Regular Article

Selection Errors of Random Route Samples

Johannes J. Bauer[1]

Abstract

Random route samples are widely used in face-to-face surveys. Most previous studies of random route sample quality compare the data collected by random route samples with data from reliable sources, such as the German Microcensus. While these studies usually find few differences in the distributions of demographic variables, it is possible that other substantive variables of interest are biased if random route samples select households with unequal probabilities. This article takes a different approach to assessing the quality of random route samples. Since random routes are used when no complete list of respondents is available, it is assumed that all units have the same selection probability. This assumption is tested, by simulating all possible random routes within a German city and calculating the probability of selection for each household. The simulation results show that all three sets of tested random route instructions lead to strong deviations from a uniform distribution and two create systematic biases.

Keywords

survey sampling, random route sampling, random walk sampling, selection, probability, sampling error, selection error, coverage error

[1]Ludwig-Maximilians University Munich, Munich, Germany

Corresponding Author:
Johannes J. Bauer, Ludwig-Maximilians University Munich, Konradstr 6, Munich 80801, Germany.
Email: johannes.bauer@soziologie.uni-muenchen.de

Introduction

Statistical random sampling enables inferences within a defined population. Sampling procedures ensure that the selection probability of each case is known. However, if selection probabilities deviate from expected values—if the sampling methodology fails—estimations can lead to incorrect conclusions about the population.

The random route method is a broadly applied procedure for sampling households for in-person surveys. It is a standard empirical investigation method found in a number of textbooks (Bortz and Döring 2006; Brace and Adams 2006; Diekmann 2010; Schnell, Hill, and Esser 2008; Häder 2010; Kent 1993; Kromrey 2009). Influential examples of the application of random routes are the Gallup World Poll which uses random routes for the selection of respondents in Latin America, former Soviet Union countries, nearly all of Asia, the Middle East, and Africa; the European Quality of Life Survey (EQLS 2003, 2007); the Eurobarometer, a repetitive sample since 1973 of about 16,000 respondents throughout the European Union (EU); the European Community Household Panel (ECHP) which uses random routes to select its German respondents; the BIBB/IAB (Federal Institute for Vocational Education and Training/Institute for Employment Research) consisting of three 0.1 percent samples of the German population; VA Consumer Analysis; the German family survey 1988, 1994, and 2000; and the Socio-Economic Panel (SOEP) (EQLS 2003, 2007; European Commission 2012; German family survey 1988, 1994, 2000; Gallup World View 2012; Haisken-DeNew and Frick 2005; Parmentier and Dostal 2002; VA 2011; Verma and Clémenceau 1996). While many large random route studies concentrate on Europe or individual European countries, the sampling procedure has also relevance in American research. U.S. companies like Gallup, ABC News (2009a, 2009b, 2010), and the Washington Post use random routes for international and third-world surveys (The Washington Post 2010; Gallup 2012).

The random route sampling enables contact to populations without a register. Since this technique is used when no complete list of households is available, it is intended to create equal selection probabilities. This article evaluates whether the procedure does in fact result in uniformly distributed probability samples of households or not.

Random routes are usually used as part of a multistage selection process (Arbeitskreis deutscher Markt- und Sozialforschungsinstitute e.V. [ADM] 1999:69-92; Hoffmeyer-Zlotnik 1997a;). In the early stages of selection, sample points, which typically correspond to urban units, small cities, or voting districts, are randomly chosen. Within these selected areas, interviewers

receive a starting position, which is an address or a household. Beginning at his[1] starting location, the interviewer starts walking along the route determined by the specific random route instructions. These routing specifications are an algorithm telling which streets and crossroads to enter. A typical instruction for this is given by Noelle (1963):

> To locate (...) households, walk in the same direction on the same side of the street until you reach a crossroad or junction on your right side. Arriving there you turn right and proceed toward the right-hand side of the road. Follow this new street until you reach the next road that goes off to the left side and call at the houses on the left-hand street side. (Translated by the author, P. 128)

While following this route, the interviewer samples households. In the typical version, *standard random*, the interviewer selects and questions every *x*th household he passes, where the selection interval defined by the specific value of *x* can vary between surveys. When he has sampled the required number of listed households, the route ends. Another common version of household selection in random routes, *address random*, asks the interviewer to record all households passed by on his prescribed route. The desired number of households is randomly selected from this list afterward. The intention of an address random is to separate selection and questioning of households (ADM 1999:78; Schnell, Hill, and Esser 2008:285; Diekmann 2010:384). By applying these procedures, researchers try to assure that all households have the same probability of being taken into the sample.

There are a few studies that also try to assess the quality of random route samples, comparing the distributions of variables like gender, age, and education in data collected by random route samples with other sources, preferably very reliable sources like the German Microcensus (Bien, Alt, and Krebs 1991; Hoffmeyer-Zlotnik 1997b; Reuband 1998). Such surveys typically focus on demographic variables, as they are available in nearly all surveys and related to many other variables, such as income, health, and so on, which are normally not themselves directly comparable across data sources. As some variables in these surveys deviate from the reliable source, the results are usually attributed to incorrect behavior by or the potential influence of interviewers. However, the survey design does not enable the separation of errors due to interviewers and errors caused by the sampling method. Therefore, the relation between interviewer and error is based on fieldwork experience and theoretical discussion.

But even if no significant differences in the investigated variables were found, it is still conceivable that variables which were not investigated are substantially biased. For example, if a random route sample overrepresents

a specific area in a city, it is still likely that the distribution of gender will not be biased due to the erroneous selection, as there should be no strong relation between unintended overrepresentation and respondent gender. However, this regional overrepresentation could correlate with other substantive variables such as income or employment status and cannot be discovered by the previously described method, if these variables are not part of the comparison.

Such misleading representations of a population are caused in random route samples, if the sample mechanism selects respondents with unequal probability. As true selection probabilities for respondents are unknown, it is not possible to apply weights to correct an unequal selection. Therefore, a major assumption of a random route sample is that every household has the same probability of being taken into the sample. This article tests this central requirement. If the assumption is violated, the chance of systematic survey errors caused alone by the sampling method must be taken into account.

The following research approach is primarily concerned with testing the theoretical assumption that random routes lead to uniformly distributed selection probabilities of households. By excluding incorrect interviewer behavior from the following simulations, it is possible to focus only on the theoretical idea of random routes. Sources of error such as interviewer manipulation during the different phases of the data collection are not part of the analysis. For research about this topic, see, for example, Manheimer and Hyman (1949), Boyd and Westfall (1955, 1965, 1970), Hanson and Marks (1958), Hoag and Allerbeck (1981), Diekmann (2002), Sodeur (2007), and Eckman and Kreuter (2011). For studies that are able to separate sample and interviewer effects in face-to-face surveys, see O'Muircheartaigh and Campenelli (1998) or Schnell and Kreuter (2005).

