighting class i and w_i are the weights in weighting class i. An explanation for the higher design effect given the erroneous optimization algorithm implemented by the survey organization is that this selection process favored certain units while neglecting certain other units. As a consequence, the required equal selection probability was not achieved.

4 Conclusion

Theoretically, the PIAAC sample for Germany should have been selected with equal probabilities for all individuals. However, due to an error in the selection procedure, this target could not be realized. Instead, an erroneous optimization algorithm was applied which led to inclusion probabilities that were too complex to calculate for us in the available time. But since the optimization procedure was a random procedure, it was possible to determine the probabilities with the help of a simulation. The selection procedure was repeated 10,000 times and the number of times being included in the sample for each individual was reported. This number divided by 10,000 yields a good approximation of the inclusion probabilities. The disadvantage of the incorrect optimization algorithm for our sample is the higher design effect compared to the one based on equal inclusion probabilities. This design effect due to unequal inclusion probabilities was 1.22, i.e. the effective sample size was $n_{eff\_p} = n_{net}/Deff_p$ = 5,319/1.22= 4,360. In other words: The precision of the estimates is – only because of this error – just as high as if 4,360 interviews of a simple random sample would have been conducted. This is 82% of the original sample size.

Nevertheless, the PIAAC sample is a full probability sample and complies with all requirements of the Consortium. The sample passed the adjudication procedure with the following statements: "Through Consortium review of the preliminary SDIF, an anomaly was detected in the age distribution of the sample, with spikes at ages 30, 40, and 50. Germany investigated the reason for this pattern and discovered an error in the sample selection algorithm at the last stage of selection. Germany provided evidence that the sample remained probability-based despite this error and corrected the selection probabilities to reflect the actual selection algorithm used. However, they were unable to calculate exact selection probabilities, so the probabilities are based on a simulation" (see OECD 2013, Appendix 7, p. 69).

Quite generally, a good approximation for the true inclusion probabilities with 10,000 simulation runs is only meaningful if the sampling fraction f is high enough. In the case of the PIAAC-sample $n = \sum_g n_g = 10,240$ out of $N = \sum_g M_g = 23,117$ cases had to be selected, so $f=n/N=0.44$. Otherwise, with a (much) lower sampling fraction the simulation with 10,000 replicates just would have led to white noise and it would have been impossible to determine inclusion probabilities this way.

References

Cox, L. H. (1987). A Constructive Procedure for Unbiased Controlled Rounding. *Journal of the American Statistical Association, 82*(398), 520-524. doi: 10.1080/01621459.1987.10478456

Fisher, R. A. & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. London: Oliver and Boyd.

Hammersley, J. M. & Handscomb, D. C. (1964). *Monte Carlo methods (Vol. 1)*. London: Methuen.

Kish, L. (1994). Multipopulation Survey Designs. *International Statistical Review* 62, 167-186.

Lynn, P., Japec, L., & Lyberg, L. (2006). What's So Special About Cross-National Surveys? In J. Harkness (Ed.), *Conducting Cross-National and Cross-Cultural Surveys.* ZUMA-Nachrichten Spezial Band 12, 7-20.

Lynn, P., Liebig, S., & Weinhardt, M. (2014). *Sampling for the European Social Survey – Round VII. Germany.* Retrieved from http://www.europeansocialsurvey.org.

OECD (2009). *PIAAC Sampling Plan (Main Survey) Part I: Sample Design and Selection Plans.* Revised Report (March 2009). Information for the National Survey Design and Planning Report (NSDPR). Chapter 4.

OECD (2013). *Technical Report of Adult Skills* (PIAAC). Paris: OECD.

Rizzo, M. L. (2007). *Statistical computing with R*. CRC Press.

Robert, C. & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer.

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014): *PIAAC Germany 2012: Technical Report*. Münster: Waxmann.