Content

Letter from the Editors

We are proud to publish the first English language issue of mda. Now in its 8th year, mda is not a newly founded journal. The journal originally started as the German language journal on quantitative methodology, called "methoden, daten, analysen" back in 2007. By making the transition to publishing in English we are able to offer the international community a new opportunity for publishing and reading peer-reviewed, open access articles free of publication fees and other barriers to research and researchers.

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration. We welcome manuscripts from all countries and continents, both national studies and international comparisons are of great interest to our readers.

Mda is released in print and online as open-access journal. All content is freely available and can be distributed without any restrictions, ensuring the free flow of information that is crucial for scientific progress. For more information on publishing in mda, and for subscribing to the journal please visit our website www.gesis.org/mda.

The Editors of methods, data, analyses

Henning Best, Marek Fuchs, Bärbel Knäuper,
Edith de Leeuw, Petra Stein

methods, data, analyses | Vol. 8(1), 2014, pp. 5-24 DOI: 10.12758/mda.2014.001

The Accuracy of Pre-Election Polling of German General Elections

Rainer Schnell, Marcel Noack
University of Duisburg-Essen

Abstract

Pre-election polls are the most prominent type of surveys. As with any other survey, estimates are only of interest if they do not deviate significantly from the true state of nature. Even though pre-election polls in Germany as well as in other countries repeatedly show noticeably inaccurate results, their failure appears to be quickly forgotten.

No comparison considering all available German data on actual election results and the confidence intervals based on pre-election polls has been published. In the study reported here only 69% of confidence intervals covered the election result, whereas statistically 95% would have to be expected. German pre-election polls even just a month ahead are therefore much less accurate than most introductory statistical textbooks would suggest.

Keywords: Pre-Election-Polls, Empirical coverage, Confidence intervals for binomial data, Design effects, Sonntagsfrage

1 Introduction

Pre-election polls account for a large proportion of political media coverage. The interest in election forecasts is based on the assumption that election results can be precisely predicted (Crespi 1988: 4). Contrary to this assumption, the available literature records a long series of failures that is not limited to either specific countries, or time periods. Common examples are the American presidential elections in 1948 and 1996 (Mitofsky 1998), the election of the British House of Commons 1992 (Lynn/Jowell 1996: 22), the French presidential election in 2002 (Durand et al. 2004), the Italian parliamentary election 2006 (Callegaro/Gasperoni 2008) and the 2005 Bundestag election (Groß 2010: 9).[1] However, the fact of its repeated failure does not appear to be common knowledge. For example, several contemporary German textbooks of statistics present naïve and uncommented calculations of confidence intervals based on pre-election polls.[2] Those computations rely on the same erroneous assumptions on confidence intervals and their interpretation as the sometimes reported *margins of error* in media coverage of pre-election polling. All these computations ignore the additional problems of surveys on human populations in general (Groves 1989: IV) and the specific problems of pre-election polls (Wüst 2010). Since these problems introduce more uncertainty in estimates for population parameters, the accuracy of pre-election polls in Germany is much lower than the naïve margins of error computations suggest as we will show.[3]

Direct correspondence to
Rainer Schnell / Marcel Noack, University of Duisburg-Essen, Research Methodology Group, Lotharstr. 65, 47057 Duisburg, Germany
E-mail: rainer.schnell@uni-due.de / marcel.noack@uni-due.de

1 Research on the development of election forecasts over time is available for some countries. For Portugal 1991-2004, see Magalhães (2005); for the United Kingdom 1950-1997, see Sanders (2003); for the USA 1979-1987, see Crespi (1988); for Germany 1947-2009, see Groß (2010).

2 For example Behnke et al. (2006: 397-399), Bosch (2012: 180-181), Fahrmeir et al. (2007: 393), Gehring/Weins (2009: 266-268), Klammer (2005: 124), Luderer (2008: 98) or Oestreich/Romberg (2012: 243-244).

3 This discrepancy between textbooks and empirical facts is hard to explain. One possible mechanism is due to the ambiguity of the German word *Wahlprognose*. The international scientific literature distinguishes between exit polls and pre-election polls. In German, the words *Wahlprognose* and *Hochrechnung* are used for both kinds of surveys. Since the high level of precision of exit polls in Germany leaves no room for further improvement (Hilmer 2009: 258), this accuracy is probably falsely attributed to all kinds of election polls.

In the following, we present a comprehensive statistical review on the performance of German pre-election polls of general elections between 1957 and 2013 based on specific voting intentions (*Sonntagsfragen*).[4]

2 Methodological Problems of Pre-election Polls

The purpose of any survey is the estimate of a population parameter μ by a sample statistic $\hat{\mu}$. In this context, the central concept is the *Total Survey Error* model (TSE). The most commonly used criterion of quality within the TSE is the *Mean Squared Error* (MSE),

$$\text{MSE}(\hat{\mu}) = \text{Bias}^2 + \text{Variance} \tag{1}$$

which is the sum of the squared Bias (difference between the expectation of the estimate $\text{E}(\hat{\mu})$ and the population parameter μ) and the variance of the estimate (Schnell 2012: 387).

The main sources of error for the MSE are specification error, frame error and non-response error on the side of bias; sampling mostly affects the variance of the estimate. Measurement errors and data processing errors are equally relevant for both bias and variance (Biemer/Lyberg 2003: 59).

Any of these error sources can have such a severe impact that any conclusions drawn from the data have to be considered as false (Alwin 2007: 3). Therefore the objective of a good survey design is to minimize the sources of these errors, taking into account the available resources and other limiting factors (Biemer/Lyberg 2003: 38). For this reason, detailed information on the design and execution of a survey are essential in order to assess its quality.

The mode of sampling is of central importance for the errors of surveys. Hence, the methodological literature on pre-election polls agrees that quota sampling should be avoided (Lynn/Jowell 1996). With the exception of the IfD Allensbach, quota samples are therefore rarely used in Germany. In accordance with recommendations of the *Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute* (ADM) in 1979, random samples today constitute the norm for such election forecasts (Groß 2010: 49). Those are mostly CATI interviews via random digit dialing. Since the 1950s, the response rate in academic studies has nearly halved (Schnell 2012: 164). The low response rate is due to a decrease in both cooperation

4 In the German political science literature, a distinction beween opinion, projection and prognosis (*Stimmung*, *Projektion*, *Prognose)* has been introduced by Wüst (2003). The first two constructs reflect specific voting intentions and a more or less theoretical weighting of the opinion. Prognosis is reserved for exit polls. However, most publications of polling institutes and German political scientists refer to pre-election poll results based on specific voting intentions as prognosis.

and availability of respondents. Reducing the number of non-contacts to less than 5% requires long field periods as well as a high number of attempts of contact. Pre-election polls do not necessarily implement either. The German television *Politbarometer*, for instance, operates on a field period of four days (Schnell 1997: 117).

Few of the publications of pre-election poll results include information considered as necessary by the professional *standards for disclosure* (AAPOR 2010, similar: ICC/ESOMAR 2008): information on sponsor and surveying institute, the exact phrasing of questions and response categories, details on the sampling frame and problems of coverage, mode of sampling, sample size, standard error, type of weighting, design effects, instructions to the interviewers, notification letters, screening procedures, incentives, detailed information on response rates as defined by AAPOR, interviewer training and interviewer workload and assignment.[5]

Another problem in pre-election polling is known as *political weighting* of the raw data. No algorithm for the computation of these correction factors has been published (Groß 2010: 110). Using Politbarometer-data, Groß (2010: 110) estimated the impact of political weighting on published results between 1986 and 2005. He showed a small mean difference between published and raw data, but a considerable variance of this difference. Further methodological details required for the evaluation of survey results are withheld by the institutes. Information on response rates, contact strategies of the interviewers, or sampling design are reported rarely, or not at all (Groß 2010: 109-111).[6]

Further technical details will nearly always be missing. This includes, for example, the strategy of dealing with hard-to-reach respondents, whose voting behavior can differ from that of easy-to-reach respondents (Crespi 1988: 43). Ever since the "Literary Digest Disaster" of the US election in 1936, problems of coverage and selective non-response bias have been discussed in the methodological literature as the possible causes of failure of election forecasts (Lusinchi 2012; Walsh et al. 2009: 317; Frankovic et al. 2009: 575-587).

Furthermore, individuals who are only available via mobile phone might cause sampling problems.[7] Even when they had a positive and known selection probabil-

---

5 Paragraph 11b of the ESOMAR standards states: "Where any of the findings of a research project are published by the client, the latter shall be asked to consult with the researcher as to the form and content of publication of the findings. Both the client and the researcher have a responsibility to ensure that published results are not misleading" (ICC/ESOMAR 2008). Infratest and Emnid are institutional members of ESOMAR, whereas only some people working for Forsa and Allensbach are members. As a consequence Forsa and Allensbach are not bound by the guidelines. By comparison, the ADM-standards are less mandatory (ADM 1999). As opposed to ESOMAR standards, ADM institutes are not factually responsible for publications of the sponsor. Although the required details as mentioned in the AAPOR standards could easily be published on the pages of the ADM or the institutes, this rarely happens.

6 Walsh et al. (2009: 317) report the same for the USA.

7 On the consequences of these so-called "cell phone onlys", see AAPOR (2009: 31).

ity, the interview situations still cannot be compared and reported voting behavior might differ between respondents on landlines and on mobile phones. Information on response rates distinguishing between mobile phone and landline numbers in German pre-election polling is rarely published.

In general, systematic differences between respondents and non-respondents will cause biased estimates.[8] Therefore, the exclusion of very small subgroups can have a high impact on the results. The details needed to estimate these effects are unfortunately hardly ever reported in the case of election coverage by means of pre-election polls. Since the technical details needed for a methodological analysis of a pre-election poll are seldom published, currently no comprehensive methodological analysis of pre-election polls is possible in Germany.[9] This paper will therefore be limited to a statistical analysis of the quality of pre-election polling as forecasting method.

3 Data

The following analyses are based on a dataset of a total of 232 published pre-election polls on the German general elections between 1957 and 2013. This dataset is

8 This non-response bias is given via

$$\bar{y}_{\text{Res}} - \bar{y}_{\text{All}} = \frac{n_{\text{Non}}}{n_{\text{Non}} + n_{\text{Res}}}(\bar{y}_{\text{Res}} - \bar{y}_{\text{Non}})$$

with the respective values for all respondents, respondents (Res) and non-respondents (Non) (for an example, see Groves (1989: 134)). One possibility of estimating the maximum bias would be via the response propensities ρ using the R indicator approach $R(\rho)$ with

$$B_{\max}(y,\rho) = \frac{(1 - R(\rho))S(y)}{2\bar{\rho}} \geq \left|\frac{\text{Cov}(y,\rho)}{\bar{\rho}}\right|$$

with

$$R(\rho) = 1 - 2S(\rho)$$

given, where $S(\rho)$ is the standard deviation of the response propensities, $S(y)$ the variance of the dependent variable in the population, $\bar{\rho}$ the mean of the response propensities and $\text{Cov}(y, \rho)$ the covariance of the response propensities and the dependent variable. The probability ρ that a sampled individual actually answers, is estimated with a number of auxiliary variables x_j, for example, via logistic regression (Schouten et al. 2009: 105). The bias is greater, the stronger the correlation between response propensities ρ and the variable of interest y (Schouten et al., 2009: 107). However, it has to be noted that the selection of the auxiliary variables x_j is of great importance. If the non-response mechanism does not correlate with the auxiliary variables used to estimate the response propensities, the bias will remain unnoticed (Schnell, 2012: 174). Using irrelevant auxiliary variables will miss any existing bias.

9 In the US, the work of Crespi (1988) is still the most extensive methodological analysis of pre-election polls. Recent minor additions can be found in Lau (1994) and DeSart/Holbrook (2003), as well as in Keeter et al. (2000) and Keeter et al. (2006).

a subset of a dataset provided by Groß containing 3610 polling results published between 1949 and 2009 (Groß 2010: 121-126). To reduce the chance of potential last minute swings, only pre-election polls with a sufficiently small temporal distance between poll and election were used. Because of that, polls published more than one month before an election were excluded.[10]

Sample size is a necessary information for the computation of standard errors and confidence intervals. For the majority of the remaining 204 studies, this critical information was not included (n=108). Through extensive archival research, sample size for additional 84 of those prognoses could be determined.

Most of the remaining 24 studies were older than 25 years. Hence, no further details on the studies could be found.[11] For these studies, we used the median sample size of the studies before 1990 (n=1000). Since they met the inclusion criteria of our study, we appended 28 recent polls covering the general election in 2013 to the dataset.[12]

At least 19% of the polls are based on quota samples.[13] Quota samples are no probability samples, therefore inference for quota samples cannot be justified statistically. Pre-election polls based on quota samples are only treated as random samples for the purpose of comparison.

4 Methods

Survey estimates should be reported together with their corresponding confidence intervals (CI). The precision of the estimation is given by the width of the CI.[14] The narrower the CI, the more precise the estimate. If every possible sample, of fixed size, is drawn from the same sampling frame, and a CI is calculated for each independent sample, a well-defined proportion of CIs contains the true parameter. That well-defined proportion is called the coverage probability or confidence level (Särndal et al. 1992: 55). If all statistical assumptions required for the calculation of

10 On request, the original dataset was kindly provided by Jochen Groß.

11 For 15 of these studies, the publications also do not mention the polling companies, which greatly complicates the research.

12 The data were extracted from the web page: www.wahlrecht.de.

13 This is not always apparent from the publications. Since 19% of the polls have been published by a German company which nearly always uses quote samples (namely Allensbach) 19% quota samples is a conservative estimate.

14 Another approach would be the usage of prediction intervals (for a review see Krishnamoorthy/Peng 2011). The difference between those two types of intervals is the intended use. Prediction intervals try to predict a future observation (Devore/Berk 2012: 404). Confidence intervals are statements on the uncertainty of population parameter estimates. Therefore, prediction intervals are not appropriate for our kind of analysis. Of interest here is the latter kind of inference.

CIs are met, a CI can accurately be determined analytically, that is without drawing all possible samples.

Assuming the election results is the population parameter, it can be checked whether the parameter is contained within the corresponding CIs. The number of CIs containing the parameter can be counted. If the assumptions are met, the proportion of CIs containing the parameter should be equal to the coverage probability. If reports of polling results mention sampling errors at all, they almost always report CIs for simple random samples, assuming a binomial distribution.[15] Statistically, this is erroneous in several respects.

Pre-election polls in Germany are hardly ever based on simple random samples, but on complex sampling designs. Nearly always, a complex design will result in a higher standard error than a simple random sample of the same size (Schnell 1997: 272-284). There are essentially two causes for the loss of precision. First of all, most complex samples are cluster samples, so that the population is divided into disjunctive units (areas, schools, number blocks in CATI) before sampling. From each unit, a number of persons, or all, are drawn. However, individuals in a spatial unit tend to be more similar to each other, than individuals chosen independently from the population. This homogeneity within the cluster needs to be taken into account for the estimation.[16]

Furthermore, interviewers generally conduct several interviews. Given that interviews conducted by one particular interviewer are more similar than interviews conducted by different interviewers, these homogeneities cause additional loss of precision (Schnell 1997; O'Muircheartaigh/Campanelli, 1998; Schnell/Kreuter, 2005). This effect increases with the number of interviews per interviewer. Since the number of interviews per interviewer is especially high for CATI surveys, this effect is particularly strong.[17] The impact of the interviewer on the variance of the estimate can be even more severe than the effect resulting from spatial clustering (Schnell/Kreuter, 2005: 401). Unfortunately, this is largely ignored when analyzing CATI surveys.

---

15 The best-known example of pre-election polling in Germany is the public-service television Politbarometer. On their homepage: http://politbarometer.zdf.de, 15.11.2013, the CI for a sample of 1250 respondents and a share of 40% of the votes is indicated as +/- 2.7%.

16 This problem was systematically discussed at first by Kish (1965: 164); an early application to pre-election polling can be found by Converse/Traugott (1986: 1095).

17 This effect (deft) is usually simply estimated with

$$deft = \sqrt{1 + \rho(b-1)}$$

(Kish, 1965: 162), where ρ is the homogeneity within the cluster (more precisely: the intraclass correlation coefficient) and b is the mean of the number of observations within the cluster.

Statistically, the loss of precision of complex designs is called the design effect (deft). Deft is defined as the ratio of the standard error of a complex sample and the standard error of a simple random sample of the same size:

$$deft = \frac{\hat{\sigma}_{\theta,\text{complex}}}{\hat{\sigma}_{\theta,\text{SRS}}} \tag{2}$$

Using estimates of deft, adjusted CIs can be calculated, which give a correct coverage probability.

The corrected intervals are wider than the usually calculated naïve 95%-CIs, by the factor deft:

$$\left[p_i - 1.96 * deft * \sqrt{\frac{p_i(1-p_i)}{n}}; p_i + 1.96 * deft * \sqrt{\frac{p_i(1-p_i)}{n}} \right] \tag{3}$$

The naïve CIs, calculated on the assumption of a simple random sample, therefore lead to believe in a higher precision than actually given.

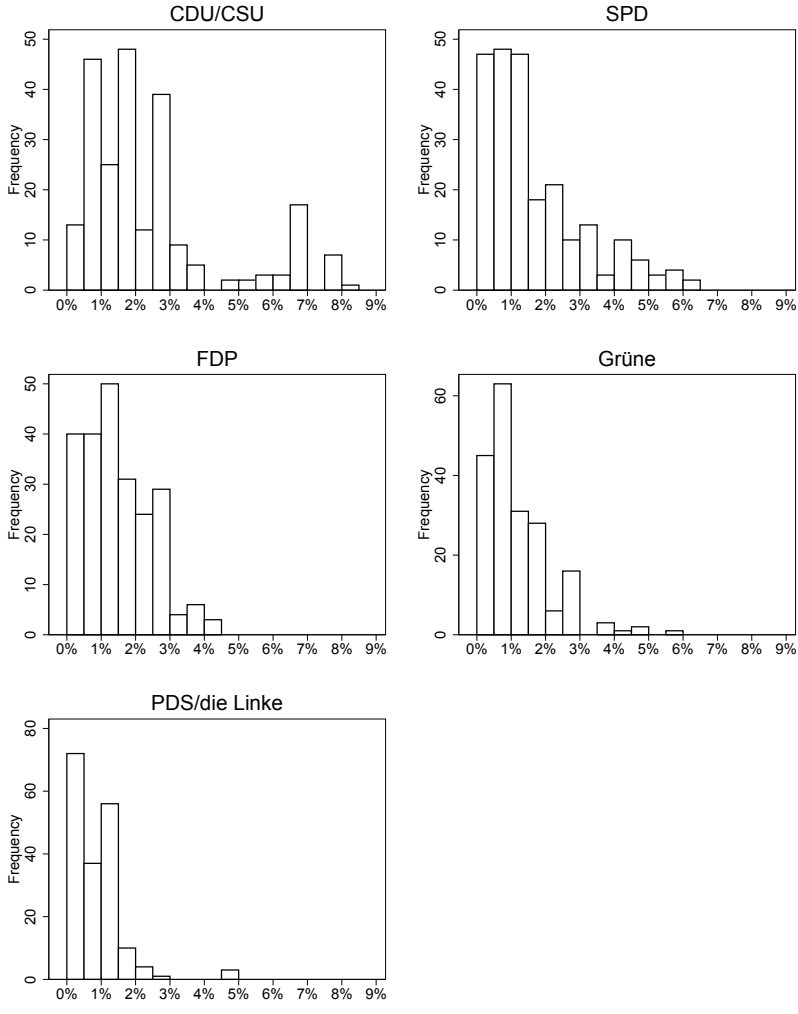
For the calculation of design effects, microdata of the variables of interest, as well as the variables that define the clusters are needed. These data are hardly ever available for pre-election polls. For this reason, an average design effect is occasionally used (UN 2005: 129). Design effects vary considerably; therefore we use the average of 118 estimations from the German Defect Project (Schnell/Kreuter, 2005: 400) with 1.4 (standard deviation=0.3) as a conservative estimate.[18] These intervals are used in the following figures.

5 Results

Of primary interest is the absolute error of the result of the pre-election result compared to the result of the general election. For each party, this is calculated as the absolute value of the difference between the survey result and the election result. Figure 1 shows the distributions of these differences. Obviously, distributions for all parties are right-skewed. Furthermore, there is a second local maximum for the CDU/CSU at 7%. This is due to the general election of 2005, when every poll mispredicted the result of the majority party (CDU/CSU). Naturally, the absolute

18 For comparison: in the Allbus 2008, questions for voting preferences for specific parties show design effects between 1.43 and 1.65 (CDU/CSU, SPD, FDP and Grüne) given the sampling point as cluster, and 1.71 to 2.03 for the interviewer as cluster. Since, as opposed to the Defect study, the Allbus 2008 is not based on an interpenetrating sampling design (Bailar 1983), the confounded effects of interviewer and sampling point cannot be separated.

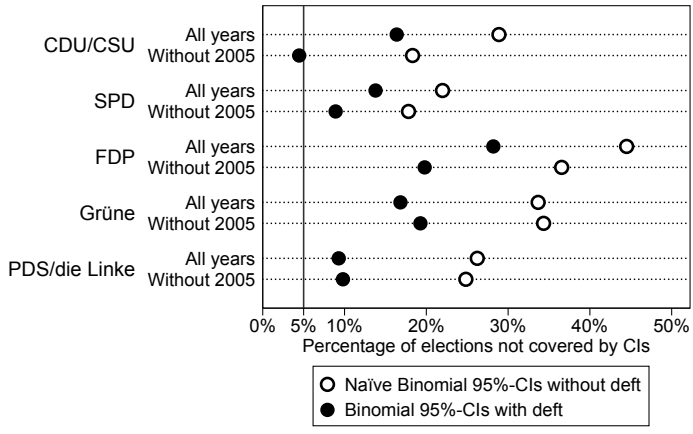
Figure 1: Absolute error of pre-election polls, 1957-2013



error for the small parties (FDP, Grüne and die Linke) is smaller than for the major parties. If the difference between prognosis and result is normed to the size of the party, the resulting relative error is considerably greater. A departure of 2% in the prediction of a party that achieved 6% corresponds to a third of its voters. 9 out of the 145 prognoses (6.2%) for parties with election results under 6% produce relative errors of this magnitude.

Please note: it is expected that at most 5% of the election results are not contained in the CIs; therefore, it is surprising that 6.2% of the poll results exceed a

Figure 2: Empirical non-coverage



third of the respective party size. The absolute error of the pre-election polls is therefore considerably greater than would be expected by a statistically naïve estimation.

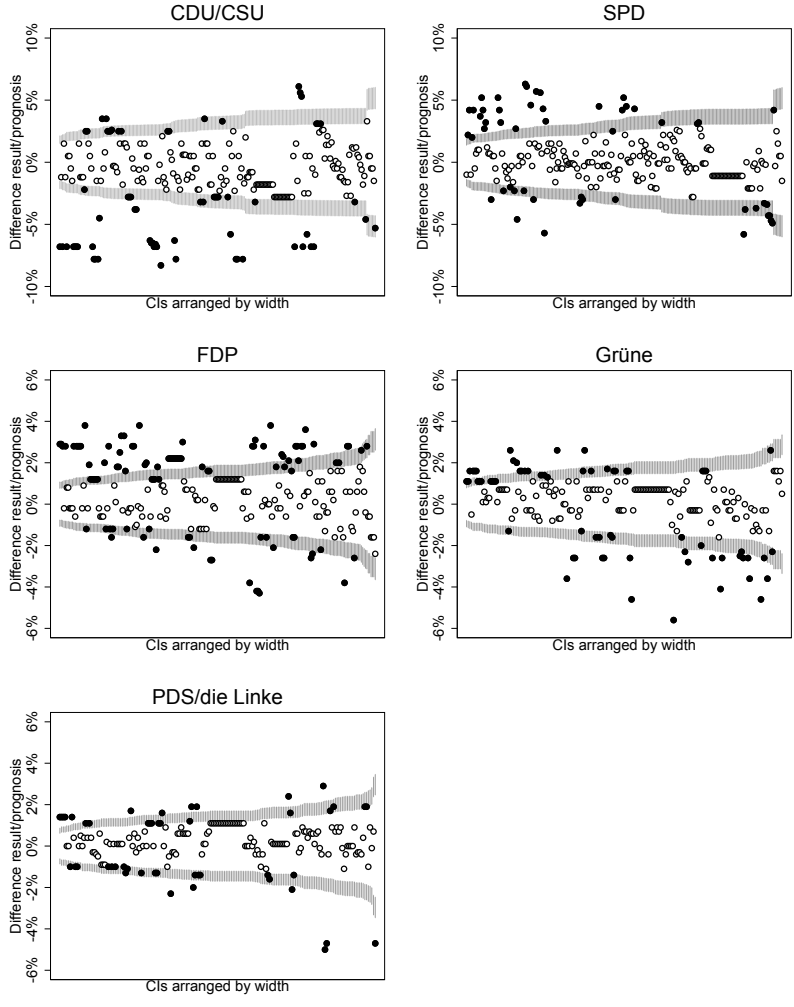
Of central importance for this article is the comparison between the usually applied naïve CIs and the election results. Looking at the coverage of the election results by the calculated naïve CIs, the result is clear (cf. Figure 2): The naïve estimates of the CIs are useless. The aspired confidence level of 95% is missed by far for all parties. Instead of the expected 5%, depending on the party, a minimum of 22% of the CIs do not contain the election result. For the FDP, half of the CIs are affected: instead of 95%-CIs, it comes closer to 50%-CIs (more accurately: 56%-CIs, since 44% of the election results are not contained in the CIs). A coin toss would therefore produce results not much worse than the naïve CIs.