Methodology

Basic Approach

To examine the equal-probability assumption, it is necessary to calculate the selection probability of each household for the considered random route instructions. As the first-stage selection of sampling points is fixed, the simulations concentrate on the selection of households within sample points, in particular within a small city.

Every method of random route sampling instructs interviewers to start at a randomly selected point and proceed along a given route. Given the starting address, the selection interval, and the routing instructions, the selected
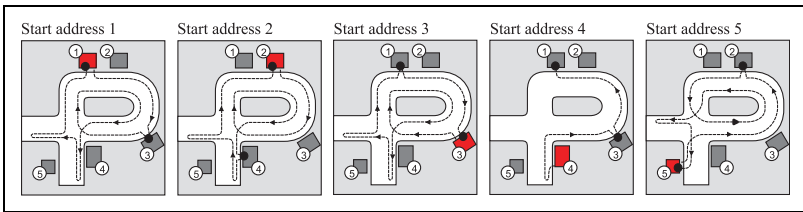
Figure 1. Example for calculation of selection probabilities (black dots mark the contacted households).

households are determined by the street network. If an interviewer starts at a specific location, the route he has to walk is defined by the street network and his instructions. Knowing all households along the interviewers' route, the simulation selects every *x*th unit. This way, given a specific starting location, households that will be contacted can be calculated. In order to get selection probabilities, the simulated interviewers start at each possible starting point once and, as a consequence, work off all possible routes. The overall number of contacts for each household leads to the selection probability of each unit. If all units have the same probability of being questioned, the sampling method has met the requirement of equal probability sampling.[2]

Figure 1 shows an example of how the selection probability can be computed within an artificial small town. In this example, the interviewer is instructed to follow the routing algorithm given in the Noelle quote mentioned previously, turning right at the first corner, left at the second, and so on.[3]

In the first panel of Figure 1, the interviewer begins at the first address, in the upper left corner. His first selection is household 3 and he proceeds to unit 1. In the second panel, he begins at the second unit, selects units 4 and 2. The five panels of Figure 1 demonstrate the sampling outcomes based on every possible starting point in this small example. Looking across all panels in Figure 1, households 1 and 3 are approached three times, whereas number 2 is contacted twice and houses 4 and 5 are selected only once. Altogether the combined number of contacts is 10. Given these quantities, the probabilities for each household can be calculated. Households 1 and 3 have a selection probability of 3/10, household 2 of 2/10, and numbers 4 and 5 of 1/10.[4]

To apply this procedure to a real city, it is necessary to obtain the street network as well as the location of all households. The gathering of data took place between February and April 2010. The data about the road system were taken from the open source webpage, openstreetmap.org, which offers access

to geographical data. As the precision of the data varies depending on the region, a very well documented city was required. Kösching, a small city in Bavaria Germany, close to Ingolstadt, provided very detailed and accurate information about its street network. The quality of the street network was additionally verified by comparisons with cartographical maps and satellite pictures taken from http://maps.google.com as well as an inspection within the city itself.[5] To conduct the following simulations, the place of residence for all inhabitants of Kösching was required, more precisely the number of households for every side of all street sections. These street sections are defined as the part of a street that directly links two crossroads. Therefore, all streets were divided by their crossroads, resulting in the respective amount of sections. The number of households was then noted for each side of all these sections. This gathering of households was done by the researcher himself in order to get very reliable data. Altogether, there are 2,480 households.

To give the results a more general basis, three different random route algorithms are applied. These random route instructions are the central element determining the selection probability. In order to get professional and practical used random route algorithms, I contacted several major market research institutes and asked for their specifications. TNS Infratest as well as the Gesellschaft für Konsumforschung (GfK) kindly shared their random route instructions with me.

For more technical details about the implementation, see online Appendices I to IV (which can be found at http://smr.sagepub.com/supplemental/).

Random Route Instructions

Noelle (1963) proposed the algorithm already presented in the Introduction section. This random route procedure is also found in the teaching material of several sociology institutes at German universities (University of Erlangen 2001; University of Greifswald 2008; University of Kiel 2010; University of Oldenburg 2010) as well as in the textbook of Kromrey (2009). The interviewer begins by facing the street and then turning left. He enters the next crossroad to his right and from there the interviewer alternates between taking crossroads on his left and right.

Whenever he turns right, he switches to the right-hand side of the street; whenever he turns left, he switches to the left-hand side of the street. When the interviewer reaches a street area without households, a city limit, or the end of a blind alley, he turns around and walks back on the same side of the street until he reaches the household where he entered the street. In addition,
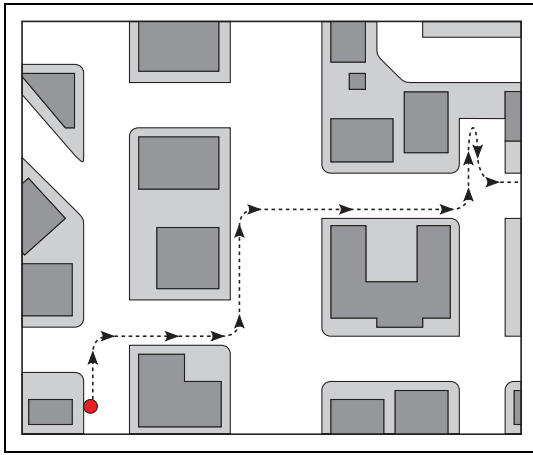
Figure 2. Example for the execution of the Noelle route.

when entering the street on the right-hand side the interviewer also leaves right; the same logic applies to the left-hand side. While walking back, the interviewer is directed not to select households he has already passed. If there are no buildings on the side of the street, where the interviewer is, he walks on the opposite side. As this route is designed for a standard random, the interviewer questions every *x*th household he passes. The start address household is neither counted nor selected except when the interviewer walks past it again. After the interviewer has listed a predefined number of households, the route ends. An example for this procedure is given in Figure 2.

TNS Infratest determines the starting direction by instructing the interviewer to proceed in the direction of decreasing house numbers of the start street. However, the household data at hand gives only information about the location of the households within the street network, but not the house numbers. For this reason, it was not possible with my data sets to determine the prescribed direction. Because the selection of the start direction from TNS Infratest is a replacement for a random mechanism, a true random selection, where each street direction has the same probability, can be used as a substitute. As interviewers can only start in two directions, letting them walk routes in both directions replaces a random selection and leads to the expected probability. A short discussion of this specification will be given in the fourth section.

Beginning at the start address, the interviewer is to walk straight ahead. If there is no straight or almost straight street available, the interviewer enters the street on his right-hand side. If there is no straight and right-hand side, he

Figure 3. Example for the execution of the TNS Infratest route.

is instructed to choose left-hand side roads. Reaching a street area without households, a city limit, or the end of a blind alley, the interviewer turns and walks back on the other side of the street. This rule also applies when there are no buildings on the side of the street where the interviewer is. While walking back he keeps counting households. The side of the street the interviewer walks on is given by the side of his starting address. When he starts on the left-hand side of a street, he has to stay on this side. While the interviewer follows these rules, he selects every xth household. The instructions of TNS Infratest were also given in the textbook of Weischer (2007). An exemplary application is shown in Figure 3.