The coverage probability increases greatly when using CIs with design effects. Of those CIs, between 9% and 28% do not contain the election result. These CIs are closer to the usually falsely reported confidence level of 95%, but still far from achieving it.

Figure 3 shows the binomial CIs with and without design effect in comparison for each party. The naïve binomial CIs are distinctly smaller than the corresponding, correctly calculated binomial CIs with design effect.

A consequence of the higher coverage probability is a considerably greater width of the CIs. Figure 4 shows the mean CI widths (CIW) as a dot chart. Half of the correctly computed CIs for the CDU/CSU and SPD have a mean width of more than 7%. FDP, Grüne and Die Linke are roughly at about +/-2%. For most practical applications, this accuracy is not sufficient. If you want to know if a party would

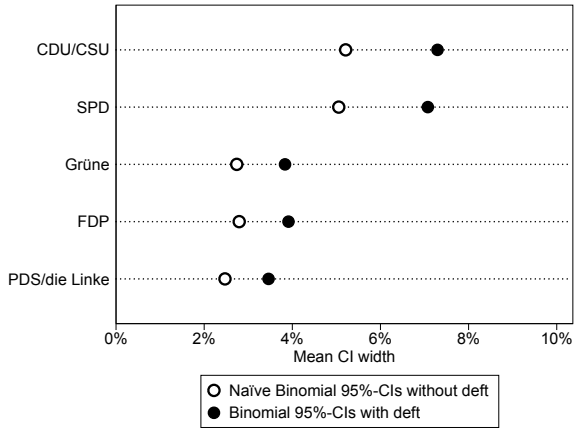
Figure 3: Width of naïve binomial 95%-CIs without deft (inner) in comparison to binomial 95%-CIs with deft (outer) and election results. The naïve CIs without deft corresponding to the results marked as ○ contain the election results; the naïve CIs without deft corresponding to the results marked with ● do not contain the election results.



pass the 5%-electoral threshold, an estimate with a CI from 3% to 7% is factually useless.

This unsatisfactory performance of German pre-election polls becomes more apparent for the number of polls which predicts – given the naïve margin of error – all parties correctly. Statistically, this requires the computation of simultaneous

Figure 4: Dot chart of mean 95%-CI widths



multinomial confidence intervals.[19] For the multinomial CIs, 162 of the 232 Polls (70%) show CIs which all contain the election results. If naïve CIs without deft are used only 67 of the 232 polls (29%) show CIs which all contain the election results. To sum up: Less than a third of the polls would predict all parties within their alleged precision.

The simple fact that small samples, as being used in most polls, cannot deliver the required accuracy for small parties seems to be ignored outside statistics. In general, the width of a CI can be determined given the sample size. If the approximate percentage of votes and the design effect are known, the sample size required for the desired precision can be computed.[20] For a proportion of p=0.4, the width

19 CIs computed for pre-election polls usually assume binomially distributed characteristics. Pre-election polling in Germany has to deal with more than two parties. Therefore, the assumption of binomial distributions is inappropriate, when the results of a pre-election poll are investigated for all parties simultaneously. In this case, it would be appropriate to apply simultaneous multinomial CIs (Ulmer 1989, 1994). Calculating simultaneous multinomial CIs is more demanding than calculating binomial CIs. The easiest approach is the method suggested by Goodman (1965: 250-251). Here, the simultaneous CIs are adjusted according to the number of CIs calculated. For four parties, this would result in a correction factor of 2.498, and 2.576 for five parties. As correction factor, the z-value of $z_{1-\alpha/2}$, as used for a single CI (1.96 for a 95% interval) is replaced by a z-value of $z_{1-\alpha/(2k)}$, where k equals the number of parties. Combined with the assumed design effect of 1.4, the resulting CI for five parties is:

$$\left[p_i - 2.576 * 1.4 * \sqrt{\frac{p_i(1-p_i)}{n}}; p_i + 2.576 * 1.4 * \sqrt{\frac{p_i(1-p_i)}{n}} \right].$$

20 Since the factors ρ, *deft* and $z_{1-\alpha/2}$ are constant, the width of the CI is determined exclusively by $\sqrt{p(1-p)n^{-1}}$.

of a simultaneous CI for n=1000, and a design effect of 1.4 will be 8.5%. To halve the width of the CI, the sample size has to be quadrupled (e.g. Bortz: 2005: 105). Consequently, the width of a CI for 4000 respondents is 4.3%. 16000 respondents provide a CIW of 2.1%, 64000 respondents a width of 1.1%. The width of the CIs is therefore a linear function of the square root of the number of respondents, which transforms the problem of precision to a financial problem. Given the current options of fieldwork in Germany, a sample of 16000 to 64000 respondents cannot be completed within one or two weeks, as required by pre-election polling (Schnell 2012: 385-386).

Even if the resources of all major companies could be pooled, this survey would fail due to the inadequate costs: a pre-election poll of this scale would cost more than €500000.[21] For a still inaccurate estimate, this is not likely to be acceptable to any sponsor.

6 Alternative Explanations for the Failure of Pre-Election Polls

There are two possible alternative explanations for the results of this study. Obviously, opinion changes in the electorate between the end of fieldwork and the election could produce seemingly erroneous results. A less obvious explanation for our result is an increase in accuracy of the pre-election polls during the observed period from 1957 until 2013. The performance of a scientific technique should improve over time. Therefore, worse results would be expected for older polls. Both mechanisms will be examined in more detail.

The literature on pre-election polls sometimes mentions a *last-minute swing* to explain discrepancies between poll results and election results (Roth 2008: 174).[22] Given this hypothesis, a decreasing amount of error would be expected for pre-election polls closer to the election date. This hypothesis is supported by the US results reported by Crespi (1988: 135-136, 166). His results show a significant correlation of r=0.21 between pre-election poll error (difference poll/election result) and the time interval in days before the election. However, although temporal proximity to the election represents the best predictor, his multiple regression model for the difference between polling and election results explains only 12% variance (Crespi 1988: 167). For German data, Groß (2010: 204-212) observes much longer temporal distances of up to one year, and reports a weak, but significant curvilinear

21 For approaches using pooled micro-data of pre-election polls see Park et al. (2004) and Jackman (2005).

22 Occasionally, the mechanism of the "spiral of silence" is mentioned. However, the meta-analysis of all available empirical studies by Glynn et al. (1997) do not give much support for this hypothesis.

Figure 5: Absolute error of the poll results depending on the number of days to the election. The scatterplot-smoother is Loess with f=0.8

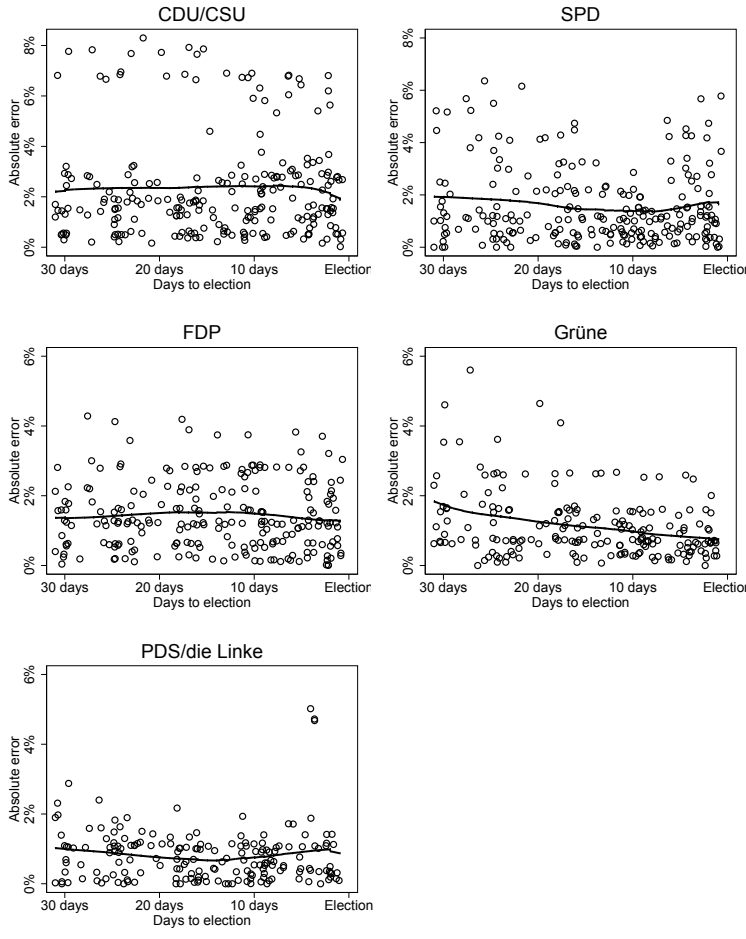
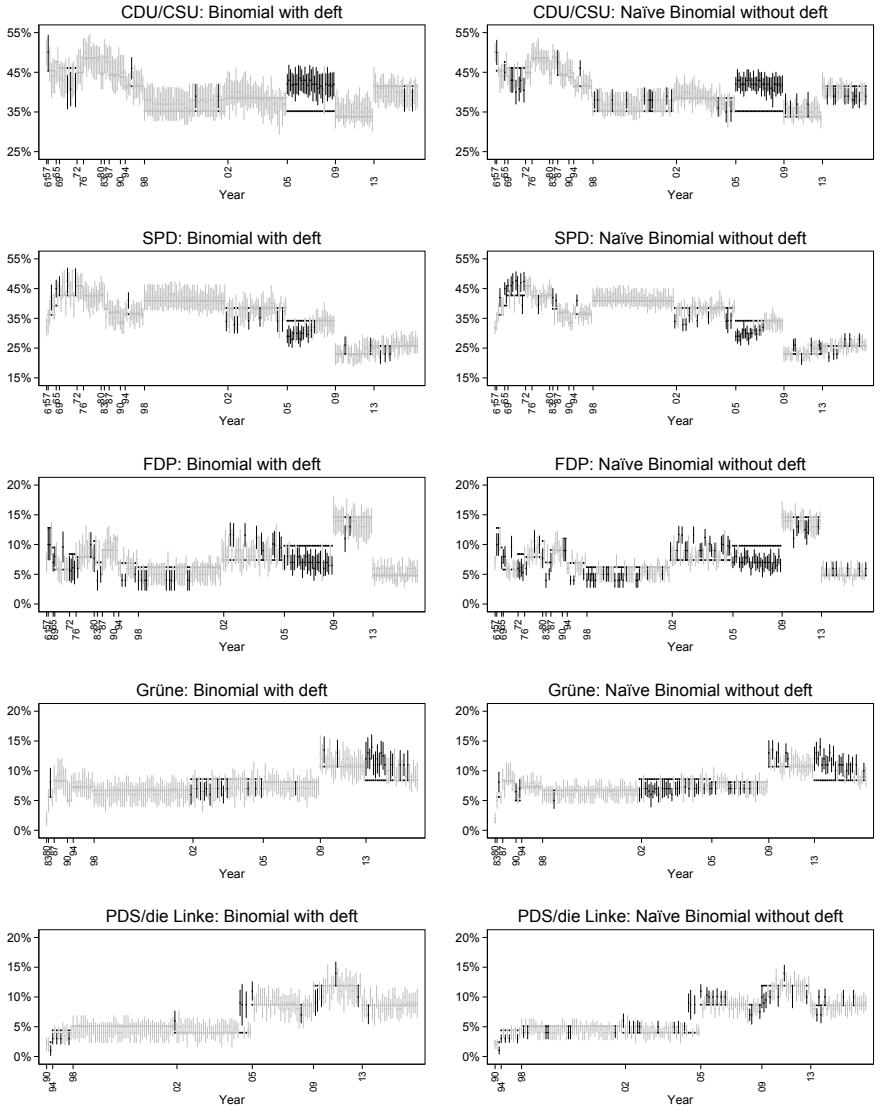


Figure 6: Performance of naïve binomial 95%-CIs without deft (right), in comparison to binomial 95%-CIs with deft (left) over time. Pre-election polls are arranged chronologically. Gray CIs contain the election result, black CIs do not contain the election result. Point estimates are shown as dots.



correlation between temporal distance and error. Our data neither shows a linear nor a nonlinear relationship (cf. Figure 5).[23] Last-minute swings do not seem to be of primary importance for the inaccuracy of the pre-election polls.

The hypothesis of increasing poll accuracy is not supported by the data. This is shown in Figure 6.

7 Conclusions

The comparison of reported margins of error with the actual errors of German pre-election polls between 1957 and 2013 shows disillusioning results: the observed inaccuracy is considerably greater than the published margins of error suggest. The computations of the usual binomial CIs, as taught in most introductory statistical textbooks, is misleading at best. The actual coverage is far below the desired 95%. For some of the small parties, the result is only marginally more accurate than a coin toss. At least for Germany, pre-election polls are not a useful forecasting tool. Applying the statistically more appropriate binomial CIs with design effects, the coverage increases, but at the cost of enlarging the already wide CIs.

Therefore, the results reported here suggest the following conclusions:

- Pre-election polls are not suitable as introductory statistical textbook examples. The formulas to calculate naïve CIs for binomial distributions that are widely used in those textbooks are inappropriate and produce results that are not in accordance with the empirical coverage probabilities.
- The size of the correctly computed CIs (binomial CIs with design effects) make them useless for practical purposes.
- German polling companies rarely report the necessary information for the evaluation of their polling results.

The ad-hoc theoretical weighting of the polling results is neither documented, nor helpful: Although in some cases a reduction of error by theoretical weighting cannot be excluded a priori, systematic evidence favoring theoretical weighting has not been published.

Sampling errors represent only one component of the MSE mentioned in section 2. It is, however, the only component that is quantifiable without a special survey design. Under simplified assumptions, other components may also be esti-

23 A weak effect can only be observed for one of the small parties (Grüne). This effect is due to the election in 2013. Even with these outliers, the temporal proximity explains less than 10% of the variance for this party. The effects remain stable even if not absolute errors, but relative absolute errors are used.

mated, but this would still require more complex designs. The TSE model is therefore used as a regulating idea, rather than an analytical model (Schnell 2012: 388). Assuming that all other components of the TSE do not affect the polling results, the electoral results should be covered by about 95% of the correctly calculated CIs. The data in Figure 2 clearly contradicts this assumption.

The observed low coverage rate of the confidence intervals of German pre-elections could be due to biased estimates, larger variance of the estimators or changing population parameters.[24] Since we eliminated the standard explanation with last minute swings in section 6, biased estimates and increased variances are likely. In our view, the failure of the pre-election polls is primarily due to the limits of measurement of the dependent variable (*Sonntagsfrage*) and the confounding with a second variable of interest, the likelihood of voting. Finally, interviewer effects may be the cause of the increased variance of the estimates (Schnell/Kreuter 2005).

The standard model for pre-election polling in Germany is based on small samples and neither uses a tested theoretical model for coverage errors, nonresponse, electoral participation nor a model for the final decision of undecided voters. Empirically, this model fails far more often than it succeeds.

Acknowledgement

We want to thank Jochen Groß for providing the initial data set, Dipl. Bib. Heidi Dorn for providing the missing sample sizes for 84 studies and the two anonymous reviewers for their very helpful comments.

References

AAPOR. (2009). An evaluation of the methodology of the 2008 pre-election primary polls: A report of the ad hoc committee on the 2008 presidential primary polling. Lenexa: American Association for Public Opinion Research.

AAPOR. (2010). *AAPOR code of professional ethics & practices*. American Association for Public Opinion Research. Retrieved December 16. 2013 from http://www.aapor.org/AM/Template.cfm?Section=AAPOR_Code_of_Ethics&Template=/CM/ContentDisplay.cfm&ContentID=4248

ADM. (1999). *Standards zur Qualitätssicherung in der Markt- und Sozialforschung*. Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute. Retrieved December 16. 2013 from http://www.adm-ev.de/fileadmin/user_upload/PDFS/QUALI.PDF

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken: Wiley.

24 We are thankful to a reviewer for making this point clear.

Bailar, B. (1983). Interpenetrating subsamples. In N. L. Johnson & S. Kotz (Eds.), *Encyclopedia of Statistical Sciences, Vol. 4*. (pp. 197-201) New York: Wiley.

Behnke, J., Baur, N., & Behnke, N. (2006). *Empirische Methoden der Politikwissenschaft*. Paderborn: Schöningh.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken: Wiley.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler.* 6. Edition. Heidelberg: Springer.

Bosch, K. (2012). *Statistik für Nichtstatistiker.* 6. Edition. München: Oldenbourg.

Callegaro, M., & Gasperoni, G. (2008). Accuracy of pre-election polls for the 2006 italian parliamentary election: Too close to call. *International Journal of Public Opinion Research*, 20, 148-170.

Converse, P. E., & Traugott, M. W. (1986). Assessing the accuracy of polls and surveys. *Science*, 234, 1094-1098.

Crespi, I. (1988). *Pre-election polling. Sources of accuracy and error.* New York: Russel Sage Foundation.

DeSart, J., & Holbrook, T. (2003). Campaigns, polls, and the states: Assessing the accuracy of statewide presidential trial-heat polls. *Political Research Quarterly*, 56, 431-439.

Devore, J. L., & Berk, K. N. (2012). *Modern mathematical statistics with applications.* 2. Edition. New York: Springer

Durand, C., Blais, A., & LaRochelle, M. (2004). The polls-review: The polls in the 2002 french presidential election: An autopsy. *Public Opinion Quarterly*, 68, 602-622.

Fahrmeir, L., Künstler, R., Pigeot, I., & Tutz, G. (2007). *Statistik - Der Weg zur Datenanalyse*. 6., Revised Edition. Berlin/Heidelberg: Springer.

Frankovic, K. A., Panagopoulos, C., & Shapiro, R. Y. (2009). Opinion and election polls. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of Statistics: Sample Surveys - Design, Methods and Applications, Vol. 29A*. (pp. 566-595). Amsterdam: Elsevier.

Gehring, U. W. & Weins, C. (2009). *Grundkurs Statistik für Politologen und Soziologen*. 5., Revised Edition Wiesbaden: VS-Verlag.

Glynn, C. J., Hayes, A. F., & Shanahan, J. (1997). Perceived support for one's opinions and willingness to speak out - a meta-analysis of survey studies on the "spiral of silence". *Public Opinion Quarterly*, 61, 452-463.

Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7, 247-254.

Groß, J. (2010). *Die Prognose von Wahlergebnissen. Ansätze und empirische Leistungsfähigkeit*. Wiesbaden: VS-Verlag.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Hilmer, R. (2009). Exit polls - genauer geht's nicht. In H. Kaspar, H. Schoen, S. Schumann & J. R. Winkler (Eds.), *Politik - Wissenschaft - Medien. Festschrift für Jürgen W. Falter.* (pp. 257-267). Wiesbaden: VS-Verlag.

ICC/ESOMAR. (2008). *ICC/ESOMAR international code of market and social research*. International Chamber of Commerce/ESOMAR. Retrieved December 16. 2013 from http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf

Jackman, S. (2005). Pooling the polls over an election campaign. *Australian Journal of Political Science,* 40, 499-517.

Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly*, 70 (Special Issue), 759-779.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-148.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Klammer, B. (2005). *Empirische Sozialforschung - Eine Einführung für Kommunikationswissenschaftler und Journalisten*. Konstanz: UVK.

Krishnamoorthy K. & Peng, J. (2011). Improved closed-form prediction intervals for binomial and Poisson distributions. *Journal of Statistical Planning and Inference,* 141, 1709–1718.

Lau, R. R. (1994). An analysis of the accuracy of "trial heat" polls during the 1992 presidential election. *Public Opinion Quarterly*, 58, 2–20.

Luderer, B. (2008). *Klausurtraining Mathematik und Statistik Für Wirtschaftswissenschaftler: Aufgaben - Hinweise - Lösungen*, 3., Revise Edition. Wiesbaden: Vieweg+Teubner.

Lusinchi, D. (2012): "President" Landon and the 1936 literary digest poll: Were automobile and telephone owners to blame? *Social Science History*, 36, 23-54.

Lynn, P. & Jowell, R. (1996). How might opinion polls be improved?: The case for probability sampling. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159, 21-28.

Magalhães, P. C. (2005). Pre-election polls in Portugal: Accuracy, bias, and sources of error, 1991-2004. *International Journal of Public Opinion Research*, 17, 399-421.

Mitofsky, W. J., (1998). Was 1996 a worse year for polls than 1948? *Public Opinion Quarterly*, 62, 230-249.

Oestreich, M., & Romberg, O. (2012). *Keine Panik vor Statistik! Erfolg und Spaß im Horrorfach nichttechnischer Studiengänge*. 4., Updated Edition. Wiesbaden: Vieweg+Teubner.

O'Muircheartaigh, C., & Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 161, 63-77.

Park, D. K., Gelman, A., Bafumi, J.(2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12, 375-385.

Roth, D. (2008). *Empirische Wahlforschung - Ursprung, Theorien, Instrumente und Methoden*. Wiesbaden: VS-Verlag.

Sanders, D. (2003). Pre-election polling in Britain, 1950-1997. *Electoral Studies*, 22, 1-20.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Schnell, R. & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 389-410.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Opladen: Leske+Budrich.

Schnell, R. (2012). *Survey-Interviews. Standardisierte Befragungen in den Sozialwissenschaften*. Wiesbaden: VS-Verlag.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 101-113.

Sison, C. P., & Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90, 366-369.

Ulmer, F. (1989). *Wahlprognosen und Meinungsumfragen und der Ablasshandel mit den Prozentzahlen: der Lotteriecharakter des repräsentativen Querschnittes - Sonderdruck aus Heft 30./31. Jahrgang der Zeitschrift für Markt-, Meinungs- und Zukunftsforschung*. Tübingen: Wickert-Institute/Demokrit-Verlag.

Ulmer, F. (1994). *Der Dreh mit den Prozentzahlen*. Wuppertal: Bergische Universität GH Wuppertal.

United Nations. (2005). *Household sample surveys in developing and transition countries*. New York: United Nations Publication.

Walsh, E., Dolfin, S., & DiNardo, J. (2009). Lies, damn lies, and pre-election polling. *American Economic Review*, 99, 316-322.

Wang, H. (2008). Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *Journal of Multivariate Analysis*, 99, 896-911.

Wüst, A. M. (2003). Stimmung, Projektion, Prognose? In A. M. Wüst (Eds.), *Politbarometer.* (pp. 83-107). Wiesbaden: VS-Verlag.

Wüst, A. M. (2010). Exit Poll. In D. Nohlen & R.-O. Schultze (Eds.), *Lexikon der Politikwissenschaft, Band 1 A-M*. 4., Updated and Revised Edition. (pp. 242-243). München: C. H. Beck.

methods, data, analyses | Vol. 8(1), 2014, pp. 25-52 DOI: 10.12758/mda.2014.002

Sampling the Ethnic Minority Population in Germany. The Background to "Migration Background"

Kurt Salentin
Bielefeld University

Abstract

The paper discusses techniques for sampling the "migrant background" population in Germany, which comprises all first-generation immigrants, all non-citizens born in Germany, and all children with at least one parent fulfilling one of these criteria. Random walk sampling and random digit dialing techniques are feasible for sampling this population as a whole, but inefficient for subgroups. Telephone directories provide biased representations of the population, and the large proportion of non-pubs disqualifies their use. The Central Register of Foreigners excludes naturalized immigrants and introduces a socio-economic bias toward the less successful. Snowballing overrepresents persons with larger ethnic networks. The center sampling technique may encounter particular problems in Germany due to settlement patterns and legal issues affecting certain immigrants. Local authority Population Registers provide the best representation of the population.

Foreign citizenship fails to identify the target population as it largely underestimates numbers and distorts the social structure. Place of birth is a suitable criterion to identify the Aussiedler population (ethnic German immigrants from eastern Europe and the former Soviet Union). In most cases, however, foreign names best serve the purpose of unbiased sampling. Therefore, name-based sampling in the Population Registers is the method of choice. However, the decentralized administration of Population Registers makes this a costly endeavor and although there is a certain legal sampling interface, there are still legal obstacles to optimal implementation of this sampling procedure.

Keywords: sampling; Germany; ethnic minoritiy; immigrants; population register; network sample; telephone directory

1 Ethnic categories in empirical research

This contribution sets out to provide an overview of the possibilities for determining the "migration background" of population subsets in Germany. The concept of migration background is a specifically German variant of the general sociological construct of foreignness, which describes a condition of perceived difference between groups defined by cultural, geographical, biological, and/or linguistic criteria. Following Weber (1968, 385ff.), migration background is an ethnic category because it derives difference from common descent. Two analytically distinct paradigms play a role in societal discourses and research questions: (a) The *immigration* paradigm assumes that persons who come into a country from outside differ from the established population in some socially meaningful sense because of circumstances preceding international migration. The difference may make them useful, dangerous, or deserving of protection, or in some other manner the object of collective responses. Here the assumption of difference is associated with a belief that immigrants (and even more so their descendants) will become less different through assimilation, although not necessarily always in a linear, automatic, and irreversible manner (Alba & Nee 1997). (b) The *ethnic minorities* paradigm assumes that difference and consequently inequality remain stable over time. Because many ethnic minorities were created by past immigration processes (Font & Méndez 2013, 19) the two paradigms are not mutually exclusive. But they may also result from historic frontier changes dividing a group's settlement area, immigration of a new majority, or state-formation, for example during decolonization. Political debates often circle round the question of whether immigrants have become ethnic minorities. The question is contested because it implies an admission of a persistent social problem and a negative prognosis: ethnic minorities are perceived more strongly than immigrants as essentially different from the majority, weak, and disadvantaged. Despite the different development assumptions, both paradigms describe a relationship between difference and social problems.

The social sciences investigate whether the posited differences exist, whether they change over time, and what consequences they have. Ethnic categorization is crucial at two junctures in empirical research: Firstly, in the sphere of investigation of social inequality, information on the origins of individuals is required in order to discover whether the life chances of ethnic minorities differ from those of the majority (even quite some time after migration) and whether ethnic minorities are

Author's note:

The author would like to thank Christian Babka von Gostomski, Jost Reinecke, two reviewers, and the editors of this journal for valuable suggestions. Translation: Meredith Dale

Direct correspondence to
Kurt Salentin, Institute for Interdisciplinary Research on Conflict and Violence, Bielefeld University, Bielefeld, Germany.
E-mail: kurt.salentin@uni-bielefeld.de

treated differently from the majority in social intercourse on the basis of actual or supposed difference relating to origin, skin color, language, or religion. Many phenomena simply cannot be understood without testing ethnicity as a hypothesis of social difference. Secondly, diversity research, which investigates the effect of the ethnic composition of a socio-geographically defined subpopulation on social cohesion within it (Putnam, 2007; most recently Petermann & Schönwälder 2012; Sturgis, Brunton-Smith, Kuha, & Jackson, 2013), requires corresponding aggregate data in order to calculate diversity metrics for socio-spatial units. Collecting data for such studies requires suitable sampling procedures. This contribution discusses individual-level sampling as the more frequent application, but the discussion is equally applicable to higher levels of aggregation.

The problems of the ethnicity concept described in the classic contribution by Petersen (1980) automatically also apply to its statistical recording. Ethnic categories are vague and multidimensional, and at the same time essentialist, constructed, and not fully amenable to objective characterization, often apparently arbitrary and almost always politically contested, embedded in country-specific circumstances, and subject to rapid change; their semantics are language-specific and their labels change constantly and quickly become pejorative. The sheer diversity makes even a partial overview of the concepts and operationalizations found across the globe an impossible undertaking in the space available, so I will restrict my discussion to a selection of the most important.

Operationalizations of the immigration paradigm (summary: Waters 2014, 17ff.) always relate to the border crossing. A category distinction is frequently made between foreign-born and local-born, based on the assumption that socialization in different contexts before and after the act of migration causes differences in behavior patterns, skills and resources, attitudes, etc. A finer differentiation is provided by the generation model, where the first-generation migrants are identical with the foreign-born, and the local-born comprise the second and subsequent generations. In some cases researchers also distinguish intermediate stages on the basis of age at arrival, such as the generations 1.5 and 1.75, which experienced a "mixed" socialization (Rumbaut 2004). The ethnic minority paradigm uses other categories of its own (for the United States and United Kingdom see Waters 2014, 12ff.). Here the criteria of differentiation are orientated on physiognomy, geographical origin, language, and/or religion. The term "race" is found largely in Anglo-Saxon countries to denote a temporally stable multidimensional categorization according to religious, geographical, cultural, and/or biological criteria such as skin pigmentation (Petersen 1980, 235-36). In continental Europe this concept is rejected as biologistic; in Germany its misuse by the Nazis makes it absolutely unacceptable. The work of authors like Weber (1968), who names belief in common descent as the constitutive feature, and Barth (1969), who describes ethnic identity as a contingent outcome of the interaction of social groups, has highlighted the constructed – and

precisely not biological or otherwise primordial-nature of the differences meant by the term "ethnic group", which may nonetheless have empirically persistent consequences.

How ethnicity is understood in different national contexts, how and whether these ideas can be harmonized, and how they can be translated into sampling procedures in research has to date only been investigated in the scope of regional comparative studies that all point to considerable compatibility problems (Latcheva et. al 2006; Groenewold & Bilsborrow 2008; Groenewold & Lessard-Phillips 2012; Font & Méndez 2013).[1] A systematic international comparison has yet to be conducted.

The German concept of migration background represents one approach to the problem of statistical testing of societally perceived differences between the majority population and population groups created by migration. The approach originates from official statistics, but is also applied in social research. As I will show in detail below, it draws on verifiable features of family migration history and avoids both contested biologistic components and volatile elements such as language use or self-categorization, which are only suitable as dependent variables in assimilation analysis. Alongside a series of specific problems, which I will also come to, migration background is ultimately also subject to the same reservation as any other ethnic categorization: Its use in research can have unwanted effects, as the framing effect risks preparing the ground for an ethnicization of the societal discourse. Here I would merely point to the overview published by the German Institute for Human Rights (Deutsches Institut für Menschenrechte 2008), the passionate debate in France (Cusset, 2008; Le Bras, Racine, & Wieviorka, 2012) and Brubakers' warning against reification (2012).

After defining the target population (section 2), sampling frames and selection criteria are discussed (section 3). The article concludes by considering which options would be optimal and whether they are feasible. The paper claims no validity outside the Federal Republic of Germany. While procedures suited exclusively for subpopulations such as school students or working population with migration background are omitted, the following fundamental discussion should also be helpful for work on such subgroups and for access to other selection frames. Equally, the scope of the article precludes detailed discussion of the legal framework, cost aspects, administrative handling, and software questions, for which the cited literature should be consulted.

1 Hoffmeyer-Zlotnik and Warner (2010) collate items measuring ethnicity in 45 international surveys. But they do not discuss sampling aspects.

2 Target population

The German Federal Office of Statistics (Statistisches Bundesamt 2012, 6) defines "persons with migration background" as "all immigrants who entered the current territory of the Federal Republic of Germany after 1949" (criterion 1), "all non-citizens born in Germany" (criterion 2), and "all Germans born in Germany with at least one parent born abroad or born in Germany as a non-citizen" (criterion 3).[2] One could quibble over the details: It is not apparent why non-citizens pass the "migration background" to all descendents without end, but naturalized citizens do so only to the first subsequent generation. Nonetheless, this definition possesses advantages that increasingly lead researchers to accept it: It is unambiguously operationalizable, functions (unlike most definitions of ethnicity) without self-assessment or controversial attributes such as "race", and runs no risk of turning dependent variables like linguistic competence into elements of the target population definition (and thus of the sampling). Incidentally, even within Germany the official statistical definition of "migration background" varies. A detailed overview is provided by Verband Deutscher Städtestatistiker (2013); here I discuss only the definition used by the Federal Office of Statistics.

This category currently represents 19.5% of the total population, with a rising trend; the total number is 16.0 million (Statistisches Bundesamt, 2012, on the basis of the 2011 microcensus). The proportion is highest among the under-sixes, at almost 35%, falling to less than 10% among the over-75s; in the typically surveyed age group of the over-15s it amounts to 17.6%. Given the extent of heterogeneity of region of origin, it is often necessary to narrow in on individual countries of origin. Alongside 3.2 million *Aussiedler* and *Spätaussiedler* (20.5% of persons with migration background) and 2.96 million people of Turkish origin (18.5%), we are dealing with a multitude of small and very small groups.[3] We must therefore differentiate between the *global* migration background defined by the three criteria above and country-specific categories. A *country-specific* approach is required, for example, to distinguish citizens of EU member-states from third-country nationals. This has consequences for sampling methodology.

The introduction of this concept marked a turning-point. Until the late 1990s only citizenship had been considered relevant in Germany, and any type of ethnic categorization had invited accusations of racism in the context of German history.

2 "alle nach 1949 auf das heutige Gebiet der Bundesrepublik Deutschland Zugewanderten", "alle in Deutschland geborenen Ausländer", "alle in Deutschland als Deutsche Geborenen mit zumindest einem zugewanderten oder als Ausländer in Deutschland geborenen Elternteil"

3 *Aussiedler* are ethnic German immigrants from eastern Europe and the former Soviet Union. They are automatically entitled to German citizenship. *Spätaussiedler* denotes those who arrived in Germany after January 1, 1993. In this contribution *Aussiedler* is used in the general sense covering both.

The migration background concept is based on the crucial insight that the question of social difference did not become obsolete after large numbers of immigrants became naturalized and disappeared from the category of "foreigner." Introducing a definition that includes the descendants of immigrant represents an admission of the necessity of an ethnic dimension. But the authorities were not prepared to expand the reach of the category to include autochthonous minorities. Certain groups living in Germany enjoy a legal status as minorities and are granted special protection as such: the Danes, the Friesians, the Sorbs, and the German Sinti and Roma (Polm 1995). As German citizens not covered by the migration background concept, they fall into a statistical blind spot. Although there are no calls for better documentation of the situations of the first three groups (living in the areas bordering the Netherlands, Denmark, and Poland respectively), the relative lack of data about the Sinti und Roma represents a problem (European Union Agency for Fundamental Rights 2009; Strauß 2011).

Within the population with migration background the official statistics distinguish depending on country of birth between persons with and without personal experience of migration, which is identical with the categorizations of local/foreign-born and first/subsequent generation. A finer differentiation of the sequence of generations is not provided, nor is it possible in the available sampling frames.

3 Sampling frames and demarcation criteria

A sampling procedure must distinguish the sampling frame from which a sample is drawn from the criteria by which migration background is defined (see Table 1), even if it is not possible to realize every combination. The discussion of selection criteria should be helpful to researchers with access to lists of customers, patients, school students, prison inmates, or employees, or to other sampling frames.

Following the logic of the migration background concept, the focus of this contribution lies in identifying minorities created through immigration. I will therefore, as already mentioned, not discuss differentiation criteria that depend on assimilation processes, such as language use or ethnic self-identification. While these are indispensable for the identification of older autochthonous minorities, they are suitable only as dependent variables in the analysis of post-migration integration processes, not as criteria in the sampling process.

Furthermore, I only discuss criteria that are actually available for sampling, and exclude widely used survey items such as place of birth of parents or grandparents.

Table 1: Sampling frames and demarcation criteria

	Demarcation criterion		
Sampling frame	Place of birth	Citizenship	Name
Person-centered network	Snowballing, respondent-driven sampling, quota sampling		
Aggregation center	Center sample technique		
Settlement	Random route with screening		
Telephone directory			Name-based selection in telephone directory
Population Register	Population Register sample by place of birth	Population Register sample by citizenship	Name-based selection in Population Register
Central Register of Foreigners		Central Register of Foreigners sample	

3.1 Sampling frames

Most of the sampling frames discussed below can be regarded as more or less representative *models of the residential* or *target population*. These must be distinguished from person- and object-centered networks centered on individuals or aggregation centers, in which the target population is overrepresented. Strictly speaking networks are not sampling frames, because no lists of persons exist in advance.

Person-centered networks

In themselves, person-centered networks have no specific criteria-defined composition, aside from personal acquaintance. But assuming a certain degree of social homogeneity, we may surmise that the networks of immigrants will include more immigrants of the same origins than those of other persons. Simple snowball sampling, of the kind employed to research rare populations, then involves filtering these networks; in the case at hand by characteristics such as citizenship, or country or region of origin (for the principle see Goodman, 1961). As a rule, quota samples also share the traits of snowball samples, because although interviewers seek their subjects according to sociodemographic characteristics, they do so by successively following the networks or contacts of previous interviewees. This is also associated with a hope that making contact through acquaintances will improve the willingness to participate. One problem arises through the correlation between integration in social networks and probability of inclusion in the sample. Individuals with many contacts will be overrepresented, while isolated individuals are unlikely to be

selected. Schupp and Wagner (1995) describe how, after initial trialing, the snowball method was abandoned for the migrant sample of the German Socio-Economic Panel because of this effect. In a direct comparison between territorial and snowball samples in a World Bank study, McKenzie and Mistiaen (2007) demonstrate that persons in ethnic networks orientate more strongly on their origins. In a sample of Senegalese transnational households with members who migrated to Spain (Beauchemin and González-Ferrer, 2011), snowballing in Senegal was also unfruitful; further, a comparison of the target subjects in Spain with a nominally similar sample from the Spanish population register showed that snowballed subjects possessed stronger ties to the country of origin. Schnell, Hill, and Esser (2005, pp. 303f.) list further general criticisms of quota sampling.

Respondent-driven sampling (RDS; Heckathorn, 1997), which permits mathematical compensation of unequal network participation to achieve probability samples, was conceived as a means to rectify the skewed probability of inclusion. This requires information on the size of the network of the individual whose contacts enter the sample in the respective next step, as well as relational information on the recruitment process, because the network structure must be mapped during analysis. This information is, however, difficult to document anonymously during the survey, because it requires the respondent to reveal names and addresses of contacts. As an alternative, Schonlau and Liebau (2010) describe a method operating with anonymous coupons, where subjects have to contact the interviewer on their own initiative. However, McKenzie and Mistiaen (2007) suggest that migrants are generally more suspicious of strangers and less willing to reveal contact data: contradicting the "snowball" metaphor, generally few new addresses are supplied and many subjects simply refuse to be recruited. Their finding of bias compared to a comparable territorial sample despite RDS correction suggests that while RDS may be able to compensate for differences between persons with more or fewer intra-ethnic contacts, it cannot do so between persons with networks of different ethnic composition. Because other implementation and weighting problems are also unresolved (Schonlau & Liebau, 2010), this method has not to date found broader application in German-speaking countries.

Aggregation centers

In many studies samples are interviewed at *intercept* or *aggregation points*: places frequented by specific minorities, such as shops, government offices, cultural centers, places of worship, or in the vicinity of railway stations. Because of the obvious selectivity of the simple variant toward persons with stronger ethnic ties (for example McKenzie & Mistiaen, 2007), a team led by Gian Carlo Blangiardo has spent twenty years developing a method known as the *center sample technique*, which creates probability samples out of *intercept point samples* (Blangiardo, Migliorati,

& Terzera, 2004; Baio, Blangiardo, & Blangiardo, 2011). The researcher creates a list of known aggregation centers, whose visitors must comprise the heterogeneity of the population of interest, however distorted. In principle other selection criteria apart from geographical origin, such as religious or linguistic characteristics, can be also used to define minorities within the minority. But in fact the available aggregation centers determine the characteristics of the sample, and the researcher's freedom of choice is limited. The relative importance of a single aggregation center is determined by observing the number of visitors; this information flows into the weight given to the interviews conducted there. The subjects themselves must report the frequency with which they visit the aggregation centers, from which, in combination with the aggregation center relevance, a compensatory *ex-post* weighting is calculated. The technique functions only under the precondition that there are no social categories that completely avoid the *intercept points*, as these would have a probability of inclusion of zero. Blangiardo and others (including Groenewold & Bilsborrow, 2008) have proven the method's practicability in several countries, including with undocumented populations. Whether that also applies in a country like Germany, where there is greater manifest pressure of persecution on such groups than in other European states or the United States, cannot currently be said. Nor should there be any illusions about the efficiency of the technique. The German asylum process, the dispersion procedure for *Aussiedler*, and the regionally scattered economic structures attracting labor migrants have combined to geographically disperse many migrant groups. This makes at least national *intercept point* samples a laborious undertaking.

Population-like entities: settlement and telephone directory

For a long time the most popular quasi-model of a residential population comprised the settlements in which it lived. A good approximation of a random sample of the population can be achieved by contacting subjects directly in their homes guided by routing instructions (*random walk* or *random route*) (on the weaknesses: Schnell, 1991). Sometimes the term *area sampling* is also used. Given its relatively large proportion, the population with global migration background is well represented in the resulting samples, without any special measures. The screening effort, which is inverse to the proportion of the population, remains manageable. Anyone wishing to sample persons with global migration background is well advised to apply a standard method for the residential population, estimating costs for a several-fold gross sample (and has no need to read on). Optimization by multi-stage dispropor-

tionate stratification of territorial units means that the gross sample can be smaller than five- or sixfold.[4]

If, however, country-specific groups are to be identified, random route samples become inefficient. For example, for every person of Italian origin (population in Germany 780,000), 128 contacts would be required. And for many groups it is by no means easy to clarify membership of the target population by screening. Optimizing the random walk rules by concentrating fieldwork in areas known to have higher proportions of the target group is less efficient than one might expect, because immigrants in Germany are comparatively unsegregated (Schönwälder & Söhn, 2009). And it produces undesirable consequences. Concentration may be associated with distortions of the social structure and other aspects of selectivity. Restriction to a small number of areas also produces cluster effects (representing an often overlooked reduction of the effective sample size) – a problem that always occurs when clusters are formed in a sampling frame.

There have certainly been applications of area sampling for very small migrant populations (for example, Groenewold & Bilsborrow, 2008), but in multi-stage selection procedures, in which territorial units are stratified by population share. However, in the field sampling plans were quickly revised because of the disproportionate effort involved and snowball elements added or target households arbitrarily substituted, with the result that no probability sample was achieved. Without extremely generous budgets, therefore, immigrant samples using random walk rules are only practicable with considerable concessions in terms of sample quality.

The situation concerning sampling by controlled random dialing of a landline number (Gabler-Häder design) is very similar (Gabler & Häder, 1997). Firstly, 13% of residents of Germany aged 16 and above have no landline number, in which figure single-person households, men, under-30s, low-income groups, and people living in eastern Germany and Berlin are overrepresented (Infas, 2010; Mohorko, de Leeuw, & Hox, 2013, Tables A1, B1). The proportion shows a slightly rising trend (European Commission, 2010, p. 52; Gabler & Häder, 2009). Secondly, the screening effort required for smaller populations is considerable, quite apart from identification problems. The issues are similar for dialing cellphone numbers, although this in general compensates the growing *coverage bias* of the landline network (Mohorko et al., 2013), and for *dual frame* approaches (Callegaro, Ayhan, Gabler, Haeder, & Villaret, 2011). For those reasons these methods will not be discussed further.

4 There is not the space here to go into further requirements, such as language of instruments and staff.

Telephone directory

Ever since machine-readable telephone directories became available, they have been used for sampling, with the possibility of focusing on groups of specific origin using name-based methods (see below). The attractions of this approach are ease of access at very low cost and national coverage in a homogeneous data set. The permissibility of using participant data for surveys is unclear, because under German law personal data may not be processed without consent (see section 4 [1] of the Federal Data Protection Act and several provisions of the Telecommunications Act). Distortion is caused by households that have a landline but no telephone directory entry (Deutschmann & Häder, 2002; Häder, 1996; v. d. Heyde, 1997). The characteristics of unlisted subscribers are known: disproportionately low-income households, couples with a child under the age of 18, households in cities with more than 500,000 inhabitants, and newer telephone numbers (i.e. mobile households, younger people, and tenants rather than owner-occupiers). Households in southern Germany are more likely to have their number listed than those further north. The electronic telephone directory contains fewer entries than the printed version.

Rather less is known about the telephone directory entries of immigrants. In studies of people of Turkish origin conducted by the former Zentrum für Türkeistudien, Sauer and Goldberg (2001, p. 29) find overrepresentation of middle age groups, singles, employed, self-employed, and large households in the telephone directory vis-à-vis the microcensus. Comparing telephone directory samples of French, British, Italian, and Spanish people with the microcensus, Santacreu Fernández, Rother, and Braun (2006) find discrepancies (in some cases massive, and varying between groups) in the distribution of gender, marital status, age, age at migration, migration period, education, and employment status. Salentin (2002) examines the extent to which a Population Register sample of people of Turkish and Serbian origin can be found in the telephone directory, and finds this to be possible for 65% of the people of Turkish origin but only 40% of those from Serbia. Younger people are more likely not to be listed. In the case of immigrants, it is not clear to what extent origin as such affects likelihood of telephone directory entry over and above the sociostructural characteristics.

The strongest argument against the telephone directory is its progressive deterioration. In 1998, according to the suppliers, telephone directory CDs contained 40 million entries. By 2002, with still about 34 million entries, more than 30% of all lines were unlisted in the electronic telephone directory (Deutschmann & Häder, 2002). The 2012 telephone directory CD contains only 26 million entries, while the number of households has increased from 37.5 million in 1998 to 40.4 million in 2011.[5] If we estimate the number of non-private entries in the 2002 data

5 https://www.destatis.de/DE/ZahlenFakten/Indikatoren/LangeReihen/Bevoelkerung/lrbev05.html, accessed December 14, 2012.

and assume for the sake of simplicity a constant number over time, we find that just 36% of households are listed in 2011/2012. While that may be only a rough estimate, it raises grave doubts as to the suitability of the telephone directory as a selection frame.

Apart from the names, telephone directory entries contain no indicators of migration background. On the other hand, the existence of the telephone number facilitates telephone surveying, which makes telephone directory sampling an attractive and popular option in connection with that form of survey. With few exceptions they produce household samples that require a subsequent selection of target person.

Population Register

Each community (district or town) in Germany maintains its own Population Register. Regional registers are not accessible to researchers and there is no national register (see below). Each local authority Population Register contains almost the entire population living within its territory, regardless of citizenship. They exclude only foreign diplomats, members of foreign armed forces, and some undocumented migrants. The authorities differentiate those with legally precarious or non-existent status into: 1. "Clandestines", who evaded border controls when entering the country (and are therefore not included in the Population Register) or hold expired residence permits (overstayers); 2. "Pseudolegals", who acquired a residence permit on the basis of false claims and are likely to be officially registered like the holders of legitimately acquired residence status; and 3. "Persons registered as required to leave" but permitted to stay temporarily, largely rejected asylum-seekers, whose presence is technically illegal but tolerated, and are in principle officially registered (Schneider, 2012). Just because an undocumented person is listed at some address in the Population Register does not, it must be said, mean that they are also contactable. On the basis of detainments listed in the police crime statistics, Schneider (2012) estimates the number of clandestine immigrants in Germany at between 150,000 and 350,000. Depending on the basis of the estimates, Vogel and Aßner (2011) arrive at a corridor of 140,000 to 340,000 or 115,000 to 385,000 for 2010 (for criticism of such estimates, see Schönwälder, Vogel, & Sciortino, 2004). There are no estimates of the size of the "pseudolegal" population (Vogel & Aßner, 2011, p. 22). According to the Federal Office for Migration and Refugees (Schneider, 2012) there were 87,000 persons registered as required to leave Germany in 2010. The Population Register also excludes an unknown number of people who move within Germany without registering, creating a mismatch between resident and registered population, as well as people who move abroad without deregistering, which leads to a net overcounting of the population with migration background. The 1987 census revealed overcounting of individual nationalities of up to 10%. Despite certain

discrepancies, the Population Register is the best available representation of both the overall population and the population with migration background; all in all it can be said to exclude only a relatively small part of the immigrant population.

The use of Population Register data is governed by the Registration Act. Universities are classified as "other official bodies" and may be supplied with more information than other users, including name, address, date and place of birth, and current citizenships. With certain restrictions, this permits conclusions to be drawn about migration background. Data on former citizenships is either not kept or not released. It is thus very easy to identify non-citizens, but only circuitously naturalized citizens (see below). Most Germans with at least one other citizenship fulfil at least one of the criteria of migration background and can be identified directly, assuming they have informed the Population Register of the other citizenship. Although first-generation immigrants can be identified on the basis of place of birth, a finer differentiation of generation status is not possible. The Population Register contains information on date of arrival at the locality but not the date of arrival in Germany. Information on generation status must be requested directly in surveys.

Mixed-nationality marriages cannot usually be identified in the Population Register on the basis of different citizenship within a family. The Population Register does not provide information about family relationships between spouses and other adults. There is one exception: In conjunction with data on minor children, information including nationality can be obtained on the legal guardians, usually meaning the parents.

Under federal law the states decide which agencies are responsible for Population Register affairs. Certain states have established centralized portals or state agencies for the purpose of supplying information that are largely mirrors of the local authority data collections. But these central instances issue only restricted information on individuals. Requests involving more than a single person still requires either the approval of the local authority, or are not permitted at all (the latter being the case in Bavaria, Baden-Württemberg, Hesse, Lower Saxony, North Rhine-Westphalia, and Schleswig-Holstein), so the centralized agencies are of no assistance for sampling purposes. The data pool in the state of Hesse serves exclusively for criminal investigations, and the provision of information to researchers is excluded. A federal population register has been proposed, but can no longer be expected to be established in the foreseeable future. Therefore the procurement of Population Register samples remains as complex and time-consuming as described by Albers (1997). As before, the permissibility of data release must still be negotiated with each individual local authority (Kommune) and the hurdles of heterogeneous data structures and file formats overcome. In the past fees also incurred considerable costs for supplied or processed addresses, as well as (often unforeseeable) costs for programming work. Here improvement is in sight, as an amendment

comes into force in 2015 that provides for information to be supplied free of charge to public bodies, although only from the local authority agencies themselves, not from state portals.

The consequence is that geographically extensive sampling can currently only be conducted with an extraordinary expenditure of resources. If a multi-stage selection procedure is used there is a trade-off between expense and representativeness. Regional concentration leads to cluster effects.