Using the route of the GfK the start direction is given by street numbers. The interviewer is instructed to follow in the direction of rising numbers. As in the TNS Infratest route, this process is replaced by a random direction selection. At the beginning of the routing algorithm, the street must be followed until the end is reached. When the interviewer gets there, he walks back until he arrives at the other end. From this point on, he constantly enters roads to his right-hand side and stays on the right side of the street. To prevent the interviewer from proceeding in circles, the instructions do not allow walking on the same side of a street twice. If the interviewer reaches a street without households in an area of 500 meters, a city border or the end of a blind alley he turns and walks back on the other side of the street. In contrast to the routes of Noelle (1963) and TNS Infratest, the collection of households of the GfK is designed as an address random. All households the interviewer

Figure 4. Example for the execution of the GfK route.

comes past are listed. After the interviewer has recorded a predefined number of households, a random sample of households is taken from this list and approached for an interview. Each household within one list has the same probability m/M, where m is the number of randomly selected households and M the overall size of the list. An example for these instructions is given in Figure 4.

Methodological Specifications

Using these three sets of rules (Noelle, Gesellschaft für Konsumforschung, and TNS Infratest) and the street and household data for Kösching, I simulated the performance of the random route sampling techniques with all possible starting points. Applying these instructions on the Kösching data requires a specification of the overall number of households to be frequented by an interviewer as well as the number of household to be passed until a contact is made. Shown calculations will begin with questioning every 5th household, altogether 20 interviews for standard random procedures.[6] Therefore, for standard random specification, the simulated interviewer walks past 100 households. In order to enable a comparison at the address random simulation for the route of the GfK, the first results are presented with this quantity. Additional simulations at standard random procedures are done for selection intervals with every 1st, 2nd ... 20th household as well as for an overall number of 1, 2, 3 ... 19 interviews; for address random the number of households

passed by varies between 1 and 600. Given an ideal situation where all households have the same selection probability, all units should have a probability of $100/2{,}480 = .040$ percent.

It was sometimes necessary to stop the routing algorithms, as the random route specifications give no directions on how to proceed. That was usually observed when the simulated interviewer was trapped in a circular pathway and not able to complete the required amount of interviews. In addition, a premature break-off was inevitable when the routing instructions excluded any further options to continue the route. This occurs in particular with the GfK specifications, as the route prohibits the interviewer to enter side of streets he has already been on. In such situations, a real interviewer would be instructed to call his manager to get new directions. As that is not possible within this simulation, these circumstances are treated as break-off conditions and result in fewer than the target number of households being selected.

Results

Route: Noelle (1963)

Applying the Noelle route with a selection interval of 5 and a number of 20 contacts, the simulated data show that many households have a very low, some a very high, probability. 1,908 households (76.9 percent) have a probability which is lower than the equal selection value of .040 percent. 581 households do not have a selection probability at all. On the other hand, there are 173 households with probabilities greater than 0.150 percent. Two of those probabilities are close to 1 percent (.949 percent, .956 percent), which are more than 400 times higher than the household with the lowest positive probability (.002 percent).

A Gini coefficient close to 1 indicates a strong deviation from an equal probability distribution, while values close to 0 would indicate a more equal distribution. The Noelle algorithm produces a Gini coefficient of .78. Hence it shows a clear violation of the equal probability assumption.

In addition, looking at the map in Figure 5, where each dot is a household and higher selection probabilities are represented by a darker color, especially the south-eastern areas as well as some scattered streets include more overrepresented households. This representation is caused by the left–right instructions. The routing algorithm creates a fixed path that leads along the black colored households. If an interviewer gets on this track at some point of his route, he will have to follow it until he reaches the required number of
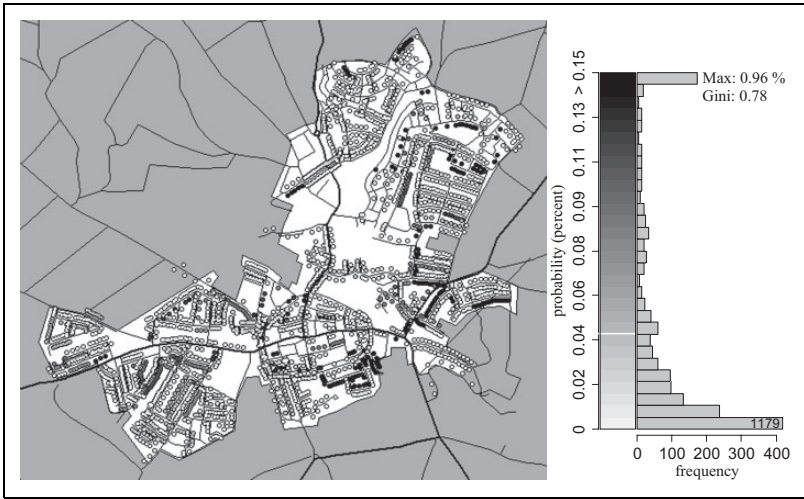
Figure 5. Simulation of selection probabilities (Noelle route; contacting every 5th household, altogether 20 households).
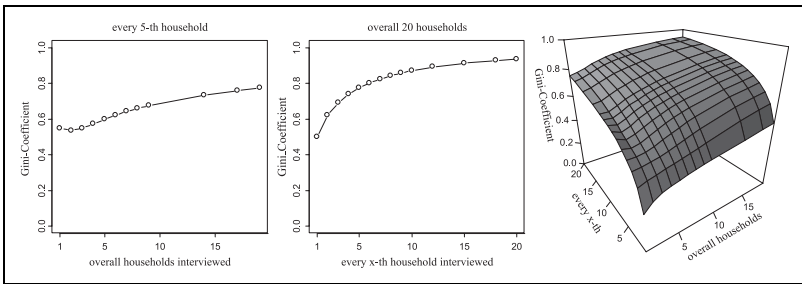

Figure 6. Deviation from a uniform distribution plotted against the overall number of households and the number of units to be passed until the next contact.

interviews. Therefore, the more households the interviewer has to pass, that is, the larger the product of selection interval and number of interviews, the higher the overrepresentation along this pathway.