Central Register of Foreigners

The Central Register of Foreigners holds a range of data on all persons without German citizenship living in Germany. It is fed by notifications from the local foreigner registration offices and accumulates a successive dataset that is corrected at infrequent intervals. Problems such as a cumulative overrecording and technical difficulties caused by variables that in some cases constitute only pointers to data held by the local foreigner registration offices need not be discussed in detail here, as there is no legal basis for using the Central Register of Foreigners and as such no grounds for it to serve as a sampling frame for academic research. But even given privileged access the register is of restricted value: its records often fail to match the Population Register (Vogel & Aßner, 2011, p. 24); when a person is naturalized their data are immediately deleted; and as explained below, naturalized citizens and non-citizens differ structurally, creating considerable differences between the Central Register of Foreigners population and immigrants as a whole. Babka von Gostomski and Pupeter (2008, p. 154) summarize the value of samples from the Central Register of Foreigners: "There is therefore no basis for generalizations to all persons with migration background in Germany."

3.2 Demarcation criteria

Citizenship

Operationalizing the characteristic of citizenship for migration background is technically uncomplicated in many databases (criteria 1 and 2), but plainly unsuitable for *Aussiedler*, who are usually German citizens. However, a considerable proportion are identifiable through dual citizenship of their country of origin in Eastern Europe, Central Asia, or Russia (Salentin 2007). The same applies to the children of Aussiedler, who also belong to the target population under criterion 3. German citizens make up 54.9% of the population with migration background (8,771,000 of 15,962,000 persons, Statistisches Bundesamt, 2012, pp. 56ff.). For most immigrated minorities apart from *Aussiedler*, the proportion of German citizens is likely to be smaller, with wide variations; citizens of EU member-states and other industrialized countries are less likely to apply for citizenship, refugees more likely

(Woellert, Kröhnert, Sippel, & Klingholz, 2009, on the basis of the 2005 microcensus). For example, by 2010, 41.42% of people of Turkish origin in Germany had taken German citizenship.[6]

The ensuing problem, alongside quantitative underrecording, is a qualitative distortion of the social structure of the target group if the scope is restricted to non-citizens. A wealth of studies based on the microcensus, the German Socio-Economic Panel, and other samples confirm that naturalized citizens exhibit better socioeconomic parameters and more strongly assimilated attitudes than non-citizens from the same region of origin. They have better school and vocational education, higher occupational status, higher income, and are less likely to be unemployed (Diehl & Blohm, 2008; Gresch & Kristen, 2011; Haug, 2002; Liljeberg, 2011, 2012; Salentin & Wilkening, 2003; Santel, 2008; Seibert, 2008; Seifert, 2011; Woellert et al., 2009). They speak better German (Galonska, Berger, & Koopmans, 2004), are more likely to choose German names for their children (Gerhards & Hans, 2009), less likely to adhere to traditional lifestyles, less likely to live in highly segregated residential environments (Haug & Swiaczny, 2003; Janßen & Schroedter, 2007), and are less religious (Diehl & Koenig, 2009; Liljeberg, 2012). They are happier and gradually cease basing social comparisons on their own past (Brockmann, 2012). The observed differences are plainly in part a consequence of naturalization, for example in the case of income, as Steinhardt (2008) is able to demonstrate. But viewed longitudinally, stronger assimilation is itself a trigger for naturalization (Maehler, 2012). In any case, non-citizen samples systematically exclude the more successful immigrants, for "taking into consideration the different areas and indicators of integration, one can say that naturalized citizens are much better integrated than non-naturalized" (Weinmann, Becher & Babka von Gostomski, 2012, p. 6). In short, naturalization is a dependent variable of integration research that must not be allowed to affect the sampling. Samples based on foreign citizenship produce artifacts. For that reason selection by citizenship is no longer acceptable today.

Where dual citizenship is identified this generally indicates migration background. This information can be drawn from the Population Register. But this is of little help for sampling. Germany has a tradition of preventing multiple citizenship after naturalization, although the rules have recently been relaxed. Also, informa-

6 Own calculation after Statistisches Bundesamt 2012, pp. 56ff. This includes children of at least one parent who immigrated or was born in Germany, who have been German by birth since the jus soli principle was introduced in 2000, and the children of naturalized citizens. Here it was assumed, on the basis of the structure used by the Statistisches Bundesamt (2012, p. 7), that the unlisted figure for Turkey for Category 2.2.2.2.2 (p. 62) (German with at least parent who immigrated or was born in Germany) corresponds to the difference between Category 2.2.2 persons who did not themselves immigrate), and the sum of Categories 2.2.2.1 (non-citizens who did not themselves immigrate) and 2.2.2.2.1 (naturalized citizens who did not themselves immigrate).

tion on additional citizenships is inconsistently recorded. One reason for this is that the acquisition of an additional citizenship is under certain circumstances illegal.

Place of birth

The place or country of birth is, according to criterion 1, a reliable indicator of migration background. Under the Federal Expellee Act, birth in the German territories ceded after World War II is a precondition for recognition as an expellee. For expellees who possess only German citizenship and have no Eastern European sounding names (see below), this makes place of birth the only possibility of identification. That in turn means that their descendants can no longer be identified at all unless parental data can be accessed. While place of birth is equally viable for other migrant groups, it is unfortunately either not recorded or not accessible in many data sets. Utilization also requires country-specific directories of places of birth, and uncoded records cannot usually simply be processed technically (Salentin, 2007, with information on the administrative background), thus incurring programming expenses. There is currently only limited reported experience with sampling based on place of birth (Haug & Sauer, 2006; Ouakkar, 2011; Salentin, 2007; at an experimental stage also Zdrojewski & Schirner, 2005, and an as yet unpublished regional study on familial social support among immigrants from the former Soviet Union by Claudia Vogel and Elena Sommer at the University of Vechta).

Name

The idea that in most countries the names of immigrants differ from those of autochthons is nothing new. In the United States social scientists began identifying minorities by their names in the 1930s (for example Taylor, 1930). However, all name-based methods encounter a number of fundamental problems:

1. Depending on the historical context, immigrants may assimilate their forenames and family names. Swanson (1928, p. 468) reports from the United States: “Karlsson was frequently written Colson, Hedenskog became Hadenscogg, Pehrsson was anglicized into Parsons, and even such a typical Swedish name as Åkerblom in the adjutant general’s reports took the Celtic form of O’Kerblom.” This dimension of assimilation correlates with economic status, as already observed by Beynon (1934, p. 605), who assumes a bias toward unqualified and unemployed caused by the name criterion. In Germany *Aussiedler* are more likely to change *family* names, whereas a correlation between sociostructural integration, education, religiosity, and assimilative choice of first name has been demonstrated for labor migrant families from the Mediterranean region (Gerhards & Hans, 2009).

2. In most societies family names are inherited patrilineally, with the result that exogamy causes a blurring of name boundaries (Mateos, 2007, p. 255). This effect is difficult to quantify. If one examines the self-categorization as *Hispanic* among bearers of typical Spanish names in the 2000 U.S. census (where, however, subjective assimilation processes are also at play) considerable discrepancies are found. While well over 90 percent of those with family names like Velazquez, Juarez, Huerta, and Cervantes identify as *Hispanics*, the figures are considerably lower for Fernandez (80.7%), Delacruz (74.85%), or Duarte (76.56%) (United States Census Bureau, n. d., own calculation).
3. Where names remain constant across several generations, a discrepancy with actual assimilation will inevitably arise: at some point the scientific interest in regarding any bearer of a formerly "foreign" name as "foreign" will no longer be justifiable. In Germany this applies to the names of the Huguenots and the "Ruhr Poles" (Humpert & Schneiderheinze, 2002, p. 189), as well as even older French, Danish, and Dutch names in the border regions, to mention but a few.[7] After all, we do not regard Beethoven as Dutch.[8] Typicalness of names is a time-dependent variable, not an ahistorical constant. In fifty years time the Turkish *Yildiz* (rather than *Yıldız*) will be just as German a name as *Kozlowski* (from *Kozłowski*) already is. The findings of onomastics, a discipline located at the intersection of linguistics, history, and human geography, are therefore useful but not absolute. A principle of temporal/territorial endemicity is the order of the day: A name must be regarded as typical for a country if it existed there before the immigration movement under consideration, however foreign it may sound and whatever its linguistic history. An immigrated name, by contrast, is one that only arrived later. The endemicity of German names could, for example, be tied to the borders of one or both German states in 1950, in order to differentiate the names of labor migrants from the post-war recruitment phase.
4. Countries with identical or related languages generally also have similar names. The more similar the name distributions of authochthons and allochthons, or of two allochthonous groups, the worse the performance of name-based methods (Humpert & Schneiderheinze, 2000, p. 40; Martineau & White, 1998; Mateos, 2007, p. 250).
5. First names have characteristic life cycles (Berger, Bradlow & Braustein, 2012; Berger & Le Mens, 2009; Héran, 2004; Lambert, 2005; Rouxel, 2004)

7 Huguenots escaping persecution in France settled in Germany in the late-seventeenth century; several hundred thousand Poles migrated to the industrializing Ruhr region in the second half of the nineteenth century.

8 Ludwig van Beethoven's forebears came from Flanders, then part of the Netherlands (and today part of Belgium). The Dutch comedian Philip Simon likes to provoke German audiences by referring to Beethoven as a Dutch composer.

and migrate internationally more freely than persons, which means they differentiate less well than family names (Humpert & Schneiderheinze, 2000). Their choice is subject to diverse social influences (Fryer & Levitt, 2004). First names are therefore, despite their smaller total number, no less complex to research and in fact more likely to lead to misclassification and social bias.

Three techniques are available to infer geographical origin from a name:

1. In the reference list or dictionary method (overview: Humpert & Schneiderheinze, 2000; Mateos, 2007) the names in the sample are compared exactly against a list of known geographical origin. Because of the origin of many reference datasets, this is also known as the onomastic method. For the set of names that occur in more than one origin group (on the extent of this, see Humpert & Schneiderheinze, 2002, pp. 190ff.), the probability of their belonging to any particular group can be stated in terms of their relative frequency (Degioanni & Darlu, 2001). Ad hoc reference datasets are sometimes compiled pragmatically according to the principles described by Beynon (1934, p. 605), who speaks casually of "obviously Hungarian names"; sometimes "experts" (members of the target population) are consulted, or specialized service-providers who systematically trawl sources and administer large datasets.[9] The method has the advantage of delivering fairly clear and reliable identification, but drives up the effort and cost of full classification, because of the huge number of names that need to be catalogued. In most countries certain names occur very frequently, very many others only rarely. Fox and Lasker (1983) identify a Pareto distribution for name frequencies. In France before World War II, for example, Darlu, Degioanni, and Ruffié (1997, p. 616) estimate the number of family names at 500,000; the Meertens Instituut cites 300,000 for the Netherlands in 2012,[10] while Kohlheim and Kohlheim (2009, p. 62) speak of more than 500,000 different German names. Because exhaustive lists from reliable sources are available for very few countries, a reference list method always leaves gaps.

The best-suited datasets are openly accessible directories of the residential population before the start of the immigration movements of interest, such as the UK Census of 1881 for the United Kingdom, the French national population register (*répertoire national d'identification des personnes physiques*) provided by INSEE (including name frequencies for every year since 1891 down to the level of département), or the Dutch census (*volkstelling*) of 1947, and with certain restrictions also the German telephone directory (*Reichstelefonbuch*) of 1942. But for most countries there are no reliable and complete directories that allow a distinction between allochthons and autochthons. Borrowing from onomastic studies can

9 The author is aware of Humpert & Schneiderheinze (Duisburg) and Jörg Michael (Hannover).

10 http://www.meertens.knaw.nl/nfb/, accessed December 18, 2012.

prove helpful, to the extent that they (a) use sources that are not too old, (b) contain frequency data, (c) foreground aspects of migration history rather than linguistics. Alternatively, lists of the present population, such as telephone directories, can be used. The difficulty in this latter case is to distinguish names that have already immigrated. In view of the immense diversity of names this overtaxes even so-called experts, who often tend as a result to decide by "feeling." An algorithm can probably accomplish the same task more reliably (see below).

For epidemiological purposes, authors have applied indicators of predictive power from medical testing to the reference list method (Cook, Hewitt, & Milner, 1972, p. 40): sensitivity (proportion of group members correctly classified), specificity (proportion of members of other groups classified as such), and proportions of false positive and false negative classifications (Razum, Zeeb, & Akgün, 2001). Many factors influence the values derived (overview: Mateos, 2007); the multitude of published studies precludes further discussion here. However, presupposing knowledge of the frequency distributions, a simple recommendation can be formulated: A small sample can be acquired with only small losses by choosing a few names with maximum sensitivity and specificity; only if larger populations must be classified is it necessary to resort to less sensitive and specific name lists and reckon with larger screening losses.

2. The n-gram method originating from computer linguistics uses language-specific differences in the frequency of particular sequences of letters, of which words, sentences, names, and other strings are composed (basics: Beesley, 1988; Cavnar & Trenkle, 1994; Schmitt, 1991). For example, the name Meier is broken into the trigrams *mei*, *eie*, and *ier* or the bigrams *me*, *ei*, *ie*, *er*. The n-gram technique is the standard solution for the *language identification* problem for texts in the Internet, although it may misclassify even full texts (as described by Dunning, 1994). By comparing the frequencies of different n-grams in names in different regions, the probability of origin from a particular region can be calculated. The technique has already been in service for some time in commercial database applications,[11] while Schnell et al. (2013) and Susewind (2013) describe sampling applications.

Compared to the reference list method, the n-gram technique has the advantage that it also identifies, with no extra work, alternative transcriptions from non-Latin alphabets and spelling variants that are not yet in the reference dataset, such as Wellen*stain* for Wellen*stein*. But it cannot be persuaded to accept a German name like Brentano, because it knows only n-grams like ano (rather than names as such). Although a systematic comparison of the n-gram and reference list methods has yet to be conducted, it can be assumed that with a very large reference name list the dictionary method will perform better, while n-grams also function well with

11 For example at Intelligent Search Technology Ltd., http://www.name-searching.com/identity-resolution.html.

smaller datasets (whereas the marginal utility of researching many different names falls sharply, because they scarcely alter the n-gram frequency profile). More generally, the short string length of names relativizes the benefit of the n-gram technique, as it produces more frequent misclassifications than with longer passages. Also, the process of splitting into n-grams destroys valuable information about the length of the name.

Another computer-linguistic method is the Soundex algorithm (Russel, 1918) and its successors, long used in U.S. Census contexts, which group homophonic names and thus enable a phonetic search. But because they greatly simplify and are configured for pronunciation in a specific language, they are of little use for name identification.

3. No studies applying the great progress made in bioinformatic sequence analysis over the past two decades to name analysis have yet been published. The sequence of letters in names can in principle be investigated using the same methods applied to nucleotides in DNA. Thus techniques based on *edit distance* algorithms (after Levenshtein, 1966) are suited for error-tolerant reference list comparison preserving information on string length. For classification of origin, multiple string comparison methods (Gusfield, 2008, Chap. 14) may prove more useful. In biology, these are used to assign individual proteins to known protein families according to the nucleotide sequence, and are analogously able to assign names to particular regions. The sequence analysis methods of social science (overview: Abbott & Tsay, 2000) are not directly applicable here, as they would seek to discover through cluster analysis those commonalities that are already known for names.

4 Summary and discussion

Today, the state of research in Germany allows reasonably precise statements to be made about the properties of immigrant and minority samples in relation to sampling frame and applied selection criterion. Snowball samples cause bias in relation to social integration. The correction in *respondent-driven sampling* raises problems of trust in application that will often be unresolvable. Access through the classic route of survey research, random selection of homes or telephone number, is in principle possible, especially if multi-stage selection methods are applied. The expected distortions do not exceed the usual extent for surveys of the residential population. But without truly generous budgets, the researcher will be dealing with regional restrictions and cluster effects. For most small target groups the method is economically impractical. The telephone directory is increasingly shrinking to a residual list of older connections used by geographically immobile persons, who demonstrate a multitude of peculiarities compared to the population as a whole. It is therefore increasingly difficult to argue that weighting can compensate the obvious

biases. Otherwise, telephone directories offer only the name as selection criterion. For explorative purposes telephone directory samples stand out for their low cost and easy availability.

Weighting is also required in *respondent-driven sampling* and with the *center sample* technique. Weighting assumes the elements of underrepresented combinations of categories to be representative of the entire corresponding category of population. Any violation of that assumption can actually worsen the bias of a sample. It certainly cannot compensate all the global or selective biases in a sample.

Apart from the undocumented, the Population Register includes all relevant groups. In theory it provides all the characteristics that identify migration background. Nonetheless, its use encounters a real difficulty: Although name and place of birth may be *supplied*, *selection* according to these characteristics is legally controversial. While many local authorities class this as permissible, legal experts consulted by the author regard it as a "gray zone." Although researchers may undertake post-hoc categorization of samples if in doubt, the attractive route of direct selection from the Population Register appears not unproblematic at the present time. Selection by citizenship is regarded as acceptable, but provides no viable substitute. Pending clarification of the legal situation, researchers are left to negotiate individually whether use of the two most useful characteristics is possible. Furthermore, the decentralized nature of the Population Register continues to create effort and expense. If there was a samplable national population register, one would have to worry less about other sources. Without such a solution, nationwide surveys are more or less unaffordable for small projects. For studies at city level or in selected settlement types the Population Register is the means of choice. This assessment of the German situation confirms the observations of Méndez and Font (2013, 276f.), who regard population registers as the best sampling frame in Europe. According to their criteria, the drawbacks of the German Population Register are legal uncertainty, lack of information about the country of birth of the parents of adults (which is crucial for clarifying generation status and is available for example in Sweden, Denmark, and the Netherlands), and age at immigration. In view of heightened public wariness in Germany about the collection of data that is not essential for administrative purposes, no change is to be expected here in the foreseeable future. But in comparison with the United Kingdom, France, and Italy, the options available to German researchers are actually comparatively good.

The Central Register of Foreigners excludes by definition significant parts of the population with migration background, including the best-integrated, so today one would no longer wish to call for it to be opened for research purposes. Findings in other countries suggest that ex-post weighting makes *intercept point samples* well suited to reach very specific populations that are not recorded in lists, as long as absolutely all members of the target group visit aggregation centers.

It is well known that citizenship only incompletely represents migration background. The qualitative difference between non-citizens and immigrants weighs more heavily than the quantitative, as the best-integrated immigrants tend to be the ones that naturalize. Place of birth is indispensable for identifying *Aussiedler* and functions as a validating criterion for all first-generation immigrants. However, because any German-born child of a first-generation immigrant (or of later non-naturalized generations) belongs to the target population, place of birth abroad is insufficient as a sole criterion. The criterion that performs best overall is the name, which is why name-based methods have become established as the "standard instrument" (Haug, Müssig, & Stichs, 2009, p. 41) for immigrant surveys. Depending on the case, various methods are available for inferring origin from name. Considerable scope for technological innovation remains, and all the methods are more or less error-prone, meaning that gross samples must always be overdimensioned. All the name-based techniques serve well as heuristic approaches, and considerably reduce the cost and complexity of screening.

Nonetheless, the outcome of this review is sobering. There is in theory an ideal solution for sampling the population with migration background, namely via a multiplicity of Population Registers and name recognition, supplemented by citizenship and place of birth. But firstly, such samples are costly (and integration researchers must argue this assertively vis-à-vis funders). Secondly, the legal basis for gathering them is questionable. There is presently no acceptable methodological repertoire to match the considerable public interest in integration. It remains to hope that political decision-makers understand this difficulty.

This contribution has not undertaken an international comparison, firstly for considerations of space, but also because for many countries there is insufficient literature on the availability of data on ethnicity in the available sampling frames, legal considerations affecting access, and experience from research practice. I would welcome an expansion of the systematization of sampling frames and demarcation criteria presented here to cover the situation in other countries.

References

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. Sociological Methods & Research, 29(3), 3-33.

Alba, R., & Nee, V. (1997). Rethinking assimilation theory for a new era of immigration. *International Migration Review* 31(4), 827-74.

Albers, I. (1997). Einwohnermelderegister-Stichproben in der Praxis: Ein Erfahrungsbericht. In S. Gabler & J. H. P. Hoffmeyer-Zlotnik (Eds.), *Stichproben in der Umfragepraxis* (pp. 117-126). Opladen: Westdeutscher Verlag.

Babka von Gostomski, C., & Pupeter, M. (2008). Zufallsbefragung von Ausländern auf Basis des Ausländerzentralregisters. *mda*, 2(2), 149-177.

Baio, G., Blangiardo, G. C., & Blangiardo, M. (2011). Centre sampling technique in foreign migration surveys: A methodological note. *Journal of Official Statistics*, 27(3), 451-465.

Barth, F. (1969). *Ethnic groups and boundaries: The social organization of culture difference*. Bergen and Oslo: Universitetsforlaget.

Beauchemin, C., & González-Ferrer, A. (2010). Sampling international migrants with origin-based snowballing method: New evidence on biases and limitations. *Demographic Research*, 25, 103-124.

Beesley, K. R. (1988). *Language identifier: A computer program for automatic natural-language identification of on-line text*. Proceedings of the 29th Annual Conference of the American Translators' Association, pp. 47-54.

Berger, J., & Le Mens, G. (2009). How adoption speed affects the abandonment of cultural tastes. *PNAS*, 106(20), 8146-8150.

Beynon, E. D. (1934). Occupational succession of Hungarians in Detroit. *American Journal of Sociology*, 39(5), 600-610.

Blangiardo, G. C. (2008). *The centre sampling technique in surveys on foreign migrants: The balance of a multi-year experience*. Joint UNECE/Eurostat Work Session on Migration Statistics, Geneva, Switzerland, March 3-5, 2008.

Blangiardo, G. C., Migliorati, S. & Terzera, L. (2004). Center Sampling: from Applicative Issues to Methodological Aspects. Bari: Atti della XLII Riunione Scientifica (Università di Bari, 9-11 giugno 2004).

Brockmann, H. (2012). Das Glück der Migranten – eine Lebenslaufanalyse zum subjektiven Wohlbefinden von Migranten der ersten Generation in Deutschland. Berlin: Deutsches Institut für Wirtschaftsforschung.

Brubaker, R. (2012). Categories of analysis and categories of practice: A note on the study of Muslims in European countries of immigration. *Ethnic and Racial Studies*, 36(1), 1-8.

Callegaro, M., Ayhan, O., Gabler, S., Haeder, S., & Villar, A. (2011). Combining landline and mobile phone samples: A dual frame approach. Mannheim: GESIS.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. Proceedings of SDAIR-94, 3d Annual Symposium on Document Analysis and Information Retrieval. Las Vegas.

Cook, D., Hewitt, D., & Milner, J. (1972). Uses of the surname in epidemiologic research. American Journal of Epidemiolology, 95(1), 38-45.

Cusset, Y. (2008). La discrimination et les statistiques « ethniques »: éléments de débat. *Informations sociales 4*, 108-116.

Darlu, P., Degioanni, A., & Ruffié, J. (1997). Quelques statistiques sur la distribution des patronymes en France. *Population*, 52(3), 607-634.

Degioanni, A., & Darlu, P. (2001). A Bayesian approach to infer geographical origins of migrants through surnames. *Annals of Human Biology,* 28(5), 537-545.

Deutscher Bundestag. (2011). *Entwurf eines Gesetzes zur Fortentwicklung des Meldewesens (MeldFortG)*. Berlin: Bundestagsdrucksache 17/7746.

Deutsches Institut für Menschenrechte. (2008). *Datenerhebung zum Erweis ethnischer Diskriminierung: Fachgespräch des Deutschen Instituts für Menschenrechte*, 12. Juni 2008. Berlin.

Deutschmann, M., & Häder, S. (2002). Nicht-Eingetragene in CATI-Surveys. In S. Gabler & S. Häder (Eds.), *Telefonstichproben: Methodische Innovationen und Anwendungen in Deutschland* (pp. 68-84). Münster: Waxmann.

Diehl, C., & Koenig, M. (2009). Religiosität türkischer Migranten im Generationenverlauf: Ein Befund und einige Erklärungsversuche. *ZfS*, 38(4), 300-319.

Diehl, C., & Blohm, M. (2008). Die Entscheidung zur Einbürgerung: Optionen, Anreize und identifikative Aspekte. In F. Kalter (Ed.), *Migration und Integration* (pp. 437-464). Wiesbaden: VS-Verlag.

Dunning, T. (1994). *Statistical Identification of Language*. Las Cruces, New Mexico: New Mexico State University.

European Commission. (2010). *Special Eurobarometer 335: E-communications household survey*. Brussels.

European Union Agency for Fundamental Rights (2009). *EU-MIDIS at a glance: Introduction to the FRA's EU-wide discrimination survey.* Vienna.

Font, J., & Méndez, M. (2013). Introduction: The methodological challenges of surveying populations of immigrant origin. In: Font, J., & Méndez (Eds.), M., Surveying Ethnic Minorities and Immigrant Populations (pp. 11-41). Amsterdam: Amsterdam University Press.

Font, J., & Méndez, M., (eds.,). (2013). *Surveying ethnic minorities and immigrant populations: Methodological challenges and research strategies.* Amsterdam: Amsterdam University Press.

Fryer, R. G., & Levitt, S. D. (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, 119(3), 767-805.

Gabler, S., & Häder, S. (1997). Überlegungen zu einem Stichprobendesign für Telefonumfragen in Deutschland. *ZUMA-Nachrichten*, 21(41), 7-18.

Gabler, S., & Häder, S. (2009). Die Kombination von Mobilfunk- und Festnetzstichproben in Deutschland. In M. Weichbold, J. Bacher, & C. Wolf (Eds.), *Umfrageforschung: Herausforderungen und Grenzen* (pp. 239-252). Wiesbaden: VS-Verlag.

Galonska, C., Berger, M., & Koopmans, R. (2004). *Über schwindende Gemeinsamkeiten: Ausländer- versus Migrantenforschung: Die Notwendigkeit eines Perspektivenwechsels zur Erforschung ethnischer Minderheiten in Deutschland am Beispiel des Projekts „Die Qualität der multikulturellen Demokratie in Amsterdam und Berlin"*. Berlin: Wissenschaftszentrum Berlin für Sozialforschung.

Gerhards, J., & Hans, S. (2009). From Hasan to Herbert: Name-Giving Patterns of Immigrant Parents between Acculturation and Ethnic Maintenance. *ajs*, 114(4), 1102-1128.

Goodman, L. A. (1961). Snowball Sampling. *The Annals of Mathematical Statistic*s, 32(1), 148-170.

Gresch, C., & Kristen, C. (2011). Staatsbürgerschaft oder Migrationshintergrund? Ein Vergleich unterschiedlicher Operationalisierungsweisen am Beispiel der Bildungsbeteiligung. *ZfS*, 40(3), 208-227.

Groenewold, G., & Bilsborrow, R. E. (2008). Design of samples for international migration surveys: Methodological considerations and lessons learned from a multi-country study in Africa and Europe. In C. Bonifazi, M. Okólski, J. Schoorl, & P. Simon (Eds.), *International migration in Europe: New trends and new methods of analysis* (pp. 293-312). Amsterdam: Amsterdam University Press.

Groenewold, G., & Lessard-Phillips, L. (2012). Research methodology. In: M. Crul, J. Schneider & F. Lelie (Eds.): The European Second Generation Compared: Does the Integration Context Matter? (pp. 39-56). Amsterdam: Amsterdam University Press.

Gusfield, D. (2008). *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge: Cambridge University Press. (Reprint, first published 1997)

Häder, S. (1996). Wer sind die „Nonpubs"? Zum Problem anonymer Anschlüsse bei Telefonumfragen. *ZUMA-Nachrichten*, 20(39), 45-68.

Haug, S. (2002). Familienstand, Schulbildung und Erwerbstätigkeit junger Erwachsener. Eine Analyse der ethnischen und geschlechtsspezifischen Ungleichheiten – Erste Ergebnisse des Integrationssurveys des BiB. *Zeitschrift für Bevölkerungswissenschaft*, 27(1), 115-144.

Haug, S., Müssig, S., & Stichs, A. (2009). *Muslimisches Leben in Deutschland: im Auftrag der Deutschen Islam Konferenz*. Nuremberg: BAMF.

Haug, S., & Sauer, L. (2006). Zuwanderung und räumliche Verteilung von Aussiedlern und Spätaussiedlern in Deutschland. *Zeitschrift für Bevölkerungswissenschaft*, 31(3–4), 413-442.

Haug, S., & Swiaczny, F. (2003). Migrations- und Integrationsforschung in der Praxis: Das Beispiel BiB-Integrationssurvey. *Standort – Zeitschrift für angewandte Geographie*, 27(1), 16-20.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2), 174-199.

Héran, F. (2004). Un classique peu conformiste: la cote des prénoms. *Revue européenne des sciences sociales*, 42(129), 159-178.

Hoffmeyer-Zlotnik, J. H. P., & Warner, U. (2010). *Measuring ethnicity in cross-national comparative survey research*. Bonn: GESIS.

Humpert, A., & Schneiderheinze, K. (2000). Stichprobenziehung für telefonische Zuwanderumfragen: Einsatzmöglichkeiten der Namenforschung. *ZUMA-Nachrichten*, 24(47), 36-59.

Humpert, A., & Schneiderheinze, K. (2002). Stichprobenziehung für telefonische Zuwandererumfragen: Praktische Erfahrungen und Erweiterung der Auswahlgrundlage. In S. Gabler & S. Häder (Eds.), *Telefonstichproben: Methodische Innovationen und Anwendungen in Deutschland* (pp. 187-208). Münster: Waxmann.

Infas. (2010). *Pressemitteilung: Gut jeder Zehnte ohne Festnetzanschluss im Haushalt*. Bonn: Institut für angewandte Sozialwissenschaft.

Janßen, A., & Schroedter, J. H. (2007). Kleinräumliche Segregation der ausländischen Bevölkerung in Deutschland: Eine Analyse auf der Basis des Mikrozensus. *ZfS*, 36(6), 453-472.

Kohlheim, R., & Kohlheim, V. (2009). *Duden – Die wunderbare Welt der Namen*. Mannheim: Duden.

Latcheva, R., Lindo, F., Machado, F., Pötter, U., Salentin, K., & Stichs, A. (2006). *Immigrants and Ethnic minorities in European cities: Life-courses and quality of life in a world of limitations. Final report*. Vienna: Centre for Social Innovation (http://www.equi.at/dateien/LIMITS_FinalReport.pdf).

Lambert, J.-C. (2005). *Lucas et Léa, prénoms préférés des Auvergnats*. Paris: INSEE.

Le Bras, H., Racine, J.-L., & Wieviorka, M. (2012). *National debates on race statistics: Towards an international comparison*. Paris: Fondation Maison des sciences de l'homme.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics – Doklady*, 10(8), 707-710.

Liljeberg, H. (2011). *Repräsentative Studie zum Integrationsverhalten von Türken in Deutschland: Ergebnisse einer telefonischen Repräsentativbefragung.* Berlin: LILJEBERG Research International.

Liljeberg, H. (2012). *Deutsch-Türkische Lebens- und Wertewelten 2012: Ergebnisbericht zu einer repräsentativen Befragung von Türken in Deutschland*. Berlin: INFO Research Group.

Maehler, D. B. (2012). *Akkulturation und Identifikation bei eingebürgerten Migranten in Deutschland*. Münster: Waxmann.

Martineau, A., & White, M. (1998). What's not in a name. The accuracy of using names to ascribe religious and geographical origin in a British population. *Journal of Epidemiology and Community Health*, 52, 336-337.

Mateos, P. (2007). A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place*, 13(4), 243–263.

McKenzie, D. J., & Mistiaen, J. (2007). *Surveying migrant households: A comparison of census-based, snowball, and intercept point surveys*, IZA Discussion Paper 3173. Bonn: IZA.

Méndez, M., & Font, J. (2013). Surveying immigrant populations: Methodological strategies, good practices and open questions. In: Font, J., & Méndez (Eds.), M., *Surveying Ethnic Minorities and Immigrant Populations* (pp. 271-290). Amsterdam: Amsterdam University Press.

Mohorko, A., de Leeuw, E., & Hox, J. (2013). Coverage bias in European telephone surveys: Developments of landline and mobile phone coverage across countries and over time. *Survey Methods*, http://surveyinsights.org/?p=828

Ouakkar, A. (2011). *Engagiert oder distanziert? Elterliche Überzeugungen und Praktiken beim häuslichen Lernen in autochthonen und russlanddeutschen Familien*. Degree thesis, University of Bielefeld, Fakultät für Psychologie.

Petermann, S., & Schönwälder, K. (2012). Gefährdet Multikulturalität tatsächlich Vertrauen und Solidarität? Eine Replik. *Leviathan*, 40(4), 482-490.

Petersen, W. (1980). Concepts of Ethnicity. In: S. Thernstrom, A. Orlov & O. Handlin (Eds.), Harvard Encyclopedia of American Ethnic Groups (pp. 234-242). Cambridge, Mass.: Harvard University Press.

Polm, R. (1995). Minderheit. In C. Schmalz-Jacobsen & G. Hansen (Eds.) Ethnische Minderheiten in der Bundesrepublik Deutschland (pp. 340-342). München: Beck.

Putnam, R. D. (2007). E pluribus unum: Diversity and community in the twenty-first century. The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, 30(2), 137-174.

Razum, O., Zeeb, H., & Akgün, S. (2001). How useful is a name-based algorithm in health research among Turkish migrants in Germany? *Tropical Medicine and International Health*, 6(8), 654-661.

Rouxel, M. (2004). *Prénoms: De l'influence des modes à la recherche d'originalité*. Paris: INSEE.

Rumbaut, R. G. (2004). Ages, life stages and generational cohorts: Decomposing the immigrant first and second generations in the United States. *International Migration Review* 38(3), 1160-1205.

Russel, R. C. (1918). US patent No. 1,261,167. Retrieved from European Patent Office, http://worldwide.espacenet.com/publicationDetails/originalDocument?CC=US&

NR=1261167A&KC=A&FT=D&ND=&date=19180402&DB=&locale=en_EP, October 24, 2013.

Salentin, K. (2002). Zuwandererstichproben aus dem Telefonbuch: Möglichkeiten und Grenzen. In S. Gabler & S. Häder (Eds.), *Telefonstichproben: Methodische Innovationen und Anwendungen in Deutschland* (pp. 164-186). Münster: Waxmann.

Salentin, K. (2007). Die Aussiedler-Stichprobenziehung: *mda: Zeitschrift für Empirische Sozialforschung*, 1(1), 25-44.

Salentin, K., & Wilkening, F. (2003). Ausländer, Eingebürgerte und das Problem einer realistischen Zuwanderer-Integrationsbilanz. *KZfSS*, 55(2), 278-298.

Santacreu Fernández, O., Rother, N., & Braun, M. (2006). Stichprobenziehung für Migrantenpopulationen in fünf Ländern: Eine Darstellung des methodischen Vorgehens im PIONEUR-Projekt. *ZUMA-Nachrichten*, 30(59), 72-88.

Santel, B. (2008). *Integrationsmonitoring: Neue Wege in Nordrhein-Westfalen*. Osnabrück: Rat für Migration e. V.

Sauer, M., & Goldberg, A. (2001). *Die Lebenssituation und Partizipation türkischer Migranten in Nordrhein-Westfalen: Ergebnisse der zweiten Mehrthemenbefragung*. Münster: LIT.

Schneider, J. (2012). *Maßnahmen zur Verhinderung und Reduzierung irregulärer Migration*. Nürnberg: BAMF.

Schnell, R. (1991). Wer ist das Volk? Zur faktischen Grundgesamtheit bei „allgemeinen Bevölkerungsumfragen“: Undercoverage, Schwererreichbare und Nichtbefragbare. *KZfSS*, 43(1), 106-137.

Schnell, R., Gramlich, T., Bachteler, T., Reiher, J., Trappmann, M., Smid, M., & Becher, I. (2013). Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten. *mda: Zeitschrift für Empirische Sozialforschung*, 7(1), 5-33.

Schnell, R., Hill, P. B., & Esser, E. (2005). *Methoden der empirischen Sozialforschung*. Munich: Oldenbourg.

Schonlau, M., & Liebau, E. (2010). *Respondent driven sampling*. Berlin: DIW.

Schupp, J., & Wagner, G. (1995). Die Zuwanderer-Stichprobe des Sozio-oekonomischen Panels (SOEP). *Vierteljahrshefte zur Wirtschaftsforschung*, 64(1), 16-25.

Schönwälder, K., & Söhn, J (2009). Immigrant Settlement Structures in Germany: General Patterns and Urban Levels of Concentration of Major Groups. *Urban Studies*, 46(7), 1439-1460.

Schönwälder, K., Vogel, D., & Sciortino, G. (2004). *Migration und Illegalität in Deutschland*. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB).

Seibert, H. (2008). *Junge Migranten am Arbeitsmarkt: Bildung und Einbürgerung verbessern die Chancen*. Nuremberg: IAB.

Seifert, W. (2011). *Integration von Zugewanderten in Nordrhein-Westfalen: Eingebürgerte und ausländische Bevölkerung im Vergleich*. Düsseldorf: Information und Technik Nordrhein-Westfalen.

Statistisches Bundesamt. (2012). *Bevölkerung und Erwerbstätigkeit: Bevölkerung mit Migrationshintergrund: Ergebnisse des Mikrozensus 2011*. Wiesbaden: Statistisches Bundesamt.

Steinhardt, M. F. (2008). *Does citizenship matter? The economic impact of naturalizations in Germany*. Hamburg: Hamburg Institute of International Economics.

Strauß, D. (2011). Zur Bildungssituation von deutschen Sinti und Roma. *Aus Politik und Zeitgeschichte*, 22-23, 48-54.

Sturgis, P., Brunton-Smith, I., Kuha, J., Jackson, J. (2013). *Ethnic diversity, segregation and the social cohesion of neighbourhoods in London*. *Ethnic and Racial Studies*. doi:10.1080/01419870.2013.831932.

Susewind, R. (2013). Namematching refined. Blogged research note. http://www.raphael-susewind.de/blog/2013/namematching-refined.

Swanson, R. W. (1928). The Swedish surname in America. *American Speech*, 3(6), 468-477.

Taylor, P. S. (1930). Some aspects of Mexican immigration. *Journal of Political Economy*, 38(5), 609-615.

United States Census Bureau. (n.d.). Surnames occurring 100 or more times. Machine-readable data file. Washington.

Verband Deutscher Städtestatistiker (Eds.). (2013). *Migrationshintergrund in der Statistik: Definitionen, Erfassung und Vergleichbarkeit*. Cologne.

Vogel, D., Aßner, M. (2011). *Umfang, Entwicklung und Struktur der irregulären Bevölkerung in Deutschland*. Nürnberg: BAMF.

von der Heyde, C. (1997). Random-Route und Telefon: Struktur von Telefonhaushalten. In S. Gabler & J. H. P. Hoffmeyer-Zlotnik (Eds.), *Stichproben in der Umfragepraxis* (pp. 196-206). Opladen: Westdeutscher Verlag.

Waters, M. C. (2014). Defining difference: The role of immigrant generation and race in American and British immigration studies. *Ethnic and Racial Studies*. doi:10.1080/01419870.2013.808753

Weber, M. (1968). *Economy and society*. Berkeley: University of California Press.

Weinmann, M., Becher, I., & Babka von Gostomski, C. (2012). *Einbürgerungsverhalten von Ausländerinnen und Ausländern in Deutschland sowie Erkenntnisse zu Optionspflichtigen: Ergebnisse der BAMF-Einbürgerungsstudie 2011*. Nürnberg: Bundesamt für Migration und Flüchtlinge.

Woellert, F., Kröhnert, S., Sippel, L., & Klingholz, R. (2009). *Ungenutzte Potenziale: Zur Lage der Integration in Deutschland*. Berlin: Berlin-Institut für Bevölkerung und Entwicklung.

Zdrojewski, S., & Schirner, H. (2005). Segregation und Integration: Entwicklungstendenzen der Wohn- und Lebenssituation von Türken und Spätaussiedlern in der Stadt Nürnberg. In Verbundpartner „Zuwanderer in der Stadt“ (Eds.), *Zuwanderer in der Stadt: Expertisen zum Projekt* (pp. 75-146). Darmstadt: Schader-Stiftung.

methods, data, analyses | Vol. 8(1), 2014, pp. 53-78 DOI: 10.12758/mda.2014.003

The Five Dimensions of Muslim Religiosity. Results of an Empirical Study

Yasemin El-Menouar
Bertelsmann Stiftung

Abstract

In this paper a new instrument measuring Muslim religiosity is presented. Drawing on Glock's multidimensional concept of religiosity, a quantitative paper-and-pencil study among 228 Muslims living in German cities was carried out. While previous studies have often simply translated indicators measuring Christian religiosity into Islamic terminology, this study applies Glock's model taking into account the specific characteristics of Islamic piety. In particular, the function of his fifth dimension of secular consequences was modified: Contrary to other denominations, in Islam this dimension is regarded to be as unique and independent as the other four. Empirical findings confirm this assumption. Applying principal component analysis with oblimin rotation yields a five-dimensional structure of Muslim religiosity: 1. Basic religiosity, 2. Central duties, 3. Religious experience, 4. Religious knowledge, and 5. Orthopraxis. Further statistical analysis indicates that the scales are reliable and internally valid.

Keywords: Muslim religiosity, measurement, multidimensionality, Charles Y. Glock, Islam, principal component analysis

1 Introduction

The role of Islam has become a key public and political concern in recent years and this development has resulted in a growing interest in Muslim religiosity as the subject of empirical social research. How religious are Muslims? What influence does Muslim piety exert on political opinions? Is Muslim religiosity an obstacle to social integration? There is considerable demand for answers to questions like these, and to date several surveys have been carried out among Muslims living either in the Muslim world or in the Western diaspora (e.g. Pew Research Center 2007; Brettfeld and Wetzels 2007; Hassan 2008). However, the results of these studies provide only limited insight into the aspects outlined above. There is still little knowledge about Islamic religiosity and its associations with other characteristics of Muslims. A basic prerequisite for investigating the varieties of Muslim religiosity is finding an adequate measurement instrument. Yet the measuring instruments applied so far appear to suffer from five main problems: I. A conception of Islamic piety as a one-dimensional construct, II. A one-to-one translation of Christian measures into Islamic terminology, III. Interpretation of research results within a framework of Western or Christian concepts of religiosity, IV. Use of indicators measuring more than religiosity, and V. A lack of statistical estimates of reliability and validity. In the following section these problems will be discussed in more detail.

1.1 Main problems of previous indicators measuring Muslim religiosity

I. Most research conceptualizes *Muslim faith as an inherent monolithic bloc*. Thus, Islamic religiosity appears to be a one-dimensional construct. Some studies use indicators that stress different single aspects of Muslim religiosity, e.g. belief in Allah, fasting at Ramadan, or just the religious self-assessment of the respondents (e.g. Mogahed 2009, Sen and Sauer 2006). Other studies adopt several indicators, combining them in an additive index (Kecskes 2000). However, the use of ever changing indicators to measure an imaginary one-dimensional Muslim religiosity unsurprisingly yields results that differ from study to study and even contradict each other: Whereas some observe a decline in religiosity (e.g. Meng 2004) others argue that “Re-Islamization” is on the rise (e.g. Heitmeyer 1997). Furthermore,

Direct correspondence to
Yasemin El-Menouar, Bertelsmann Stiftung, Project Manager, Programm Lebendige Werte, Carl-Bertelsmann-Str. 256, 33311 Gütersloh
E-mail: yasemin.el-menouar@bertelsmann-stiftung.de

Eilers et al. (2008) have already shown that the indicators used are not necessarily associated with each other[1].

II. Another problem is that *instruments measuring Christian religiosity are often simply translated into Islamic terminology.* The following example is that of an item taken from an international study which is used as an indicator for Islamic belief based on Stark and Glock's (1968) model of religiosity (Hassan 2008): "Only those who believe in the Prophet Mohammad can go to heaven". The original wording of the item developed by Glock to measure Christian religious belief is as follows: "Belief in Jesus Christ as Saviour is absolutely necessary for salvation". However, unlike Jesus, Mohammad has no divine status. He is seen as a role model for Muslims rather than as somebody to be believed in. In fact, orthodox Muslims are opposed to celebrating Mohammad's birthday because they perceive such an act as a form of polytheism. In short, a simple translation of indicators from the respective items of other religions can lead to measurement problems and to false interpretations of results.

Another example is the use of mosque attendance or membership of a mosque as an indicator for piety in the same way that church attendance is used in the case of Christian religiosity. It has been shown that church attendance is a good indicator for Christian religiosity (Jagodzinski and Dobbelaere 1993, Pollack and Pickel 2007). Here, religiosity coincides with church attendance. However, in Muslim piety, mosque attendance has a genuinely different role. First, it remains highly linked to gender. It is mostly men who go to a mosque, e.g. for the Friday prayer. Second, mosque attendance is not an inherent part of Muslim piety as such. A pious Muslim is connected with Allah in a direct way and does not need the mosque or the Imam as an agent intermediary. Therefore, the mosque has a genuinely different role from that of the church. Furthermore, membership of a mosque is not compulsory as it is in the case of Christianity. Most Muslims, even very pious Muslims, are not formal members of a mosque.[2]

III. A further problem that stems from the translation of items from other religious cultures is that *results are often interpreted within the framework of Christian or Western concepts of religiosity.* One example that illustrates this problem is the common misinterpretation of a central finding that can be found in many studies: Pointing to the strong and stable belief in Allah that is found consistently for the great majority of Muslims[3] (e.g Esmer 2002.) Islam is presumed to suffer from reli-

1 Huber has already dealt with this problem and developed a multidimensional instrument which can be applied for different denominations as well as religiosity without denominational adherence (Huber 2003 and 2009).

2 As Tezcan (2008) states the role of the mosque is changing.

3 According to results of the World Value Survey 2000, these show that in Turkey, Egypt and Morocco over 90 percent of respondents say that God is important in their lives (categories 8 to 10). This figure did not decrease in 2005 (basis: own calculations).

gious stagnation which is often interpreted as a lack of social progress and secularization[4] (Bracke and Fadil 2008). Moreover, various studies falsely interpret agreement on the part of Muslims with central aspects of Islam as an endorsement of orthodoxy (Hassan 2008). In Western approaches, orthodoxy is defined as "(…) the extent to which the traditional supernatural doctrines are acknowledged" (Glock 1968). However, this is not true in the case of Islamic religiosity. As Pace (1998) states, being faithful is self-evident and natural within the Muslim population. This can be considered to hold true almost universally and represents an aspect shared by the great majority of Muslims. Secular Muslims as well as very pious Muslims can show the same degree of Islamic belief but may differ concerning other aspects of Muslim religiosity. Therefore, we are unlikely to find much variance when using indicators like the belief in Allah. However, a history of consistently strong belief in Allah does not mean that religious dynamics within Islam are absent. The focus on and expectation of familiar processes well known to exist in Christianity obstruct the view on other dynamics and variation that go beyond a traditional Christian outlook.

IV. A fourth problem is the use of *indicators measuring more than Muslim religiosity.* Many items are in fact political in nature (Heitmeyer 1997) and can be traced back to a growing interest in political Islam and Islamism. For instance, the attitude toward the morality of Western societies (e.g. Brettfeld and Wetzels 2007) could be interesting as a possible correlate of Muslim religiosity, but should not form an integral part of an instrument measuring religiosity. Such an approach leads to a mix of several aspects somehow associated with Islam that fails to measure Muslim religiosity systematically based on a theoretical framework.

V. A final problem associated with the studies mentioned here lies in the *statistical methods that are employed*. In many studies, important statistical measures of reliability and internal validity of the scales are not reported at all (e.g. Hassan 2008). In addition to this general problem of missing statistical information, a further substantial problem can be found in typological measures which mostly employ Cluster Analysis in order to analyze the structure of the included items: Here, not only are measures of reliability often missing, but also the possibilities of assessing the reliability of a typology remain very limited (see also Keskes and Wolf 1993). As a

4 Today, the validity of the secularization paradigm is increasingly challenged in scientific deliberations. "When the debate shifts to 'Islam and secularisation', such critiques, which revisit and re-articulate the paradigm of secularisation from within Sociology of Religion, are hardly taken into account (...)" (Bracke and Fadil 2008). However, Pickel has shown that the validity of the secularization thesis should not be rejected prematurely. According to his results the secularization thesis still explains best the current developments in religiosity in Europe. Only the evolutionary character of the theory has to be abandoned, as Pickel states (2009 and 2010).

consequence we do not know whether these items really measure the same theoretical construct, but also the instruments from different studies cannot be compared.

1.2 Measuring Muslim religiosity: A new approach

In order to obtain an adequate instrument to measure Muslim religiosity, some important structural particularities must be taken into account. One aspect is the absence of a central religious institution, like the church, defining the 'right belief' in Islam. Instead, there are different theological views concerning the definition of the 'right belief or piety'. These different points of views are all more or less accepted within the Muslim community (Rippin 2005, Calder 2007). Therefore, a multidimensional approach is needed in order to cover different facets of Muslim religiosity. In this respect, Glock's five-dimensional model of religiosity is the most established one in the sociology of religion, even though there remains scientific controversy concerning empirical evidence[5] for his model of religiosity (Roof 1979, Huber 2003). There are studies which have employed this model. While they can be seen as the first steps towards a systematic multidimensional approach towards measuring Muslim religiosity, they still have shortcomings comparable to those discussed above (e.g. Hassan 2008 as the most important study in this respect). An adjustment of the indicators used for the specific dimensions is needed as well as a reconsideration of the role of the dimensions according to structural particularities of Islamic piety.

In this paper I use Glock's theoretical model as a heuristic tool in order to measure the different aspects of Muslim religiosity. In particular, a new role is given to the dimension of secular consequences. Whereas in other denominations this dimension is not a genuinely religious one (Stark and Glock 1968, Huber 2003), in the case of Islam it is considered to be an integral part of religiosity and as important and unique as the other four dimensions. The observance of religious norms in everyday life does not simply follow from the other four dimensions of religiosity (belief, ritual practice, experience and knowledge) but instead represents a different level of Muslim piety. This is of great importance if one seeks to capture the heterogeneity of Muslim religiosity beyond the bipolar axis of religiosity vs. non-religiosity. Islam is generally defined as a religion of orthopraxy and there-

5 Huber (2009 and 2003) recently applied Glock's multidimensional approach to the measurement of religiosity in intercultural studies. According to his results the multidimensional structure of religiosity can be confirmed across various religions. This instrument is especially useful when comparing different denominational groups or different countries with varying religious backgrounds. However, in order to capture the variance within one denominational group a more specific measure is needed (Krämer 2009). Therefore, it is important to assess the role of specific Islamic norms for everyday life and not only religious norms in general. In the latter case we would gain only little variance among pious Muslims.

fore differences should be manifested in the degree to which religious norms are observed in everyday life. This is to be understood as a counterpart to Christian orthodoxy (Ruthven 2000).