The length of the route strongly depends on the number of selections and the selection interval. As shown in Figure 6, higher values for these two parameters lead to higher Gini coefficients, and consequently to larger deviations from a uniform distribution.[7] In addition, the resulting path ends in a cycle
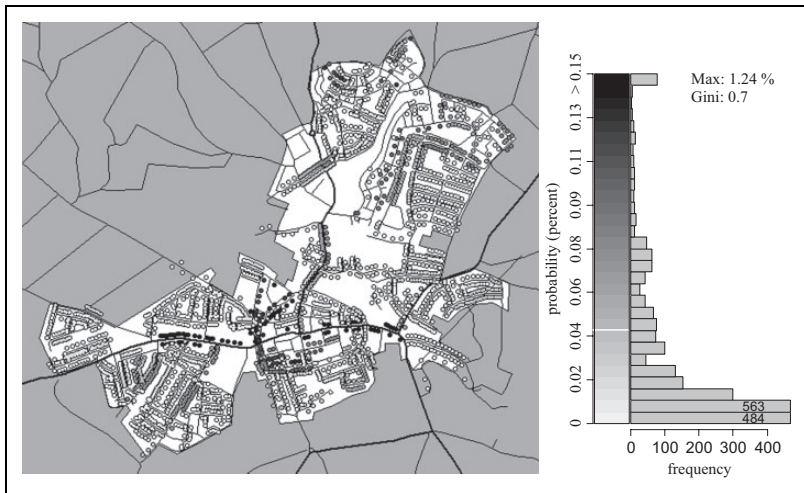
Figure 7. Simulation of selection probabilities (TNS Infratest route; contacting every fifth household, altogether 20 households).

which the interviewer cannot leave if he performs his instructions correctly. The longer the route to walk, the higher the selection error.

To summarize, the Noelle procedure generates extreme deviations from an equal probability distribution even in the optimal situation where all households in a city are known. Consequently, the Noelle route cannot fulfill this basic random route requirement.

Route: TNS Infratest

For the TNS Infratest routing instructions, the selection probabilities for questioning every 5th household altogether 20 are shown in Figure 7. The overall distribution of the TNS Infratest route looks fairly similar to the Noelle route. There are 1,964 households (79 percent) with a probability lower than .040 percent and 94 which have a selection probability of more than .150 percent. Now 25 units exceed the selection probability of 1 percent and, with a number of 144, slightly fewer households have a selection probability of zero. As in the Noelle route, the TNS Infratest route creates a strong deviation from the equal probability assumption.

Whereas the histogram is comparable to the Noelle route, the distribution of overrepresentations on the map in Figure 7 shows a systematic clustering
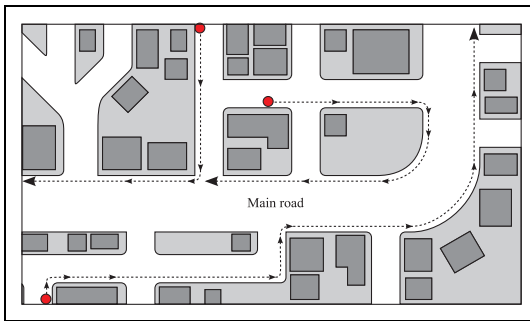
Figure 8. Example for the selection mechanism that increases the probability of main roads (TNS Infratest route).

of higher probabilities in major roads. The reason for this overrepresentation is again related to the basic routing specifications. The interviewer has to move straight ahead as long as possible. Only when the interviewer cannot proceed farther, he enters a street to his right. The likelihood that interviewers enter long streets with many entrances will increase. Typically, main roads have a lot of streets that will end there. Hence, households on main roads have a higher probability of being frequented by a random route interviewer (Figure 8).

The instructions concerning blind alleys aggravate this problem. In these cases, the interviewer has to turn back and proceed straight ahead. As main roads normally lead out of the city and because the interviewer is not allowed to cross the city border, these areas will be treated as blind alleys. So the interviewer is led to main streets when using the walk-straight-ahead mechanism and, as soon as he has entered, he will not be able to leave this street without violation of his instructions. Instead, he will walk back and forth between the two ends of those streets.

Consequently, this process leads to a very high overrepresentation of households in main roads. As this enhancement of probabilities only ceases when the interviewer has gathered his required number of households, long routes inevitably lead to high overrepresentations (Figure 9).

Route: Gesellschaft für Konsumforschung (GfK)

The histogram in Figure 10 shows the result for the route of the GfK when the interviewer has to write down 100 households. As with the other random route instructions, there is a strong deviation from uniform selection probabilities.
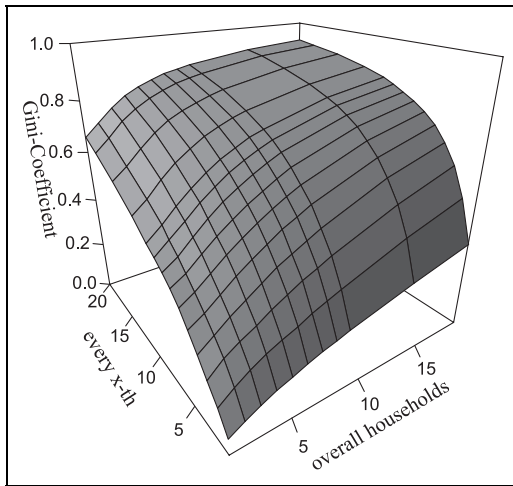
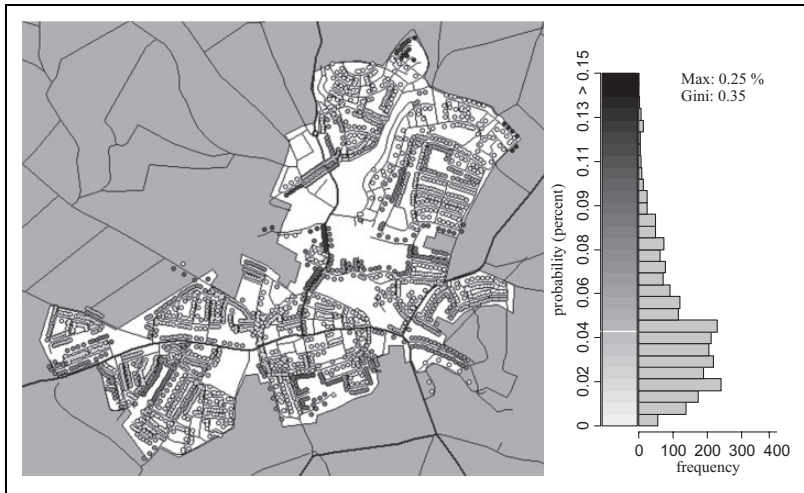**Figure 9.** Deviation from a uniform distribution.

Figure 10. Simulation of selection probabilities (GfK route; writing down 100 households).