The paper is organized as follows: First, the indicators employed for the single dimensions will be discussed. Second, results of a survey on religiosity carried out among Muslims living in selected German cities will be presented.

The dimensional structure of Muslim religiosity will be analyzed using principal component analysis (PCA). Finally, the validity of the new instrument will be tested.

2 Data

A precondition for the development of scales is to have a sample of the target population of adequate size and heterogeneity in order to cover different patterns of religiosity. Only in this way can the multidimensional structure of religiosity be unfolded. If only members of a specific religious community with relatively similar religious patterns were surveyed, this would affect the results and lead to a low dimensional solution (see also Huber 2003). This is why it is of high importance to capture the heterogeneity of the target population in the sample as well as possible.

The data used in this study were collected during a university research project in the spring of 2009. Between February and April, 228 Muslims living in selected German cities in North Rhine Westphalia (mainly Dusseldorf, Cologne and Bonn) were surveyed. A self-administered survey design was used. In order to achieve high religious heterogeneity in our sample, we employed a multi-staged sampling procedure as it is common in order to survey rare populations with no existing sampling frame (Kalton 2009). At the first stage different locations were selected where Muslims tend to congregate. The aim was to oversample religious Muslims as it is assumed that the variety of religious patterns is larger among religious Muslims compared to less religious Muslims. We therefore used two sampling strategies: 1. Sampling in Mosques after Friday prayer and at special events in order to achieve an oversample of religious Muslims. 2. Sampling at non-religious locations in order to recruit also secular or non-practicing Muslims.

1. In order to achieve an oversample of religious Muslims, we selected different Mosques and classified them depending on their associational affiliation. As the religious orientation of the different mosque associations differ, we employed a quota sample. The highest concentration of visitors occurs for the Friday prayer, so we distributed questionnaires when the visitors were leaving the Mosques. Depending on the number of visitors, we implemented a random selection procedure: If there were many persons we gave questionnaires to every third, if there were only a few then everyone received a questionnaire. For the reason that mostly male

Table 1: Demographic characteristics of the sample (N=228)

	Migration background		
	Turkish (N=144)	N. African (N=55)	Other (N=29)
Age			
below 30	29.2	36.4	34.5
30 to 44	45.1	34.5	44.8
45 to 59	13.9	21.8	10.3
60 and older	11.1	1.8	0.0
n.a.	0.7	5.5	10.3
Sex			
male	65.3	63.6	75.9
female	34.7	34.5	24.1
n.a.	0.0	1.8	0.0
Education			
still in education	3.5	3.6	0.0
no education/primary school	14.6	1.8	0.0
low (up to 9 years)	18.1	12.7	10.3
medium (10 years)	16.7	12.7	10.3
high (12/13 years)	41.7	56.4	62.1
other	0.7	5.5	13.8
n.a.	4.9	7.3	3.4

Muslims go to the Mosque for the Friday prayer, we also went to special religious events, in which also or only Muslim women take part.

2. In order to recruit also secular or less religious Muslims to cover the other end of the religious pole, we selected different locations. First, we selected districts with a high proportion of Muslim residents. Afterwards we selected different locations in these districts as Turkish or Arabic supermarkets and restaurants in order to distribute the questionnaires. Second, we distributed questionnaires to Muslim students at Dusseldorf University and at consulates of countries with a majority Muslim population.

The obtained sample has the following demographic characteristics: Most of the respondents (63.2 percent) are of Turkish descent, who make up by far the greatest proportion of Germany's Muslim population. Around 24 percent arrived from North African countries like Morocco, Tunisia or Egypt. Another 12.7 percent are of different origin. Almost half of the respondents have German citizenship.

Table 1 shows the frequency distribution of the different ethnic groups in terms of age, education and gender. The sample is biased in terms of education and sex. Highly educated Muslims are overrepresented, as are males.

Table 2: Oversample of very religious Muslims

	own study	ZfT*
very religious	29.2	17.2
rather religious	45.1	50.9
rather not religious	13.9	22.7
not religious at all	11.1	4.5
n.a.	0.7	4.6

basis: N=146, respondents of Turkish origin
*Zentrum für Türkeistudien, 2007 North Rhine-Westphalia *(Center for Turkish Studies)*

In order to check the religious heterogeneity of the obtained sample, we compared the religious self-assessment of the Turkish subsample with the results of a representative study on residents of Turkish origin (see table 2). The comparison shows that "very religious" Muslims are overrepresented in our study as was intended. However, there is also a viable proportion of less religious Muslims with 25% assessing themselves as rather not religious or "not religious at all". Therefore, the precondition of a heterogeneous sample is fulfilled in this study as explained above.

3 Theoretical considerations and the selection of indicators

Charles Glock's (1962) multidimensional model of religiosity serves as a heuristic tool in order to separate different aspects of Muslim religiosity. Glock differentiates between five relatively independent dimensions and claims that these cover all possible forms of religious expressions to be found in all world religions. These are the *ideological dimension*, which he calls 'belief dimension' in his later work. This dimension contains the agreement with basic belief contents of a religion, e.g. the belief in God. The *ritualistic dimension* is divided into the sub-dimensions ritual and devotion. The assumption is that the highly formalized rituals performed mainly in the public do not necessarily coincide with private, informal and spontaneous acts of worship. Furthermore, he distinguishes between an *experience dimension*, a *knowledge dimension* and a *dimension of secular consequences*. The latter was later excluded from the model (Stark and Glock 1968). It is not clear whether the impact a religion has on the everyday life of its adherents is part of a religious commitment or whether it simply follows from such a commitment. This could be true in the case of some denominations but not in that of Islam. As will be

shown below, religious norms regulating the everyday life of Muslims are of great importance in assessing Muslim religiosity. It represents the counter dimension for orthodoxy, which is measured by the belief dimension in the case of Christian religiosity. This important aspect has received inadequate attention in research to date. In the following sections, the indicators employed to measure the single dimensions will be discussed briefly.

Belief

The basis of religiosity is the agreement with the central contents of belief of a specific religion (Glock 1969). The main contents of religious belief within Islam are, on the one hand, the unquestioned belief in the existence of Allah and, on the other, the belief in the Quran as the pristine words of Allah (Ruthven 2000). Additionally, the respondents were asked to what extent they believe in the existence of Jinn, angels and other creatures found in the Quran.

Ritual

Following Waardenburg (2002), the central religious rituals as described by the five pillars of Islam belong to the primary signs of Islam that are accepted by Muslims worldwide even when they are not performed. The five pillars contain more than religious rituals. Additionally, they include the statement of belief and the religious donation (*zakat*), which are related to other dimensions according to Glock. In the course of the empirical analysis we will see whether these aspects belong together or not. In order to measure the ritualistic dimension, I used the frequency of performing the ritual prayer (*salat*), the pilgrimage to Mecca, fasting during the holy month of Ramadan, and celebrating the end of the fasting during Ramadan (*eid sagir*)[6].

Devotion

As indicators to measure the practice of religious devotion, I used the frequency of praying personally to Allah (*dua*) and the frequency of reciting the *basmala*. Every prayer opens with this formula and pious Muslims generally recite it before carrying out important tasks in everyday life. In this way, the believer places his action under Allah's protection and requests that they be successful (Khoury et al 1991). These are acts of worship outside formalized and social rituals. The believer carries them out in privacy and spontaneously.

---

6 Mosque attendance is not included into the dimensional measure because it is no integral part of the central rituals but a suggestion (Rippin 2005). Furthermore, mosque attendance is associated with sex and therefore is not an appropriate measure.

Experience

Glock (1969) assumes that a religious person will one day experience a religious emotion. As Stark (1965) emphasizes, the aspect of a perceived communication with a supernatural agency is a characteristic of religious experience. Particularly in *popular Islam* (Waardenburg 2002), communication with the divine is very common. As Waardenburg points out: "(...) not knowledge but participatory experience (…)" (Waardenburg 2002: 67) is of major concern in popular Islam. Extraordinary things or happenings are perceived as signs from *beyond* (ibid.). Bad or good incidences are often ascribed to Allah, who is believed to reward or punish human behaviour in this world. This corresponds with the subtype of responsive religious experience where "(...) the divine actor is perceived as noting the presence of the human actor" (Stark 1965). Followers of more orthodox traditions in Islam – especially younger generations who show a more rational approach to Islam - do not believe that Allah punishes in this world but rather in the next (Waardenburg 2002, Mihciyazgan 1994). Therefore, this dimension not only measures the degree of religiosity but is also able to differentiate between different types of religious orientations.

In order to measure the experiential dimension, I included the two subtypes of "confirming" and "responsive" religious experience as emphasized by Stark and Glock (1968). The confirming experience, characterized as a sense of the presence of the divine actor, was measured by the following item: "Do you feel the presence of Allah?" In order to measure responsive religious experience, the respondents were asked the following questions: "Have you ever felt that Allah communicates with you?", "Have you ever felt a sense of being rewarded by Allah?" and "Have you ever felt a sense of being punished by Allah?".

Knowledge

Some knowledge of religious contents is expected to be held by believers in all religions (Glock 1962). As Glock (1962) emphasizes, it is extremely difficult to decide which religious contents matter in every single denomination. This is even more difficult in the case of Islam. Given the absence of a central religious authority in Islam, the focus can vary. Generally, the contents of the Quran and the *Sunnah*[7] are the main sources of Islamic knowledge and it is expected that believers know a minimum of these contents (Waardenburg 2002). But which knowledge really matters for individual Muslims is not fixed. Therefore, I decided to let the respondents assess their knowledge concerning firstly, the contents of the Quran, secondly, concerning the life and actions of the prophet Mohammad, and thirdly, concerning Islam in general.

---

7 The *Sunnah* contains the sayings and living habits of Mohammad.

Consequences

Religious law has a predominant function in Islam (Schacht 1993). It does not only give guidance to the correct performance of the religious rituals, but also regulates the everyday life of the believers. The observance of those norms is not to be interpreted solely as a consequence of religiosity even when such norms relate to the everyday life of the believers. Their observance is to be understood as religious worship itself, which is a crucial point. For this reason, the dimension of secular consequences should be conceptualized as an integral part of religiosity in Islam. Especially within the group of pious Muslims, key differences should appear concerning specific religious norms. The Sunni Islamic tradition mostly consists of religious norms regulating individual and social life. This is the primary source of discussion among Islamic scholars, often leading to different opinions and norms concerning the same issue. This is shown for instance by the different schools of Islamic jurisprudence. Regarding dietary rules, it can be stated that most of the scholars more or less agree. The prohibition of eating pork is more or less self-evident and the great majority of Muslims have not been eating pig meat even at times when they were not performing other ritual duties. Even if the degree to which observance of the rules prohibiting the consumption of meat which is not *halal* (slaughtered according Islamic norms) and those of drinking alcohol varies from one believer to the other, these examples are accepted in fully as religious norms among the Muslim community. Similarly, the compulsory religious donation (*zakat*) is unquestioned as part of the five pillars of Islam. By contrast, issues concerning the most important aspects of Islamic morality – namely family and gender relations – are continued subjects of debate. A strict interpretation of key religious sources prohibits the interaction between women and men who are not family members. This led to the spatial segregation of the sexes in the public sphere, for instance in Saudi Arabia and in Iran; but this interpretation can also manifest itself in the separate celebration of collective ceremonies like weddings or funerals including in secular countries (Yazbeck Haddad and Lummis, 1987).

A similar issue is the prohibition of the act of touching hands between unrelated men and women, due to the possible sexual overtone of such an act. In extreme cases this leads to an avoidance of hand shaking with the opposite sexes (ibid.).

In recent decades, other topics have been a matter of concern within the Muslim community. There is an ongoing discussion on whether listening to music is *halal* or *haram*. Initially, this discussion began as a critique of Sufism and its practice on the part of Wahhabi scholars. Music plays a major role in most of the Sufi brotherhoods (Schimmel 2003). Otterbeck (2008) differentiates between three main positions regarding this issue: First, that of the moderates, who say that music in itself is not forbidden. This point of view argues that it is the aspects accompanied by music which must be assessed as a matter of *haram* and *halal* (e.g. sexual excitement), not music itself. Second, that of the hard-liners, who strictly refuse

Table 3: Indicators used for the single dimensions of religiosity

Dimension	Code	Item
Belief	B1	Belief in Allah
	B2	Belief in the Quran as the unchanged revelation
	B3	Belief in the existence of Jinn, Angels etc.
Ritual	R1	Frequency of performing the ritual prayer
	R2	Pilgrimage to Mecca
	R3	Fasting during Ramadan
	R4	Celebrating end of Ramadan
Devotion	D1	Frequency of personal prayer to Allah
	D2	Frequency of recitation of the *Basmala*
Experience	E1	Feeling: Allah is close
	E2	Feeling: Allah tells you something
	E3	Feeling: Allah is rewarding you
	E4	Feeling: Allah is punishing you
Knowledge	K1	Knowledge of Islam in general
	K2	Knowledge of the contents of the Quran
	K3	Knowledge of the life and actions of the prophet
Consequences	C1	Drinking alcohol
	C2	Eating *halal* meat
	C3	Avoiding shaking hands with opposite sex
	C4	Sex segregation at marriages and other celebrations
	C5	Muslims should not listen to music
	C6	Religious donation (*zakat*)

music in general. And third, the position of liberals, who are opposed to all forms of censorship (ibid.).

In order to measure the degree of religious norms influencing the daily actions of believers, I used indicators from the different areas mentioned above: 1. Dietary rules: eating of *halal* meat (slaughtered following Islamic rules) and consumption of alcohol, 2. Paying religious donation (*zakat*), 3. Gender issues: segregation of the sexes and avoidance of hand shaking with the opposite gender, 4. Entertainment: opinion on whether a Muslim is allowed to listen to music or not.

Table 3 gives a summary of the indicators employed in order to measure different aspects of Muslim religiosity. The abbreviations of the items, which are ordered according to their adherence to the theoretical dimensions, will be used in the following sections.

4 Results

In the first section of this chapter the dimensional structure of the indicators discussed above will be explored. In order to analyze whether the dimensional structure of Muslim religiosity resembles the suggested model by Glock, an explorative method is used. Principal component analysis[8] with non-orthogonal rotation (oblimin) was performed in order to determine the dimensional structure of the items.[9] The following criteria were used to classify the items: 1. Communalities above .5, 2. Component loadings above .5, 3. Clear relation to one component. Additionally, the reliability of the solution was analyzed employing Cronbach's Alpha. Cronbach's Alpha equal or higher than .8 is an estimate for high reliability. Items will be excluded if a better estimate of Cronbach's Alpha can thus be obtained.

In the second part, the internal validity of these components will be analyzed testing selected assumptions concerning religiosity.

4.1 Dimensions of Muslim religiosity

The solution of an overall principal component analysis supports the multidimensional structure of Muslim religiosity. Five separate dimensions with an Eigenvalue higher than one could be obtained. Table 4 shows the component loadings of the items on the single dimensions.

Most of the items clearly belong to one component, even when some items have component loadings between .3 and .4 on a second dimension. Three items appear problematic. These are, first, the item "celebration of breaking the Fast" (R4), which has component loadings lower than .5. For this reason, it will be excluded from further analysis. One reason is probably that the celebration of breaking of the fast at the end of Ramadan (*eid al-fitr*) has more or less the same status as Christmas or Easter, and participation in it is not an appropriate indicator for religiosity. The second item with unsatisfactory estimates is item C6, measuring the frequency at which the religious donation (*zakat*) is made. The dependency of this donation on one's financial situation possibly makes it a weak indicator. The communality and therefore explained variance is very low for this item. This item will thus also be excluded from further analysis. The third problematic item is item C4, asking for the necessity of gender segregation. It has quite high component loadings (.44 and .57) on two components. Due to theoretical considerations, this item will be associated to component five (see section 3).

---

8 The number of selected dimensions depends on the Kaiser criterion.

9 Additionally, a Categorical Principal Component Analysis (CatPCA) was employed in order to check the stability of the solution. The solution of CatPCA and PCA are almost the same.

Table 4: Five dimensional solution of PCA[1) 2)]

	Component loadings on dimension					
Items	1	2	3	4	5	communalities
B1	0.929					0.783
B2	0.900					0.861
B3	0.733					0.724
D1	0.598					0.608
D2	0.717					0.763
R1		0.674				0.673
R2		0.727				0.571
R3		0.740				0.729
R4	0.438	0.341				0.423
E1	0.659					0.591
E2				0.793		0.690
E3				0.854		0.691
E4				0.824		0.802
K1			0.850			0.761
K2			0.898			0.796
K3			0.829			0.710
C1	-0.687				0.578	
C2		0.500				0.586
C3					0.697	0.591
C4		0.442			0.565	0.681
C5					0.793	0.627
C6		0.594				0.425

1) component loadings above .3 displayed only
2) oblimin rotated solution, pattern matrix

Looking at the dimensional structure of the items, two of the components are the same as Glock suggests in his model. These are the experience dimension (component 4) and the knowledge dimension (component 3). Furthermore, a distinct dimension measuring religious norms could be obtained, as previously assumed. This is the fifth dimension. Here, the items related to gender relations and popular media are placed. Those related to dietary rules are associated to a different component. They are highly associated with religious rituals, as set out by the five pillars of Islam. For this reason, I have termed this dimension (component 2) “central duties”. It contains, more or less, those religious duties upon which the majority of pious Muslims agree. The norms concerning gender relations and music go further

and measure a more orthodox kind of religiosity. Therefore, component five will be called "orthopraxis". The first dimension contains items measuring religious belief as well as devotional practice and confirming religious experience. This component will be called "basic religiosity". In the following sections the obtained dimensions will be described and interpreted in depth. Further analysis to assess the reliability of the single dimensions will also be carried out.

Basic Religiosity

The first dimension contains all items of religious belief and devotional practice. This means that religious belief cannot be observed independently of practice. Belief is followed by a minimum of devotional religious practice like personal prayer beyond formalized rituals. Additionally, it would appear that some kind of religious experience is needed to confirm Islamic belief. Belief is accompanied by a feeling of an omnipresence of Allah, which is supported by the item "feeling the presence of Allah" loading on the same component. This item was related to the dimension of religious experience theoretically; yet clearly it measures a different aspect of Muslim religiosity. Therefore, no pure belief on a cognitive level could be detected. Instead, a form of core religiosity exists, expressing a general religious commitment. However, the performance of general rituals does not necessarily follow from this. It represents religiosity on an individual level. Communal aspects or collective religious rituals are not part of it. This dimension is mostly characterized by the item "there is no doubt that Allah does exist", with the highest component loading. The reliability of this scale is supported by a pretty good estimate of Cronbach's Alpha (0.90). This dimension is termed basic religiosity because it is a precondition for the other dimensions. On the other hand, the other dimensions cannot be deduced from it. For this reason, basic religiosity must be regarded as separate.

Central Religious Duties

The second dimension expresses the observance of central religious duties. It consists more or less of the observance of the "five pillars of Islam" and additional basic norms: the ritual prayer, fasting on Ramadan, the pilgrimage to Mecca and dietary rules. Accordingly, the adherence to formalized rituals goes hand-in-hand with the observance of some basic religious norms. As discussed before, these norms concerning dietary rules gained the status of natural Islamic duties. This can also be observed empirically. In contrast to the first component measuring Muslim piety on an individual level, the second measures piety on a collective or social level. The religious practices characterizing this dimension have in common, that they are performed mainly together with others. They have a communal character. Additionally, it contains only overall accepted Muslim practices and therefore is

separate from an orthodox kind of piety. For this reason, this component is called central duties. With a reliability estimate of .812, this instrument can be considered highly reliable.

Religious Experience

As Glock assumes for all religions, an independent dimension measuring religious experience for the case of Muslim religiosity could be obtained as well. As suggested by Stark (1965), I differentiated between confirming and responsive religious experience. The three items related to responsive religious experience make up their own dimension.[10] Therefore, only experiences including some kind of perceived communication with the Divine, which exceeds a vague feeling of such a presence, can be separated accurately. The statistical estimates confirm the reliability (Cronbach's Alpha = .81) of this scale.

Religious Knowledge

The dimension of religious knowledge also corresponds with Glock's model. All three items measuring the extent of religious knowledge highly correlate with the same dimension. The reliability of the scale is high, with a Cronbach's Alpha estimate of 0.83.

Orthopraxis

The consequential dimension of religiosity has an important and distinct role within Muslim religiosity. The influence of Islam in the everyday life of believers is not only a consequence of the other dimensions, as Glock puts it, but an own act of worship in and of itself. The degree to which Islam structures the everyday life of believers beyond the standardized religious rituals gives insight into different conceptions of piety within the Muslim community. Whereas some rules, such as those governing diet, are highly standardized and belong to the central duties as discussed above, other religious norms should be differentiated, as suggested by the empirical results. Religious norms organizing gender relations make up a distinct dimension that cannot be explained by the other dimensions of Muslim religiosity. These constitute a separate dimension. In contrast to Glock's suggestion to exclude this dimension from the model, its independent status is confirmed. Within the same dimension we find the item of whether a Muslim should listen to music or not. This dimension expresses the degree of orthopraxis which corresponds with orthodoxy in the case of Christian religiosity.

---

10 The item E1 measuring confirming religious experience is related to *basic religiosity* as explained above.

Table 5: The five dimensions of Muslim religiosity

basic religiosity	central duties	experience	knowledge	orthopraxis
religiosity on an individual level differentiates between believing and not believing Muslims contains: - belief - devotion - sense of omnipresence of Allah	Religiosity on a collective level differentiates between practicing and not practicing Muslims contains: - ritual prayer - fasting at Ramadan - pilgrimage to Mecca - observance of dietary rules	responsive religious experience contains: - sense that Allah… - communicates with oneself - punishes behavior - rewards behavior	There is no fixed set of knowledge expected to be known by believers contains self-assessment of knowledge: - Islam in general - contents *Quran* - contents *Sunna*	Counterpart to orthodoxy in Christianity contains observance of strict religious norms: - gender relations - music

The reliability of this dimension is not very high (Cronbach's Alpha of .64), but due to the number of indicators it is satisfactory. Clearly, more appropriate measures for this dimension need to be developed in order to improve its reliability.

As a last step, five PCA's were carried out in order to re-check the mono dimensionality of every dimension and to save the component values of every respondent on these dimensions. All five PCA's provide mono dimensional solutions with high estimates. Component 1 containing the indicators belief, devotion and confirming religious experience, which measure *basic religiosity*, explains 70 percent of variance. The component loadings vary between .76 and .93. Component 2 containing religious rituals and dietary rules measuring *central duties* of Islam explains 63 percent of variance, the component loadings vary between .76 and .85. The third component measuring responsive *religious experience* explains 73 percent of variance, the component loadings vary between .83 and .90. Component 4 measuring *religious knowledge* explains 75 percent of variance with component loadings varying between .84 and .89. The fifth and last component measuring *orthopraxis* explains 59 percent of variance, the component loadings vary between .70 and .82.

Table 5 summarizes the main characteristics of the obtained dimensions of Muslim religiosity.

4.2 Associations between the dimensions of Muslim religiosity

Each of the five dimensions of Muslim religiosity represents a different facet of piety. Therefore, each of them is separate from the others and provides insight to religiosity from a different perspective. However, they only mirror Muslim religiosity as a complete picture when looking at all of them simultaneously. Although conceptually the dimensions might be distinct, it is to be assumed that they are correlated in the real empirical world. Table 6 shows the correlation matrix of the five dimensions of Muslim religiosity. The correlation coefficients show, that some of the dimensions highly overlap on an empirical level. This is the case especially for dimension 1 and 2 which have almost 50% of common variance. Therefore, religiosity on a collective level is highly connected to individual religiosity. Also, we find a rather high correlation between the dimension of ritual duties and orthopraxy with r=.59. Orthopraxy and basic religiosity are less connected to each other with r=41. The dimensions of religious experience and religious knowledge are less correlated with the other measures of religiosity (with correlations ranging from .16 to .46), whereas religious experience is mostly connected to individual piety (basic religiosity), and religious knowledge to collective piety (central duties). This shows that religious experience as an individual experience is based on belief and devotional religious practice to a certain degree. In contrast, the observance of religious duties requires some knowledge about religious contents and the performance of formalized rituals. However, religious knowledge and religious experience are more or less uncorrelated.

The overall structure of the correlation matrix indicates that there are two main approaches of Muslim religiosity: 1. some kind of spiritual religiosity based on individual communication with the divine and religious experience, and 2. a more formalized kind of religiosity based on social rituals. This corresponds to the two main lines of interpretation of Islamic key sources by Muslim scholars (see Rippin 2007). The first is based on a mystic interpretation represented ideal typical for Sufism. The second is based on a legal interpretation of the key sources represented ideal typical for the Islamic jurisprudence.

The interrelations of the five dimensions should be investigated in more detail in further research.

Table 6: Correlation Matrix of the five dimensions of Muslim religiosity

	basic religiosity	central duties	experience	knowledge	orthopraxis
basic religiosity	1	0.70**	0.46**	0.22**	0.41**
central duties		1	0.32**	0.41**	0.59**
experience			1	0.16*	0.23**
knowledge				1	0.29**
orthopraxis					1

** p<0.01; *p<0.05

4.3 The validity of the dimensions

In order to evaluate the internal validity of the obtained dimensions, I test some assumptions about expected associations between religious factors. The assumptions are self-evident and will not be explained in detail for this purpose. Thus, the latent variables obtained by the five PCAs are used. The latent variables have a mean value of 0 with a standard deviation of 1.

Assumption 1: *The higher the value on the single dimensions of religiosity, the higher the self-assessment of religiosity.*

In order to see if there is congruence between perceived and measured religiosity, which should be the case, I compare the mean differences on the five dimensions between those assessing themselves as religious and those who do not. As is shown in table 7, the mean differences for all five dimensions are highly significant on a 1 percent level. The highest associations with religious self-assessment[11] could be obtained for the dimension of *basic religiosity* (Eta=0.69) and *central duties* (Eta=0.61). This means that it is belief and private worship which determines the definition of 'religious' most; this, in turn, supports the label of this dimension as being the base of religiosity.

Another interesting result is that the relation between religious self-assessment and the five dimensions is not ordinal in all cases. Especially in the case of *religious experience* the mean for 'very religious' (-0.09) is lower than for 'rather religious' (0.25). This means that Muslims with a high degree of religious experience tend to assess themselves as less religious compared to those with a lower degree of religious experience. I assume that the self-assessment depends on the peer group the respondents have in mind. If we consider the two approaches to Muslim religiosity

11 Question: "As how religious would you assess yourself? Very religious, rather religious, rather not religious or not religious at all?"

Table 7: Mean differences between the five dimensions of religiosity in association to the religious self-assessment

		basic religiosity	central duties	experience	knowledge	orthopraxis
very religious	mean	0.22	0.63	0.09	0.54	0.59
	N	37	39	36	38	34
	Std	0.84	0.53	0.98	0.95	1.18
rather religious	mean	0.27	0.14	0.25	0.06	0.01
	N	126	136	122	137	116
	Std	0.39	0.82	0.82	0.88	0.93
rather not religious	mean	-0.77	-0.97	-0.56	-0.60	-0.61
	N	33	35	35	36	34
	Std	1.38	1.05	1.20	0.95	0.61
not religious at all	mean	-3.71	-2.41	-1.72	-1.39	-0.80
	N	5	5	6	6	6
	Std	1.10	0.29	0.39	1.53	0.40
	Eta	0.69**	0.61**	0.43**	0.41**	0.39**

** $p<0.01$

mentioned in the former section, those Muslims defining their piety mainly through religious experience (spirituality) might see themselves as less religious compared to those strictly observing the central religious duties.

Assumption 2: *The higher the values on orthopraxis (dimension 5), the more important religious rules are for everyday life.*

The observance of religious norms in everyday life should be accompanied by assessing religious rules as being important for everyday life. This assumption can be confirmed in this study (see table 8). The association between both variables is quite high with Eta=0.55 on a highly significant level ($p<0.001$). Compared with the self-assessment of religiosity, the importance of religious rules[12] is a better predictor for orthopraxis. Nevertheless, the association with *basic religiosity* and *central duties* is still higher. This is an indication that perceived observance of religious rules must not coincide with orthopraxis. Due to different interpretations of

12 Question: "How important are religious rules in your everyday life? Very important, rather important, rather not important, not important at all?"

Table 8: Mean differences between the five dimensions of religiosity in association to the importance of religious rules for the everyday life

		basic religiosity	central duties	experience	knowledge	orthopraxis
very important	mean	0.41	0.62	0.23	0.23	0.59
	N	84	88	80	88	79
	Std	0.17	0.43	0.81	0.94	0.97
rather important	mean	0.22	0.00	0.22	0.00	-0.29
	N	80	84	77	86	73
	Std	0.41	0.79	0.90	0.91	0.84
rather not important	mean	-0.92	-1.15	-0.74	-0.51	-0.75
	N	27	34	32	34	27
	Std	1.32	1.01	1.09	1.17	0.54
not important at all	mean	-3.20	-2.17	-1.75	-0.38	-0.97
	N	9	8	8	8	9
	Std	1.43	0.40	0.37	1.41	0.30
	Eta	0.80**	0.73**	0.50**	0.26**	0.55**

** $p<0.01$

what religious rules may be, self-assessment seems to be a very subjective measure. The comparability of those subjective measures is doubted.

Assumption 3: *The higher the values on the single dimensions of religiosity, the more frequent the mosque attendance at several occasions (for the ritual prayer, to listen to religious talks, to spend spare time).*

Even when mosque attendance is not a direct religious duty for believers, it represents the physical presence of Muslims in a given place. Therefore, it can be declared as a source of Muslim identity (Rippin 2005). Muslim men in particular are expected to attend the Friday noon prayer in a mosque (ibid.). However, more and more women participate in mosque activities in addition to the classical Friday prayer. Such activities include religious talks held by Muslim scholars or by Imams, or the organization of leisure activities, which are open for male and female Muslims, even when there is gender segregation. It can be assumed that the higher the values on the dimensions of Muslim religiosity, the more frequently respondents will participate in these activities in a mosque. With regard to mosque attendance for the purpose of performing the ritual prayer, this is confirmed for all five dimen-

Table 9: Correlations of the five dimensions of religiosity and Mosque attendance (Pearsons R)

Mosque attendance	basic religiosity	central duties	experience	knowledge	orthopraxis
- to pray	0.44**	0.62**	0.18**	0.27**	0.42**
- to listen to religious talk	0.37**	0.58**	0.16*	0.32**	0.46**
- to spend spare time	0.24**	0.44**	n.s	0.25**	0.45**
N	201	215	199	217	190

** = p<0.01
* = p<0.05
n.s.= not significant

sions, whereby the correlation is highest for the dimension of *central duties* and lowest for the dimension of *religious experience* (see table 9). Overall, the correlations are lower for mosque attendance to listen to religious talks and to spend spare time. These even become insignificant for the case of religious experience. An exception is the dimension of *orthopraxis*. In this case, it is the other way around: The associations of *orthopraxis* and the frequency of going to a mosque to listen to a religious talk or to spend spare time are slightly higher than mosque attendance to perform the ritual prayer. This is an indication of stronger links to mosques on the part of Muslims following orthodox norms in general.

5 Discussion

In this paper I applied a multidimensional approach in order to measure different dimensions and thus the diversity of Muslim religiosity. Starting with Glock's model of religiosity, indicators for single dimensions (belief, ritual, devotion, experience, knowledge, and secular consequences) were derived on the basis of scientific literature on Islam. Glock's model served as a heuristic tool to separate different aspects of religiosity. In the course of the analysis, an explorative approach was taken in order to check whether the structural organization of the items follows a different pattern from the one Glock suggests. Due to the particularities of different denominations, such divergence was to be assumed. Indeed, the results show a slightly different organization of the items. Five different dimensions of Muslim religiosity could be obtained. The first dimension consists of items measuring belief and devotional practice as well as the feeling of a divine omnipresence. A sole belief dimension was thus not found in this study. Belief is highly interrelated

with private and non-formalized acts of worship and a sense of the existence of the divine. This dimension is termed *basic religiosity*. It represents a minimum commitment on an individual level and is therefore the basis of Muslim religiosity in general. However, the other dimensions do not necessarily follow from it. Distinct from the first, a second dimension measures the observance of *central religious duties* mainly covered by the five pillars of Islam. Here, we find the performance of the ritual prayer, fasting at Ramadan, the pilgrimage to Mecca, and the observance of some dietary rules, which represent highly formalized religious practices on a collective level. The third dimension follows Glock's dimension of *religious experience* and involves indicators measuring responsive religious experience. In order to measure the extent of religious knowledge, I employed a subjective measure. The respondents assessed their own knowledge concerning Islam in general, the life of the prophet (*sunna*) and the contents of the *Quran*. The results confirm that these indicators make up an own dimension, which is highly reliable. Therefore, the fourth dimension also follows Glock's suggestion of an autonomous dimension of *religious knowledge*. An important result of this study is that the observance of religious norms beyond basic dietary rules constitutes a distinct dimension of its own. Different to religious dietary rules, which are part of central religious duties, the indicators measuring this dimension capture a more orthodox form of religiosity. These are religious norms concerning gender segregation, avoidance of hand shaking and avoidance of listening to music. For strict Muslims the observance of those religious norms concerning everyday life is an act of worship in itself, and does not only follow from the other aspects of religiosity. It is a particularity of Islam that orthodoxy does not manifest itself in the actual contents of a Muslim's beliefs - but in how Islam determines his or her everyday life. Therefore, this dimension can be seen as counterpart for orthodoxy in other religions and will be called *orthopraxis*. Taking this dimension into account can help to discover differences within the Muslim population that otherwise would not be noticed. Since this study is the first investigation of this dimension, future research could focus on improving the reliability of this scale. Here, specific religious norms concerning family and gender relations could be the subject of scrutiny since these are the most important subjects in the discourse of the global Muslim community. Another orthopractical aspect worth exploring is the handling of the general *Bilderverbot* (prohibition of images) in Islam, i.e. the practice of banning certain forms of pictorial representations. Finally, the extent of literal interpretations of those norms can also provide substantial information on how orthodox Muslims are.

Whether the five dimensions of Muslim religiosity appear in the same way in a representative sample or in a different national context should also be investigated in further research. Especially the interrelations of the distinct dimensions should be investigated in more detail. The correlation matrix of the five dimensions gives hints to different approaches of religiosity among Muslims.

The last step of the analysis was to investigate the validity of the dimensions. The results show that the five dimensions are interrelated with other aspects of Muslim religiosity, as previously assumed. Therefore, validity of the dimensions could be confirmed. An interesting result should be highlighted in this context: The religious self-assessment and importance of religious rules – usually used in surveys as measures of Muslim religiosity - are most highly associated with the dimension of *basic religiosity*. While such measures are able to distinguish between believing and non-believing Muslims, they cannot capture variations within the group of believing Muslims. This could also explain why the results of many studies fall short of showing clear relationships between religiosity and other characteristics. This is no surprise considering that basic religiosity contains those aspects shared by the great majority of Muslims. The five dimensional measures presented here can thus contribute to solving the problems that arise due to the great diversity of Muslim religiosity.

7 References

Bracke, S., & Fadil, N. (2008). *Islam and Secular Modernity under Western Eyes: A Genealogy of a Constitutive Relationship* (EUI Working Papers No. RSCAS 2008/05). Italy: European University Institute.

Brettfeld, K., & Wetzels, P. (2007). *Muslime in Deutschland: Integration, Integrationsbarrieren, Religion und Einstellungen zu Demokratie, Rechtsstaat und politisch-religiös motivierter Gewalt. Ergebnisse von Befragungen im Rahmen einer multizentrischen Studie in städtischen Lebensräumen*. Berlin: Bundesministerium des Inneren.

Calder, N. (2007). The Limits of Islamic Orthodoxy. In A. Rippin (Ed.), *Defining Islam: A Reader* (pp. 222-236). London, Oakville: Equinox.

Esmer, Y. (2002). *Is there an Islamic civilization?* Retrieved January 04, 2011, from http://www.worldvaluessurvey.org/wvs/articles/folder_published/publication_520/files/5_Esmer.pdf.

Glock, C. Y. (1962). On the Study of Religious Commitment. *Religious Education*, (Special Issue), 98-110.

Glock, C. Y. (1969). Über die Dimensionen der Religiosität. In J. Matthes (Ed.), *Kirche und Gesellschaft: Einführung in die Religionssoziologie* (pp. 150-168). Reinbek bei Hamburg: Rowohlt.

Hassan, R. (2008). *Inside Muslim Minds: Understanding Contemporary Islamic Consciousness*: Melbourne University Press.

Heitmeyer, W., & Müller, J. Schröder Helmut (1997). *Verlockender Fundamentalismus: Türkische Jugendliche in Deutschland*. Frankfurt/Main: Suhrkamp.

Huber, S. (2003). *Zentralität und Inhalt: Ein neues multidimensionales Messmodell der Religiosität*. Opladen: Leske+Budrich.

Huber, S. (2009). Religion Monitor 2008: Structuring Principles, Operational Constructs, Interpretive Strategies. In Bertelsmann Stiftung (Ed.), *What the World Believes: Analysis and Commentary on the Religion Monitor 2008* (pp. 17-51). Gütersloh.

Jagodzinki, W., & Dobbelaere, K. (1993). Der Wandel kirchlicher Religiosität in Westeuropa. In J. Bergmann, A. Hahn, & T. Luckmann (Eds.), *Kölner Zeitschrift für Soziologie und Sozialpsychologie Sonderheft: Vol. 33*. Religion und Kultur (pp. 68-91). Opladen: Westdt. Verl.

Kalton, G. (2009). Methods for oversampling rare subpopuations in social surveys. Survey Methodology, (35), 125-141.

Kecskes, R. (2000). Soziale und identifikative Assimilation türkischer Jugendlicher. *Berliner Journal für Soziologie*, (10), 61-78.

Kecskes, R., & Wolf, C. (1993). Christliche Religiosität: Konzepte, Indikatoren, Meßinstrumente. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 45*, 270-287.

Khoury, A. Theodor, Hagemann, L., & Heine, P. (1991). Islam-Lexikon: Geschichte - Ideen - Gestalten (Orig.-Ausg., Vol. 4036). *Herder-Spektrum, 4036*. Freiburg i. Br.: Herder.

Krämer, G. (2008). Hohe Religiosität und Vielfalt. Muslimische Aspekte des internationalen Religionsmonitors. In Bertelsmann Stiftung (Ed.), *Religionsmonitor 2008. Muslimische Religiosität in Deutschland. Überblick zu religiösen Einstellungen und Praktiken* (pp. 68-73). Gütersloh.

Ruthven, M. (2000). *Der Islam.: Eine kurze Einführung*. Stuttgart: Reclam.

Meng, F. (2004). *Islam(ist)ische Orientierungen und gesellschaftliche Integration in der zweiten Migrantengeneration*. Bremen: Universitätsbuchhandlung.

Mihciyazgan, U. (1994). Die religiöse Praxis muslimischer Migranten. Ergebnisse einer empirischen Untersuchung in Hamburg. In I. Lohmann & W. Weisse (Eds.), *Dialog zwischen den Kulturen. Erziehungshistorische und religionspädagogische Gesichtspunkte interkultureller Bildung* (pp. 195-206). Münster, New York: Waxmann.

Mogahed, D. (2009). *The Gallup Coexist Index 2009: A Global Study of Interfaith Relations: With an in-depth analysis of Muslim integration in France Germany and the United Kingdom*. Washington D.C. Retrieved December 20, 2010, from http://www.abudhabigallupcenter.com/144842/REPORT-Gallup-Coexist-Index-2009.aspx.

Otterbeck, J. (2008). Battling over the Public Sphere: Islamic reactions to the music of today. *Contemporary Islam*, (2), 211-228.

Pace, E. (1998). The Helmet and the Turban. Secularization in Islam. In R. Laermans (Ed.), *Secularization and Social Integration*. Leuven: Leuven University Press.

Pew Research Center (22.05.2007). *Muslim Americans: Middle Class and Mostly Mainstream*. Retrieved March 16, 2010, from pewresearch.org/assets/pdf/muslim-americans.pdf.

Pickel, G. (2009): Secularization as an European Fate? – Results from the Church and Religion in an Enlarged Europe Project 2006. In G. Pickel & O. Müller (Ed.): *Church and Religion in Europe. Results from Comparative Research* (pp. 89-122). Wiesbaden.

Pickel, Gert (2010): Säkularisierung, Individualisierung oder Marktmodell? Religiosität und ihre Erklärungsfaktoren im europäischen Vergleich. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 62/2010, 219-245.

Pollack, D. & G. Pickel (2007): Religious Individualization or Secularization? Testing Hypotheses of Religious Change – the Case of Eastern and Western Germany. *British Journal of Sociology* 58/4, 2007, 603-632.

Rippin, A. (2005). *Muslims: Their Religious Beliefs and Practices* (3.th ed.). London: Routledge.

Roof, W. C. (1979). Concepts and Indicators of Religious Commitment: A Critical Review. In R. Wuthnow (Ed.), *The Religious Dimension. New Directions in Quantitative Research*. (pp. 17-45). New York et al.: Academic Press.

Schacht, J. (1993). *An Introduction to Islamic Law* (9.th ed.). Oxford: Clarendon Press.

Schimmel, A. (2003). *Die Religion des Islam: Eine Einführung.* Stuttgart: Reclam.

Sen, F., & Sauer, M. (2006). *Islam in Deutschland. Einstellungen der türkischstämmigen Muslime. Religiöse Praxis und organisatorische Vertretung türkischstämmiger Muslime in Deutschland*. Ergebnisse einer bundesweiten Befragung. Essen.

Stark, R. (1965). A taxonomy of religious experience. *Journal for the Scientific Study of Religion*, 5, 97-116.

Stark, R., & Glock, C. Y. (1968). *American piety: The nature of religious commitment.* Berkeley, Los Angeles.

Waardenburg, J. (2002). *Islam: Historical, Social and Political Perspectives*. Berlin: Walter de Gruyter.

Yazbeck Haddad, Y., & Lummis, A. T. (1987). *Islamic Values in the United States: A comparative Study.* New York, Oxford: Oxford University Press.

methods, data, analyses | Vol. 8(1), 2014, pp. 79-118 DOI: 10.12758/mda.2014.004

The Impact of Method Bias on the Cross-Cultural Comparability in Face-to-Face Surveys Among Ethnic Minorities

Joost W. S. Kappelhof
The Netherlands Institute for Social Research/SCP

Abstract

This article investigates the impact of several sources of method bias on the cross-cultural comparison of attitudes towards gender roles and family ties among non-Western minority ethnic groups. In particular, it investigates how interviewer effects, the use of an interviewer with a shared ethnic background, interview language, interviewer gender, gender matching, the presence of others during the interview and differences in socio-demographic sample composition of non-Western minority ethnic groups affect the cross-cultural comparison of attitudes towards gender roles and family ties between these groups.

The data used in this study come from a large scale face-to face survey conducted among the four largest non-Western minority ethnic groups in The Netherlands for which Statistics Netherlands drew a random sample of named individuals from each of the four largest non-Western minority populations living in The Netherlands. Furthermore, methods are introduced to estimate the potential impact of method bias on cross cultural comparisons.

The results show that measurement of both gender roles and family ties constructs are full scalar invariant across the different ethnic groups, but that observed differences in attitudes between ethnic groups especially towards gender roles are influenced by method bias. This in turn leads to biased comparisons between ethnic groups because of differences in the size of the various sources of method bias, the differential impact of the same method bias between ethnic groups and the combination thereof.

Keywords: methods bias, non-Western ethnic minorities, cross-cultural comparative survey research; incomparability of samples, interviewer effects, multi group Mimic, socio-cultural integration

Introduction

In general population surveys, non-Western minorities – or ethnic minorities as they are sometimes referred to – tend to be underrepresented (Feskens, 2009; Groves & Couper, 1998; Schmeets & Van der Bie, 2005). Ethnic minorities are difficult to survey mainly because of cultural differences, language barriers, socio-demographic characteristics, and a high mobility (Feskens et al., 2010; Feskens et al., 2006; Stoop, 2005).

To reduce nonresponse due to language barriers or cultural differences among ethnic minorities, it is often necessary to make use of Tailor-Made Response Enhancing Measures (TMREM). Examples of these TMREM are the use of translated questionnaires, bilingual interviewers, and interviewers with a shared ethnic background (Groeneveld & Weijers-Martens, 2003; Kappelhof, accepted; Kemper, 1998; Martens, 1999).

However, these TMREM may increase the measurement variability of survey estimates. For example, interviewers can systematically affect the way respondents answer survey questions, especially with respect to more sensitive questions (Tourangeau & Yan, 2007). Furthermore, the ethnicity of the interviewer and the language of the interview can systematically affect the way respondents answer survey questions as well (Van't Land, 2000). Needless to say that potential translation errors in case of translated questionnaires are another source of increased measurement variability.

These TMREM can also affect cross-cultural comparability, for example, if there are differences between the ethnic groups in the number or intensity in which these TMREM were used. Comparability issues can also arise in case the TMREM cause systematic differences between ethnic respondents groups in the way they respond to survey questions (i.e., TMREM have a differential impact). A possible reason would be, for instance, differing attitudes between ethnic groups towards what are sensitive topics (Lee, 1993).

Also, factors that are not (intended as) part of the survey design can complicate or bias comparisons between ethnic groups if the level or presence of these factors varies between these ethnic groups or has a differential effect. For instance, culturally specific or different response strategies between ethnic groups, such as acquiescence (Billiet & Davidov, 2008; Cheung & Rensvold, 2000), social desirability (Johnson & Van de Vijver, 2003) or extreme response styles (Morren et al., 2012a; Morren et al., 2011; Morren et al., 2012b), but also other factors such as the presence of others during the interview, interviewer gender or a gender match between a respondent and an interviewer (Veenman, 2002), may generate such

Direct correspondence to
Joost W.S. Kappelhof, The Netherlands Institute for Social Research/SCP, The Hague, P.O. Box 16164, The Netherlands. E-mail: j.kappelhof@scp.nl

effects. Veenman (2002) discusses a range of reasons for which the presence of others during the interview can cause respondents to adjust their answers.

Differences in sample composition of the different groups with respect to important background variables can also complicate the interpretation of observed differences between these groups (Van de Vijver, 2003; van de Vijver & Leung, 1997). This may cause problems, especially if one is interested in attempting to isolate 'true' cultural differences from differences in socio-demographic composition in which the latter may also affect survey estimates of the various ethnic groups. This can be particularly relevant if one tries to assess the effectiveness of a 'one size fits all' policy on ethnic groups that differ substantially from a socio-demographic point of view.

In the present study we investigate how these different factors affect the cross-cultural comparison of two socio-cultural integration constructs – attitudes towards *Gender Roles* and attitudes on *Familiy Ties* – between non-Western ethnic groups living in the Netherlands. Research suggests that questions about sensitive topics may elicit more measurement bias (e.g., social desirability) via interviewer-assisted modes of data collection (Tourangeau & Yan, 2007). Socio-cultural integration issues, such as *Gender Roles* and *Familiy Ties*, among non-Western ethnic groups in the Netherlands are highly relevant for policy makers. However, the questions measuring these sensitive concepts may suffer from a higher degree of social desirability bias, especially when data is collected via face-to-face surveys. The combination of the topics (gender roles, family ties) and the method of data collection (face-to-face) in our data is therefore suitable for the aim of this study.

This article sets out to investigate:

1. how interviewer effects influence the cross-cultural comparison of attitudes on Gender Roles and Familiy Ties between non-Western groups in the Netherlands; more specifically, the following aspects will be studied:
    - 1.1 how the use of an interviewer with a shared ethnic background affects the cross-cultural comparison of attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups in the Netherlands;
    - 1.2 how the language of the interview affects the comparison of attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups in the Netherlands;
    - 1.3 how interviewer gender and gender matching impact the cross-cultural comparison of attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups in the Netherlands;
2. how the presence of others during the interview affects the comparison of attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups in the Netherlands;

3. to what degree the observed differences in attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups can be attributed to differences in socio-demographic composition between non-Western populations in the Netherlands.

The data used in this study come from a large scale face-to-face survey conducted between November 2010 and June 2011. Statistics Netherlands drew a random sample of named individuals from each of the four largest non-Western minority populations living in The Netherlands. The next section of this article provides an overview of the requirements for conducting valid cross-cultural comparisons and the possible sources of bias that can complicate or invalidate these comparisons. This is followed by the description of the data and methods used to answer our research questions and subsequent results, ending with our conclusion and discussion.

1 Sources of bias that can invalidate or complicate cross-cultural comparisons in face-to-face surveys

In recent years, several books describing guidelines and best practices for conducting cross-cultural or cross-national comparative surveys have been published as well as guidelines on how to analyse cross-cultural survey data (see, for example Davidov et al., 2011; Harkness et al., 2010; Stoop et al., 2010). This is understandable, since a multitude of errors and biases can complicate or even invalidate cross-cultural or cross-national comparisons of theoretically based concepts (He & Van de Vijver, 2012; Poortinga & Van de Vijver, 1987; Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 2004).

When it comes to cross-cultural comparisons, a number of equivalence requirements need to be met before meaningful cross-cultural or cross-national comparisons of theoretical concepts can be made. First of all, the intended concept needs to be understood and have meaning in the different countries or cultures. This is commonly referred to as conceptual equivalence (Hui & Triandis, 1985; Johnson, 1998).

Johnson (1998) refers to the other requirements as forms of procedural equivalence. These forms of procedural equivalence have to do with the way the measurement instrument intended to measure the theoretical concept is constructed and they have a hierarchical structure (Vandenberg & Lance, 2000). Three types of

measurement equivalence are commonly distinguished for the measurement model (van de Vijver & Leung, 1997; van de Vijver & Tanzer, 2004).[1]

First of all there is construct equivalence. Johnson (1998, p. 9.) refers to this as follows "A measure can be identified as having this type of equivalence to the degree that it exhibits a consistent theoretically-derived pattern of relationships with other variables across the cultural groups being examined." In a multi group confirmatory factor analysis approach this relates to configural equivalence (Hox, de Leeuw & Brinkhuis, 2010; Vandenberg & Lance, 2000) .