It is, however, less severe than in the routes discussed previously. 1,441 households (58 percent) have a probability lower than .04 percent. Hence, the difference between over- and underrepresented households is not that extreme. In
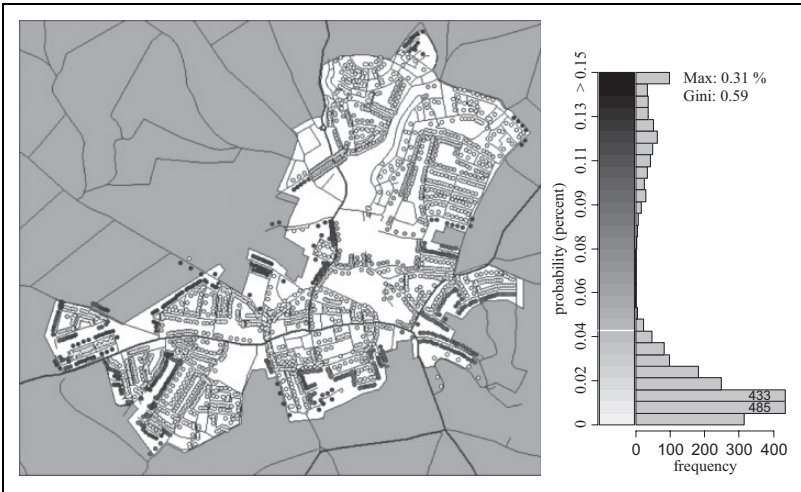
Figure 11. Simulation of selection probabilities (GfK route; writing down 400 households).

addition, there is only 1 household with a selection probability higher than .15 percent and not one household has a selection probability of zero. This reduction of zero selection percentages is caused by the selection procedure. The address random instructs the interviewer to make a list of all households passed by and not selecting every xth household. Therefore, this procedure does not skip respondents.

Despite this, the routing instructions create a systematic selection of households that overrepresents certain areas within Kösching. Because this bias does not occur until the interviewer has left his starting street, it is less visible at 100 households. However, even with this number of contacts, it is possible to see that streets along the outskirts of Kösching are overrepresented (Figure 10). If the number of households the interviewer has to write down is increased to 400, the structure becomes explicit (Figure 11).

The reason for this overrepresentation is related to the rule to constantly take the next crossroad to the right-hand side after leaving the start street. Thus, the interviewer walks around the start street until he gets to the city limit. From there, he once again walks around the streets where he has already been. When he gets back to the location where he first reached the outer area of the city, he is not allowed to proceed as he has already walked on the upcoming side of the street before. Here he is instructed to turn around and switch to the opposite side of the road. The exterior of the city being on
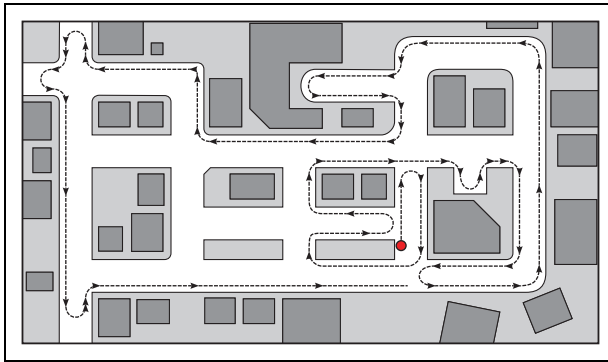
Figure 12. Example for the selection mechanism that increases the probability of outskirts (GfK route).
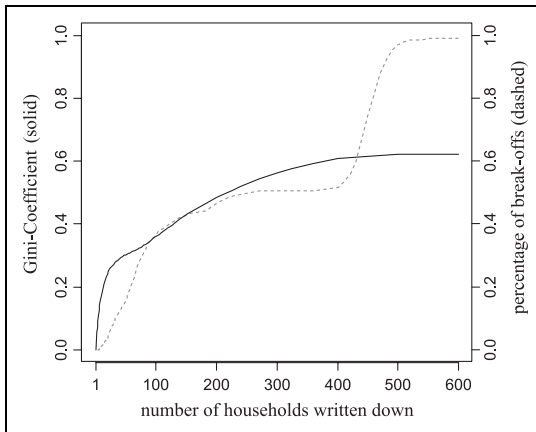

Figure 13. Deviation from a uniform distribution (GfK route).

the right-hand side, the interviewer will walk along the city limit. The more households collected and, consequently, the longer the distances the interviewers have to walk, the more likely it is that the interviewer reaches the city limit. As a result, longer roads lead to higher overrepresentations. An example for this process is given in Figure 12.

This reasoning is supported by the Gini coefficients in Figure 13 (solid line). The more households registered, the stronger the deviation from a uniform distribution. According to the graph, the Gini coefficients approach an

upper limit which is lower than at the other routing instructions. This limit is partly caused by break-off points within the simulations (see Figure 13, dashed line). The primary reason for a premature discontinuation is the restriction not to enter sides of streets that have already been covered. The GfK instructs the interviewer to contact the research direction in such problematic or nonspecific instances. As there were no instructions available as to how these cases are handled, this cannot be emulated by the simulation and the route is discontinued. For routes of short and medium length (1 till 150 households), such instances occur at starting locations in blind alleys where the interviewer walks along the dead-end streets in the direction of the entrance, turning back according to the instructions and then ending up within the blind alley. Occasionally, similar effects are obtained by paths that lead back to a street segment where the interviewer has already been. Long routes (151 till 400 households) that lead around the city where no break-off can happen, stabilize at 50 percent break-offs. However, a very long route (more than 400 household) will get the interviewer back to the street where he first entered the city limit. In this case, the routing instruction does not allow him to proceed forward, but the interviewer also cannot switch the side of the street and turn around as he has already been there before. As a consequence, nearly all the routes had to be stopped (see Figure 12). If the researcher gave the interviewer another starting point from which to complete his route, the results would be a mixture of routes with different lengths.

Due to the number of break-offs and the lack of an ambiguous rule on how to proceed, the true expected selection probabilities cannot be calculated. Still, it is very likely that the GFK route will not result in an equal probability sample, as short routes with very few break-offs already deviated from a uniform distribution. Furthermore, the huge amount of break-offs indicates additional difficulties within the GFK-instructions.

Robustness of Results

The simulated random route instructions all showed a strong deficiency in their selection probabilities. These simulations are based on assumptions that necessarily influence the results. The most influential presetting is that all households within the city are available start locations. This assumption was introduced to create an idealized situation in which random routes are certain to work. In practical applications, this is surely an improper specification because, in these situations, the sample would be taken from the register

of households at hand. Therefore, it seems necessary to discuss what happens when this presetting is relaxed.

Given that starting locations within Kösching are uniformly randomly drawn from a complete list of all households, the results should not be strongly altered. Despite that, the starting location is normally taken from the telephone book and whether people are listed in the phone book or not surely is not a random process. Still, the major requirement in order for the simulation results to hold is that the selection of a starting point is independent from its location. Thus, if the list of households within the telephone book is more or less randomly distributed throughout the city, the selection probabilities will essentially be the same. If there is no strong interrelation between potential start points and their territorial location, the simulated distribution will give similar deviations from a uniform selection. Instead, if the phone book overrepresents specific areas within a city, the individual households are likely to have different probabilities. But, given the extreme deviation from a uniform distribution, it is highly unlikely that a biased set of starting locations will lead to an acceptable selection distribution of households. Due to the observed systematic pathways causing the selection errors, the overall distribution of probabilities will, in all likelihood, not be very different. If the requirement is not given that all households have the same probability of being a starting location, if specific start areas are evidently overrepresented, the quality of the sample is still, maybe even more, questionable.

The second assumption, which shall be discussed shortly, is related only to the TNS Infratest and GfK route and has already been mentioned in the Methodological Specifications subsection. The instructed selection of starting directions by following falling or rising house numbers was replaced by a randomization of starting directions due to a lack of data. Because the original direction was defined by falling or rising house numbers and not due to a random process, the initial directions for the TNS Infratest and GfK route are actually more systematic. As all interviewers starting in the same street will take the same direction and as additional systematic specifications are expected to exacerbate the existing selection distribution, it is unlikely that the overall interpretation of the results will be altered.

Consequences of Unequal Selection Probabilities

During the review phase of this article, two reviewers raised questions regarding the consequences for surveys. As their points were similar and a response might be helpful for survey practitioners, I would like to address their comments in this article.

One reviewer claims that, for surveys, the consequences of the previous results are unclear. Another reviewer notes that, with respect to potential consequences for variables and significance tests, unequal selection probabilities will probably not affect any inferences.

The simulations shown confirm that it is at least possible, if not likely, that random route samples will lead to unequal selection probabilities of households. As equal selection probabilities were unquestionably assumed until now, the consequence is that random route studies should check for and discuss the possible impact of unequal selections on variables. Variable distributions will not be affected if there is no relation between the variables and unequal selection probabilities. However, an error is expected to occur if a correlation, even a weak one, exists. The following example can illustrate how the strength of a correlation affects inferential conclusions.

Assuming that a random route survey was conducted in Kösching and that the researcher does not know about the inhomogeneity of selection probabilities between households, he will treat the sample as simple random sample. In addition, the respondent within a household is selected randomly and the researcher is interested in the gender distribution. As gender can only take the values male and female, the number of male respondents is binomially distributed. If 1 is assigned to male respondents and 0 to female respondents, a positive correlation between gender and selection probabilities will lead to an overrepresentation of men within the sample and to a smaller variance estimator. Depending on the strength of the correlation, a significance test will lead to wrong results.

The next simulation shows the impact of different correlation strengths. For this, data sets that combine a generated gender variable and selection probabilities of households are created, while the correlation level between both variables is varied.

The respondents in Kösching are expected to be half male, half female. Hence, a list of 1,240 males and 1,240 females is generated. The first half of the variable contains all men, the second half all women. The households are merged with the generated gender variable, but they are rearranged, in dependence on their selection probabilities, in order to create a correlation. If a positive correlation is intended, households with high selection probabilities get a greater chance of appearing in a higher position and consequently are more likely to be assigned to the male category. Hence, the probability that a household is among the first 1,240 corresponds to the probability of being male. For a detailed description of the ordering algorithm, see online Supplementary Appendix V (which can be found at http://smr.sagepub.com/supplemental/).

For different correlation strengths, at every tested correlation level, 10,000 data sets are created that combine gender and selection probability. From each
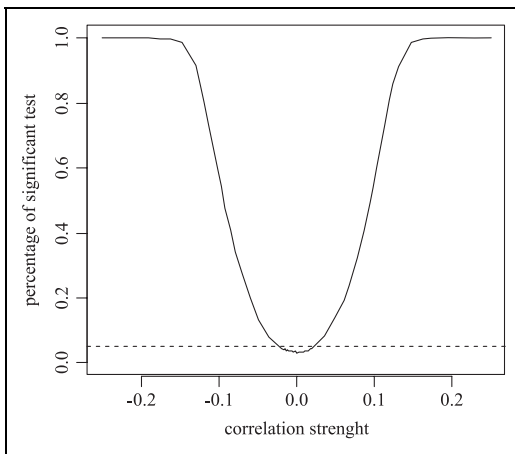
Figure 14. Percentage of significant Binomial tests (the dashed line corresponds to the significant level of .05).

of these, samples of 200 respondents are drawn. Households are selected without replacement, as respondents in studies are not interviewed twice.[8] The probability of a household from the generated data sets being entered into a sample corresponds to the calculated selection probability for the TNS Infratest route with 20 interviews and a contact interval of 5. For the selected respondents, a binomial test is used in order to determine whether a significant deviation from the expected value of 100 men is detected. Within 10,000 samples and a significance level of .05, equal selection probabilities should lead to about 5 percent rejections. Figure 14 shows the calculated percentage at different correlation levels for the TNS Infratest selection probabilities.

As expected, higher correlations come along with a high amount of erroneous results. At a correlation level of about .09, 50 percent of all tests show a significant deviation from the true value. At correlations greater than .17, all binomial tests lead to erroneous conclusions about the population.

Consequently, if random routes are applied, the researcher is to check for relations between variables and unequal section probabilities. If correlations exist, even weak ones, in many cases inference tests will lead to wrong results.

Discussion

In summary, the random route instructions of Noelle (1963), TNS Infratest, and the GfK all failed to fulfill the basic requirements for a standard random sample

in Kösching. All routes result in a strong violation of the equal selection probability assumption. In addition, the Noelle route and the TNS Infratest route both create a coverage error, as the target population does not coincide with the population actually sampled. The undercoverage is especially visible in the Noelle instruction. Given the specifications used in Figure 5, the route excludes about one-fourth of the target population in Kösching. Moreover, with regard to the systematic structure of random routes that causes these misrepresentations of households, unequal selection probabilities occurring in other cities should be taken into consideration. Particularly for the TNS Infratest route, which overrepresents main roads, and for the GFK route, which overrepresents the city border, it is difficult to exclude the possibility that a random route sample creates biased variable distributions. If selection probabilities correlate with specific variables of interest, a systematic error solely related to the sampling procedure can be expected.

Such errors can normally be reduced or corrected by weighting respondents. In the case of random route samples, this will rarely be possible. The Noelle and the TNS-Infratest routes lead to households without selection probabilities. For respondents that cannot be selected, weighting is out of option. Furthermore, to calculate weights, the selection probability of all households needs to be known. But random routes are particularly applied to populations when no complete list of households is available. Consequently, the selection probabilities of households cannot be determined and therefore calculating weights is not possible.

I would like to mention an additional random routes problem that is only indirectly related to the simulations shown. The conventional usage of "left" and "right" lacks a clear definition, necessitating a decision algorithm within the simulation that corresponds with common understanding (see online Appendix II [which can be found at http://smr.sagepub.com/supplemental/]). Despite its overall functionality, there are crossroads with ambiguous street directions. As the street direction is generally easily identifiable and, as those inexplicit situations were relatively rare, no major impacts on the results are expected. Yet, in real situations, the ambiguous definition of directions offers the interviewer a scope for choice which cannot be controlled by the routing specifications. If interviewers want to minimize their workload, they adjust their understanding of left and right and preferably enter streets with apartment buildings where they do not have to walk long distances to get to the next household. Such autonomy left to the interviewer could lead to a systematic bias.

One possible solution would be to give the interviewer a map with a plotted route. In this way, the researcher can control the complete route and will also recognize problems with the routing instructions early on. In addition, this simple

alteration could also solve or at least reduce selection errors, provided the marked route is generated by a true random mechanism. Up to now, the routing instructions did not include randomization, presumably because they offer the interviewer too many opportunities of routing fraud. If the researcher were to determine the route beforehand, randomly selected directions would become an option. This should lead to a distribution of selection probabilities that is probably closer to a uniform distribution. Of course, before such a procedure is applied, it has to be tested and adapted to practical applications in order to detect potential difficulties.

The simulations shown do not confirm the usage of the tested random route instructions. The selection probabilities vastly differ from the desired uniform selection probabilities. As the results of a sample survey strongly depend on correct sampling, and, as only little discussion about the route specifications of random routes exists, the necessity of devoting more effort to constructing routing instructions before applying random routes must be emphasized.

Notes

1. For a better readability, I speak as follows only in the male version, in each case I mean the female form too.
2. There are other requirements not analyzed in this article, notably the independence of households in a sample.
3. A complete description of these random route instructions is given in Random Route Instructions subsection. Note that the routing instructions of Noelle (1963) as well as the other specifications available do not cover all routing possibilities. For example, at Figure 1 starting location 4, the interviewer starts in a

blind alley, just as he leaves his starting address. Here, the specifications give no clear instructions what crossroad direction to choose after turning back. Furthermore, it is unclear if the starting household, which must not be recorded at the beginning, should be listed while walking back. In the example, the interviewer recorded the household and followed the standard route.

4. The shown selection probabilities ignore dependencies between the selection of households. If a starting location is randomly chosen, each starting address has a probability of 1/5. Starting at location 1, 3, and 4, households 1 and 3 have to be questioned. Consequently, household 1 has a selection probability of 3/5 as well as unit 3. Number 2 has a probability of 2/5 and households 4 and 5 are selected in one of the five cases. These numbers are proportional to the probabilities ignoring the dependencies, but sum up to a value of more than 1. Because the sum depends on the selection interval and the number of contacted households, comparisons between different specifications need a scaling, preferably between 1 and 0. As this leads to the shown probabilities ignoring dependencies, these dependencies will not be taken into account.

5. There was one street omitted in the openstreetmap.org-data. This street could also have been interpreted as a lengthened driveway leading to 3 households. Since this will not have a visible effect on the results, this missing street is ignored in the analysis.

6. In consultation with researches from TNS Infratest and the GfK, these were more or less common specifications for practical purposes.

7. Note that even if just the 1st household is selected, deviation from a uniform distribution still exists. Is the starting household close to crossroads or in blind alleys, the routing instructions can lead to unequal selection frequencies even for very short routes.

8. If replacement is allowed, even a zero correlation leads to erroneous results, as households with high probabilities become selected multiple times. If this is the case, mean and variance both strongly depend on these households. Random deviations within the distribution of these households have strong effects on the final results.

References

ABC News. 2009a. *Iraq Poll: Note on Methodology*. New York: American Broadcasting Company. Retrieved May 27, 2012 (http://abcnews.go.com/PollingUnit/story?id=7067400&page=1#.T8HRV8VUbbg).

ABC News. 2009b. *Afghanistan Poll: Note on Methodology*. New York: American Broadcasting Company. Retrieved May 27, 2012 (http://abcnews.go.com/PollingUnit/story?id=6787742#.T8HRDMVUbbg).

ABC News. 2010. *ABC News/BBC/ARD Afghanistan Poll – Note on Methodology*. New York: American Broadcasting Company. Retrieved May 27, 2012 (http://abcnews.

go.com/PollingUnit/abcnewsbbcard-afghanistan-poll-note-methodology/story?id=9
512487#.T8HRCMVUbbg).

Arbeitskreis deutscher Markt- und Sozialforschungsinstitute e.V. [ADM]. 1999.
Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis.
Opladen, Germany: Leske & Budrich.

Bien, Walter, Christian Alt, and Dagmar Krebs. 1991. ''Wie zuverlässig ist die Ver-
wirklichung von Stichprobenverfahren? Random route versus Einwohnermel-
deamtsstichprobe.'' *ZUMA-Nachrichten* 28:65-72.

Bortz, Jürgen and Nicola Döring. 2006. *Forschungsmethoden und Evaluation für
Humanund Sozialwissenschaftler*. Berlin, Germany: Springer.

Boyd, Harper and Ralph Westfall. 1955. ''Interviewers as a Source of Error in
Surveys.'' *The Journal of Marketing* 19:311-24.

Boyd, Harper and Ralph Westfall. 1965. ''Interviewer Bias Revisited.'' *Journal of
Marketing Research* 2:58-63.

Boyd, Harper and Ralph Westfall. 1970. ''Interviewer Bias Once More Revisited.''
Journal of Marketing Research 7:249-53.

Brace, Ian and Karen Adams. 2006. *An Introduction to Market & Social Research.
Planning & Using Research Tools & Techniques: Planning and Using Research
Tools and Techniques*. England, London: Kogan Page Limited.

Diekmann, Andreas. 2002. *Diagnose von Fehlerquellen und methodische Qualität in
der sozialwissenschaftlichen Forschung*. Bern, Switzerland: University of Bern,
Institute for Sociologiy. Retrieved February 23, 2011 (http://epub.oeaw.ac.at/ita/
ita-manuscript/ita_02_04.pdf).

Diekmann, Andreas. 2010. *Empirische Sozialforschung. Grundlagen Methoden
Anwendungen*. Reinbeck, Germany: Rowohlt.

Eckman, Stephanie and Frauke Kreuter. 2011. ''Confirmation Bias in Housing Unit
Listing.'' *The Public Opinion Quarterly* 75:139-50.

EQLS (European Quality of Life Survey). 2003. *First European Quality of Life
Survey*. Loughlinstown, Ireland: European Foundation of the Improvement of Liv-
ing and Working Conditions. Retrieved January 2, 2012 (http://www.eurofound.
europa.eu/docs/areas/qualityoflife/eqls2methreview.pdf)

EQLS (European Quality of Life Survey). 2007. *Second European Quality of Life
Survey*. Loughlinstown, Ireland: European Foundation of the Improvement of Liv-
ing and Working Conditions. Retrieved January 2, 2012 (http://www.eurofound.
europa.eu/docs/areas/qualityoflife/eqls2007_technical_report.pdf).

European Commission. 2012. *Eurobarometer 69.1 (2008), ZA4743 Datenfile Version
3.0.1*. Cologne, Germany: GESIS Datenarchiv. Retrieved August 14, 2012 (http://
info1.gesis.org/dbksearch19/SDESC2.asp?no=4743&search=Eurobarometer&se
arch2=&DB=d&tab=0&notabs=&nf=1&af=&ll=10).

Gallup. 2012. *Worldwide Research Methodology*. Washington, DC: Gallup. Retrieved May 27, 2012 (http://www.gallup.com/se/128147/global-research-methodology.aspx).

Gallup World View. 2012. *Gallup World Poll Methodology*. Washington, DC: Gallup. Retrieved January 2, 2012 (https://worldview.gallup.com/content/methodology.aspx).

German family survey. 1988. *Familysurvey 1. Wave*. Munich, Germany: Deutsches Jugend Institut. Retrieved February 23, 2011 (http://www.dji.de/cgibin/projekte/output.php?projekt=126&Jump1=RECHTS&Jump2=1).

German family survey. 1994. *Familysurvey 2. Wave*. Munich, Germany: Deutsches Jugend Institut. Retrieved February 23, 2011 (http://www.dji.de/cgibin/projekte/output.php?projekt=127&Jump1=RECHTS&Jump2=1).

German family survey. 2000. *Familysurvey 2000, Method Report*. Munich, Germany: Infratest Burke Sozialforschung. Retrieved February 23, 2011 (http://db.dji.de/surveys/docs/2/5/fs2000methoden.pdf).

Häder, Michael. 2010. *Empirische Sozialforschung. Eine Einführung*. Wiesbaden, Germany: VS Verlag.

Haisken-DeNew, John P. and Joachim R. Frick. 2005. *DTC. Desktop Companion to the German Socio-economic Panel Study (SOEP)*. Berlin, Germany: DIW.

Hanson, Robert and Eli Marks. 1958. "Influence of the Interviewer on the Accuracy of Survey Results." *Journal of the American Statistical Association* 53:635-55.

Hoag, Wendy and Klaus Allerbeck. 1981. "Interviewerund Situationseffekte in Umfragen: Eine log- lineare Analyse." *Zeitschrift für Soziologie* 10:413-526.

Hoffmeyer-Zlotnik, Jürgen H. P. 1997a. "Random-Route-Stichproben nach ADM." Pp. 33-42 in *Stichproben in der Umfragepraxis*, edited by S. Gabler and J. H. Hoffmeyer-Zlotnik. Opladen, Germany: Westdeutscher Verlag.

Hoffmeyer-Zlotnik, Jürgen H. P. 1997b. "Quoten Versus RandomRoute. Vergleich Shellund DJIJugendstudie." Pp. 68-80 in *Stichproben in der Umfragepraxis*, edited by S. Gabler and J. H. Hoffmeyer-Zlotnik. Opladen, Germany: Westdeutscher Verlag.

Kent, Raymond A. 1993. *Marketing Research in Action*. England, London: Routledge.

Kromrey, Helmut. 2009. *Empirische Sozialforschung. Modelle und Methoden der standardisierten Datenerhebung und Datenauswertung*. Stuttgart, Germany: UTB.

Manheimer, Dean and Herbert Hyman. 1949. "Interviewer Performance in Area Sampling." *The Public Opinion Quarterly* 13:83-92.

Noelle. 1963. *Umfragen in der Massengesellschaft. Einführung in die Methoden der Demoskopie*. Reinbeck bei Hamburg, Germany: Rohwolt.

O'Muircheartaigh, Colm and Pamela Campanelli. 1998. "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision." *Journal of the Royal Statistical Society, Series A* 116:63-77.

Parmentier, Klaus and Werner Dostal. 2002: ''Qualifikation und Erwerbssituation in Deutschland – Konzeption und inhaltliche Schwerpunkte der BIBB/IAB-Erhebungen.'' Pp. 31-44 in *IABKompendium Arbeitsmarktund Berufsforschung, vol 250, Beiträge zur Arbeitsmarktund Berufsforschung*, edited by G. Kleinhenz. Nürnberg, Germany: Bertelsmann Verlag.

Reuband, Karl-Heinz. 1998. ''Quotenund Randomstichprobe in der Praxis der Sozialforschung. Gemeinsamkeiten und Unterschiede in der sozialen Zusammensetzung und den Antwortmustern der Befragten.'' *ZA-Information* 43:48-80.

Schnell, Rainer, Paul Hill, and Elke Esser. 2008. *Methoden der empirischen Sozialfoschung*. Oldenbourg, Germany: Wissenschaftsverlag.

Schnell, Rainer and Frauke Kreuter. 2005. ''Separating Interviewer and Sampling-Point Effects.'' *Journal of Official Statistics* 21:389-410.

Sodeur, Wolfgang. 2007. ''Entscheidungsspielräume von Interviewern bei der Wahrscheinlichkeits- auswahl.'' *Methoden – Daten – Analysen* 1:107-30.

The Washington Post. 2010. *A Look at the Method behind the Post Afghan Poll*. Washington, DC: Retrieved May 27, 2012 (http://www.washingtonpost.com/wpdyn/content/article/2010/12/06/AR2010120601797.html).

University of Erlangen. 2001. *University of Erlangen Teaching Materials*. Erlangen, Germany: University of Erlangen, Institute of Sociology. Retrieved February 23, 2011 (http://www.soziologie.wiso.uni-erlangen.de/archiv/ws0102/sozmeth/panel/randomwalk.pdf).

University of Greifswald. 2008. *University of Greifswald Teaching Materials*. Greifswald, Germany: University of Greifswald, Institute of Sociology. Retrieved February 23, 2011 (http://wulv.unigreifswald.de/2008_kowiforschung/userdata/Stichprobenziehung.pdf).

University of Kiel. 2010. *University of Kiel Teaching Materials*. Kiel, Germany: University of Kiel, Institute of Sociology. Retrieved February 23, 2011 (http://www.soziologie.unikiel.de/mitarbeiter/kriwy/material/VL07_Stichpr.pdf).

University of Oldenburg. 2010. *University of Oldenburg Teaching Materials*. Oldenburg, Germany : University of Oldenburg. Retrieved February 23, 2011 (http://viles.unioldenburg.de/navtest/viles0/kapitel03_Datenmessung ∼ ∼lund ∼ ∼laufb ereitung/modul04_Auswahlverfahren/ebene02_Beispiele ∼ ∼lund ∼ ∼lAufgaben / 03_04_02_01.php3).

VA. 2011. *Verbraucheranalyse Codeplan*. Bremen, Germany: Cazaia Marktforschung. Retrieved January 2, 2012 (www.verbraucheranalyse.de/downloads/58/VA2011_Codeplan.pdf).

Verma, Vijay and Anne Clémenceau. 1996. ''Methodology of the European Community Household Panel.'' *Statistics in Transition* 2:1023-62.

Weischer, Christoph. 2007. *Sozialforschung. Theorie und Praxis*. Konstanz, Germany: UTB.