Secondly, for cross-cultural or cross-national comparison there is the requirement of equal metric units of the measurement instrument used to measure the concept. This is commonly referred to as measurement unit equivalence, metric invariance or weak factorial invariance.

Thirdly, to ensure fairness and equity of cross-cultural or cross-national comparison of concepts, measurement instruments are not only required to use the same metric, they are also required to have the same origin. This type of equivalence is also referred to as full scalar invariance, measurement invariance, strict factorial invariance or scalar equivalence (Meredith, 1993; Meredith & Teresi, 2006; Vandenberg & Lance, 2000; Wicherts, 2007).

Bias in cross-cultural or cross-national comparisons

Three sources of bias that can threaten the validity of cross-cultural or cross-national comparisons are commonly distinguished. These are construct bias, item bias and method bias (Kankaras & Moors, 2009; Van de Vijver, 2003; Van de Vijver, 2011; Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 2004). Construct bias occurs when the requirement of construct equivalence is not met. This can happen when non-identical constructs are measured across cultures or countries, or when there is only a partial overlap of the construct between the cultures or countries. Construct bias happens at the level of the measurement instrument designed to capture the theoretical concept.

Item bias happens at the individual question level and occurs when translations of questions (or items) lead to differences in question meaning or ambiguity. Item bias can also be the result of cultural specifics which can be viewed as a form of differential item functioning (DIF) (Mellenbergh, 1989). DIF is a term that stems from education testing and happens when persons of equal capability or intelligence arrive at different capability or intelligence scores based on the specific wording of an item.

---

1 Some distinguish more than three forms of measurement equivalence and make a distinction between strong (no equal residual variances) and strict factorial invariance (equal residual variances).

Method bias happens at survey level and can be introduced by a variety of factors which are distinguished in the following three categories: incomparability of samples, administration bias, and instrument bias. Incomparability of samples refers to differences in the sample composition with respect to important socio-demographic characteristics of the respondents. Administration bias refers to bias that is introduced as a result of differences in how the questionnaire is administered (e.g., interviewer effects, presence of others during the interview, interviewer characteristics), differences in questionnaire design, differences in mode of administration, etc. Instrument bias refers to bias that is introduced as a result of differences in familiarity with being interviewed, but also differences in cultural specific answer strategies.

Research into different sources of method bias

Within cross-cultural or cross-national research, method bias has received relatively little attention in comparison with construct and item bias (Van de Vijver, 2011). As far as method bias is concerned, differential answering strategies, such as acquiescence and other types of response styles, appear to have received the most attention (see for instance, Baumgartner & Steenkamp, 2001; Billiet & Davidov, 2008; Billiet & McClendon, 2000; Chen et al., 1995; Cheung & Rensvold, 2000; He & Van de Vijver, 2013; Hui & Triandis, 1989; Johnson et al., 2005; Marin et al., 1992; Morren et al., 2011; Morren et al., 2012a; Morren et al., 2012b; Ross & Mirowsky, 1984). This is not surprising, since the respondent is always a part of the survey process.

However, many studies concerned with response styles pay relatively little attention to other sources of method bias that can contribute to the observed differences in response styles, despite the fact that these data are often collected via an interviewer-assisted mode of data collection. For example, the SPVA-study – Social-economic Position of Ethnic groups – aimed to measure the socio-economic position and socio-cultural integration conducted among ethnic minorities in the Netherlands. This study was conducted face-to-face and further research on these data has shown the existence of differential response styles (Morren et al., 2012a; Morren et al., 2011). For its data collection through CAPI , the SPVA survey also used translated questionnaires, interviewers with a shared ethnic background, allowed proxy interviews and family member interpreters (Groeneveld and Weijers-Martens, 2003). So, the question is to which degree these differential response styles are the result of characteristics of the respondents themselves and to which degree they are affected by different impacts of interview language, the presence of others during the interview, gender of the interviewer, the ethnicity of the interviewers, proxy interviews and family member interpreters.

Usually, a lack of information on interviewer characteristics and interview setting prevents a more detailed analysis of these types of method bias in cross-cultural research. However, this does not mean that these factors do not bias estimates and, as a result, also lead to biased comparisons. There has been extensive research on the existence of interviewer effects and it has been shown that respondents' answers can be affected by interviewer gender, interviewer race and/or differences (or similarities) between interviewer and respondent such as gender match and race (Anderson et al., 1988; Davis, 1997; Davis et al., 2010; Finkel et al., 1991; Rhodes, 1994; Schuman & Converse, 1971; Williams Jr, 1964; Veenman, 2002; van der Zouwen, 2006). Especially the match between the race of the interviewer and that of the respondent plays a role in the answers given on culturally sensitive questions (Campbell, 1981; Cotter et al., 1982; Sudman & Bradburn, 1974; Schuman & Converse, 1971; Van Heelsum, 1997; Van't Land, 2000).Furthermore, a meta-analysis on sensitive questions in surveys by Tourangeau & Yan (2007) shows that respondents not only adjust their responses to sensitive questions in the presence of interviewers but also in the presence of others, such as family members.

The incomparability of samples can also bias cross-cultural comparisons (He & Van de Vijver, 2012; Kankaras & Moors, 2009). Several studies have analyzed the impact of different socio-demographic sample composition of the compared cultural groups on the observed cross-cultural differences (Arends-Tóth & Van de Vijver, 2008; Fernandez & Marcopulos, 2008; Leung et al., 1998). Several procedures on how to deal with the incomparability of samples, also known as observed heterogeneity, have been proposed (Boehnke et al., 2011; Lubke et al., 2003; Lubke & Muthen, 2005) as well as other procedures to separate compositional differences from 'true' group differences (DiNardo et al., 1996; Huang et al., 2005; Oaxaca, 1973).

2 Data & Methods

2.1 Data

The data used in this article come from the Dutch Survey on the Integration of Minorities (SIM) that sets out to measure the socio-economic position of non-Western minorities as well as their socio-cultural integration. It is a nationwide, cross-sectional, face-to-face CAPI survey; and the fieldwork was conducted by GfK Netherlands between October 2010 and June 2011 among the four largest non-Western minority groups living in the Netherlands plus a Dutch reference group. For this face-to-face survey, Statistics Netherlands drew five samples of named individuals: one random sample was drawn from each of five mutually exclusive population

Table 1: Response rate (AAPOR definition 1), response sample size and gross sample of SIM2011 face-to-face survey, separately for each ethnic group

Ethnic Group	Response rate (%)	Response sample	Gross sample
Turkish	52.1	815	1565
Moroccan	48.0	829	1740
Surinamese	41.0	780	1930
Antillean (incl. Aruban).	44.2	863	1974

strata; Dutch of Turkish, Moroccan, Surinamese, and Antillean[2] descent and the remainder of the population (mostly native Dutch) living in the Netherlands, in the age of 15 years and above. The present study focuses on how response enhancing measures, interview setting, interviewer characteristics and the incomparability of samples in face-to-face surveys can affect cross-cultural comparisons between non-Western ethnic minority groups. This is why the samples containing native Dutch are excluded from this study, the analysis being therefore based on four samples.

The official definition, as is used in statistical research in the Netherlands, of Dutch of Turkish, Moroccan, Surinamese, and Antillean descent includes persons that were either born in Turkey, Morocco, Surinam or the Dutch Antilles[3] or have at least one parent who was born there. In case the father and mother were born in different countries, the mother's country of birth is dominant, unless the mother was born in the Netherlands, in which case the father's country of birth is dominant. The four ethnic groups in this study make up about two-thirds of the total non-Western population, which amounts to approximately 7% of the total population in the Netherlands (CBS-statline, 2014). For the purpose of brevity, they will be referred to as Turkish, Moroccans, Surinamese and Antilleans in the remainder of this article.

The response rate (AAPOR definition 1, (AAPOR, 2011) of the SIM2011 face-to-face survey varied between the four ethnic groups and is shown in Table 1. Table 1 also includes, the gross sample and the sample size of each of the four response samples (i.e., the sample of the respondents).

In this article the SIM2011 response data file will be used. The response data file contains respondents' answers to survey questions, but also socio-demographic information on the respondent, socio-demographic information on the interviewer and interviewer observations (Table 2). Six survey questions measuring socio-cultural integration will be used in this analysis. These questions or a slightly larger

2 Including Aruba
3 or Aruba

set of questions have been used to measure socio-cultural integration of non-Western ethnic minorities in the Netherlands for over a decade (Arends-Tóth & Van de Vijver, 2008; Dagevos & Gijsberts, 2009; Dagevos & Schellingerhout, 2003; Dagevos et al., 2007). The first set of three questions aims to measure Gender role attitudes and the second set of three questions aims to measure Familiy Ties. The interviewer observation data are the result of a short form that an interviewer had to complete after each interview. In this form they had to record in which language the interview was conducted, how well they believed the respondent was able to understand and speak Dutch, but also if there were others present during the interview and if they had, according to the interviewer, influenced the answers of the respondents.

Hypotheses with respect to the research questions

Interviewer effects

Interviewer dependent correlation between the answers of respondents is not often modeled in cross-cultural or cross-national studies, but it has the potential to affect the cross-cultural comparison when the data is collected face-to-face.

Hypothesis: Observed differences between ethnic groups with respect to *Gender Roles* and *Familiy Ties* can be partly explained by interviewer effects.

The effect of bilingual interviewers with a shared ethnic background

Interviewers may have an effect on the responses and especially, the use of bilingual interviewers with a shared ethnic background can impact survey outcomes in several ways. First of all, they can have an effect with respect to potential non-response bias. They can interview respondents that would not have participated due to language difficulties in combination with functional illiteracy or cultural etiquettes. Nonresponse bias on survey outcomes would occur if these potential respondents would have a different opinion on those survey topics and they were not able to participate.

Secondly, they can have an effect with respect to potential measurement bias. Here we can distinguish two effects: the interview language and shared ethnic background. Both have the potential to increase measurement bias. For instance, the question delivery or wording of a translated questionnaire can cause a systematic difference which is, of course, intertwined with the translated questionnaire. Also, their shared ethnic background may elicit more responses that are viewed as socially desirable within the ethnic group.

The use of bilingual interviewers with a shared ethnic background in SIM2011 does not allow for this level of disentanglement of bias. For instance, *all* respondents of Moroccan or Turkish origin were interviewed by a bilingual interviewer with a

Table 2: SIM2011 data used in the analysis

Questions on socio-cultural integration

- [MANGELD] It is best if the man is responsible for the finances. (Ranging from 1= completely agree to 5=completely disagree).
- [INKJONGS] It is more important for boys than girls to earn their own money. (Ranging from 1= completely agree to 5=completely disagree).
- [VRWSTOPW] A woman should stop working when she has child. (Ranging from 1= completely agree to 5=completely disagree).
- [THUISHUW] It is best for children to live at home until they get married. (Ranging from 1= completely agree to 5=completely disagree).
- [VERTRFAMA] I trust my family more than my friends. (Ranging from 1= completely agree to 5=completely disagree).
- [KIBEZOUD] Children that live close to their parents' home should visit them at least once a week. (Ranging from 1= completely agree to 5=completely disagree).

Socio-demographic information on the respondent

- Ethnicity (Turkish, Moroccan, Surinamese and Antillean)
- Gender
- Age Group (15-24; 25-34; 35-44; 45-54; 55-64; 64+)
- Immigration generation (first generation immigrant; second generation immigrant)
- Education level (max. primary school; lower secondary; upper secondary; tertiary or more)
- Municipality size (over 250000; between 250000 and 50000; less than 50000)
- Employment status (employed, not employed, not part of the labour force)
- Has a Children (yes; no)
- Has a Partner (yes; no)
- Weight variable (design weight plus nonresponse adjustment)

Socio-demographic information on the interviewer

- Unique id number
- Ethnicity of the interviewer (Turkish, Moroccan, Surinamese, Antillean, Dutch)
- Gender of the interviewer

Interviewer observations

- Others present during the interview (no; yes, but no influence; yes, influence)
- In which language was the interview conducted (Dutch; mostly Dutch; half Dutch/ half native language; mostly native language; native language)
- What was the respondent's Dutch language proficiency level (good; fair, poor, bad)

Note. Original questions were in Dutch and these are translated by the author.

shared ethnic background. This was a necessary step not only because greater cultural familiarity due to a shared ethnic background increases the willingness to respond, but mostly because language difficulties are still quite common among the Turkish and Moroccans. This would allow the respondent to answer either in Dutch or in their native tongue.

About half of the interviews among respondents of Surinamese or Antillean origin were conducted by interviewers with a shared ethnic background, because Dutch is the mother tongue for many, if not all persons of Surinamese or Antillean origin in the Netherlands.

The SIM2011 face-to-face survey data do allow for the estimation of how the use of (bilingual) interviewers with a shared ethnic background affected the cross-cultural comparison with respect to potential nonresponse bias. In the SIM2011 data information was available on the language in which the interview was conducted, the level of the Dutch language skill and the ethnicity of the interviewer (Table 2). Here it was assumed that respondents would not have participated because of language problems or cultural differences if the interview was conducted mostly in their native language and the interviewer also assessed that the respondent's Dutch language proficiency level was poor. A comparison between the model excluding and the one including these respondents will show the impact of the increased nonresponse on the cross-cultural comparison.

Hypothesis: The use of bilingual interviewers with a shared ethnic background will have a systematic effect on the cross-cultural comparison. In particular, it will result in more traditional views with respect to *Gender Roles* and *Familiy Ties*. First of all, with respect to nonresponse bias we expect respondents who otherwise would not to participate due to language problems or cultural specific reasons to hold more traditional views towards *Gender Roles* and *Familiy Ties*. Secondly, we expect that the shared ethnic background elicits more traditional views toward *Gender Roles* and *Familiy Ties* because these are felt as more socially desirable within the ethnic group.

The effect of interview language

The SIM2011 data also allows for an estimate of the effect of interview language on the cross-cultural comparison. In this instance, the data about interview language was used to create a dummy indicating whether the interview was conducted (almost) completely in Dutch or not. Not only among Turkish and Moroccans, but also among the Surinamese and Antilleans, some of the interviews were at least partly conducted in another language as well. Obviously, the interview language will be part measurement and part nonresponse related. Furthermore, the effect of the ethnicity of the interviewer will be confounded with the interview language and also potential systematic differences introduced by a translated questionnaire

can contribute although that effect should be isolated (i.e., indicator and language dependent).

Hypothesis: Interview language has a systematic effect on the measurement of *Gender Roles* and *Familiy Ties*. If the interview language is Dutch, this will lead to less traditional views towards *Gender Roles* and *Familiy Ties*.

Interviewer gender and gender match

In the SIM2011 data, information on the interviewer gender as well as the gender of the respondent was available (Table 2). This allowed for the construction of both an interviewer gender and a *matched/unmatched* indicator to test how interviewer gender and gender match affect the cross-cultural comparison of socio-cultural issues. However, given the topics (*gender roles* and *family ties*) and the traditional views of some of these ethnic groups, we might expect men and women to react differently in the presence of a gender (un)match. For instance, women may give less traditional answers in the presence of a female interviewer whereas men may become more traditional in the presence of a male interviewer. This interaction may be masked if only a *match/unmatched* indicator is fitted. To test this hypothesis an interaction term (gender respondent with gender interviewer) was created in order to find out if there was an effect of interviewer gender and/or differential effect of gender match between men and women.

Hypothesis: Interviewer gender and gender matching will effect the cross-cultural comparability. In particular, we expect that interviews conducted by a male interviewer will result in more traditional views towards *Gender Roles* and *Familiy Ties* from the respondents compared to interviews conducted by a female interviewer, especially in the case of male respondents.

The presence (and potential influence) of others

In the SIM2011 data information on the presence of others was available (Table 2). This allowed for the construction of a *presence (*dummy*)* indicator to test how the presence of others affects the cross-cultural comparison of *Gender Roles* and *Familiy Ties*. A score of '1' (presence) was assigned to the dummy indicator if the interviewer assessed that a third party present during the interview exerted a direct or indirect influence on the way the respondent answered the questions. In all other instances (i.e., no one present or someone present but no noticeable influence) a score of '0' was assigned to the dummy.

Hypothesis: The presence of others during an interview will systematically affect the results concerning *Gender Roles* and *Familiy Ties*.

Incomparability of samples

With respect to the last research question – the incomparability of samples – we expect that part of the observed differences between the ethnic groups can be explained by differences in socio-demographic composition.

2.2 Methods

A variety of different modeling and analysis techniques have been used to detect equivalence of measures in cross-cultural research. See Braun & Johnson (2010) for an extensive overview.

In the present study multi group confirmatory factor analysis is used (MGCFA) (Joreskog, 1971) to test if the base model – full scalar invariance of the two-factor model of socio-cultural integration among the four non-Western minority groups in the Netherlands – adequately describes the data. The latent variable *Gender Roles* is measured by the following three items: MANGELD; INKJONGS and VRW-STOPW (Table 2). The latent variable *Family Ties* is measured by THUISHUW, VERTRFAMA and KIBEZOUD (Table 2).

The full scalar model is used as the basic model (Model 0) and this article does not focus on the question whether a less restrictive model (e.g., configural equivalence, metric invariance or partially measurement invariant) describes the data better, but rather focusses on the question how method bias can bias the full scalar model with respect to cross-cultural comparisons of socio-cultural integration among non-Western minorities in the Netherlands.

The MGCFA analyses have been conducted with Mplus version 6.11 (Muthén & Muthén, 2011). Both factors have ordered categorical indicators and therefore the WLSMV (Mean- and Variance-adjusted Weighted Least Square) estimator will be used to address the multivariate normality assumption (Lubke & Muthén, 2004).

In addition, several, non-nested models, corresponding to the research questions are going to be analyzed and compared, which normally leads to the use of AIC or BIC fit indices to compare the models (Kuha, 2004). However, the combination of WLSMV and the modeling of interviewer effects through clustering does not allow for models to be compared using these indices.[4] Therefore the fit of every model will be judged separately using three often used fit indices: the root mean square error of approximation (RMSEA) (Steiger, 1989), the Tucker-Lewis index (TLI) (Tucker & Lewis, 1973) and the comparative fit index (CFI) (Bentler, 1990).

4 Using a maximum likelihood estimator to compare non-nested models based on categorical data would allow the use of BIC. Mplus allows for this approach where instead of a MGCFA, a latent class approach is used with knownclass and type=mixture instead of the grouping variable. However, this does not allow for the modeling of interviewer effects using unique interviewer id as a cluster variable, because that requires type =complex.

The root mean square error of approximation (RMSEA) is an absolute fit index that examines closeness of fit. A RMSEA value of more than 0.1 is seen as an indication of poor fit, a value of 0.05 to 0.08 as acceptable and a value below 0.05 as good to very good (Hu & Bentler, 1999), although the absoluteness of these cut-off values has been criticized more than once (see for example Chen et al., 2008). The comparative indices "Tucker-Lewis index (TLI)" and "comparative fit index (CFI)" compare the fit of the model under consideration with fit of baseline-model. Fit is considered adequate if the CFI and TLI values are above 0.90, better if they are above 0.95.

Interviewer effects.

This model involves the inclusion of an unique interviewer ID as a cluster variable in the MGCFA test of full scalar equivalence (Model 1). This allows for a correction of possible interviewer-dependent correlation between the answers of respondents that were interviewed by the same interviewer. A comparison between model 0 and model 1 would give an indication as to how possible interviewer effects influence the cross-cultural comparisons of socio-cultural integration (i.e., gender roles and family ties) among non-Western minorities in the Netherlands. For the remainder of the analysis, model 1 is chosen to be the reference model, since it more accurately describes the data structure. The interviewer effects will also be included in the remaining models.

Bilingual interviewers with a shared ethnic background: nonresponse

In this instance model 1 will be used, but it will be fitted on a selection of the respondents (Model 2). The respondents that participated in their native language *and* for whom the interviewer assessed that their Dutch language proficiency level was poor were excluded. A comparison between the Model 1 and Model 2 (excluding respondents due to language problems) will show the impact of the increased nonresponse due to language problems on the cross-cultural comparison.

Interview language; the presence of others; interviewer gender and gender match.

Interview language, the presence of others, interviewer gender and gender match are sources of method bias that are not randomly assigned across experimental conditions, but are confounded with respondent's characteristics. In order to assess if and how these sources of method bias systematically influenced the cross-cultural comparison of *Gender Roles* and *Familiy Ties,* a multiple group MIMIC model (Multiple Indicators Multiple Causes) was used, in which the impact of these sources of method bias, together with eight other socio-demographic variables on the respondent, were regressed on the latent variables and indicators (see Table 2:

Socio-demographic information on the respondent). This will be referred to as Model 3 (M3) and if there is no systematic bias introduced by these sources of method bias they should not have a significant impact on the latent variables. Furthermore, a comparison between Model 1 en Model 3 will show the impact of these combined types of method bias on the cross-cultural comparison.

The incomparability of samples

The four non-Western groups in this study differ in socio-demographic composition (CBS-statline, 2014). A propensity score weighting method is used to investigate how the incomparability of the socio-demographic composition of samples (IoS) between ethnic groups affects cross-cultural comparisons (Bia & Mattei, 2008; DiNardo et al., 1996; Huang et al., 2005; Imbens, 2000; Rosenbaum & Rubin, 1983).

The selection of important socio-demographic variables for the propensity score reweighting was done in three steps. As a first step, ordered logistic regression was used to ascertain which of the eight socio-demographic background variables have a significant effect on the different categorical indicators (see Table 2: Socio-demographic information on the respondent). As a second step, a check for significant differences in the composition of the four ethnic groups with respect to these socio-demographic background variables was conducted. As a third step, only those socio-demographic background variables were selected to be included in the propensity score weighting model for which it was shown that they a) have a significant impact on at least one of the categorical indicators and b) show a significant difference between at least two ethnic groups. This led to the propensity score reweighting of the different ethnic groups with respect to four socio-demographic background variables: "Municipality size", "Employment status", "Education level" and "Immigration generation". The comparison of the model with propensity weighted samples (Model 4) with Model 1 would allow for an estimation of the effect of IoS on the observed cultural differences.[5]

5 As a check on the usability of the propensity score weighting method to disentangle 'true' cultural differences from IoS on the cross-cultural comparison of socio-cultural integration, the Oaxaca-Blinder decomposition (OBD) method was also used (Blinder, 1973; DiNardo, 2006; Jann, 2008; Oaxaca, 1973). This should yield similar results (DiNardo, 2006).

3 Results

Model 0: Full scalar invariance

The results of the three fit indices show that full scalar equivalence (M0) has an acceptable fit. This means that both factor means can be compared between the different ethnic groups in a fair and equitable way (Table 3).[6]

The factor means of *Gender Roles* and *Familiy Ties* of the different ethnic groups are shown in Figures 1 and 2 under M0. Figures 1 and 2 show the change in relative positions of the factor means of *Gender Roles* and respectively *Familiy Ties* among the ethnic groups after correcting for the various sources of method bias. For details on the numerical values of the parameter estimates and their respective standard errors, see Appendix A. It can be seen that Turkish and Moroccans have, one average, a similar, more traditional attitude towards *Gender Roles* and *Familiy Ties* in comparison to the Surinamese and Antilleans, although there is a significant difference in factor mean for *Family Ties* between Turkish and Moroccans (Tables 4 and 5). There are no significant differences between Turkish and Moroccans for *Gender Roles* as well as no significant differences between Surinamese and Antilleans for both *Gender Roles* and *Family Ties* (Tables 4 and 5). The remaining group comparisons all show significant differences between ethnic groups for both factor means.[7]

Model 1:
The impact of interviewer effects on the cross-cultural comparison

In model 1 (M1), interviewer effects are taken into account when testing for full scalar invariance. The inclusion of interviewer effects where interviewers are modelled as a clustering of observations by unique interviewer number resembles more closely the actual structure of the sample and has a good fit according to the fit indices (Table 3). As could be expected, the correction for interviewer effects mainly results in larger standard errors around factor loadings and thresholds for the indicators of both means (See Appendix A). The relative positions of both *Gender Roles* and *Family Ties* of the ethnic groups are only slightly affected, but this does not change the ordering (Figures 1 and 2). However, there is no significant difference for *Gender Role* anymore between Moroccans and Antilleans (compare M0 and M1 in Table 4). This means that the observed difference between Moroccans and Antilleans in Model 0 is the result of interviewer effects.

---

6 Response samples are weighted to the respective population distribution for gender, household size, municipality size, immigration generation, age groups (12).

7 Based on t-test comparison of means for independent groups using a Bonferroni adjusted significant level for multiple comparisons.

Table 3: Fit indices results for each model

Model	RMSEA	$CI^{0.95}_{rmsea}$	CFI	TLI
M0	0.079	0.072 - 0.085	0.940	0.961
M1	0.053	0.047 - 0.060	0.936	0.958
M2	0.055	0.047 - 0.062	0.935	0.958
M3	0.021	0.016 - 0.026	0.938	0.921
M4	0.049	0.043 - 0.056	0.952	0.969

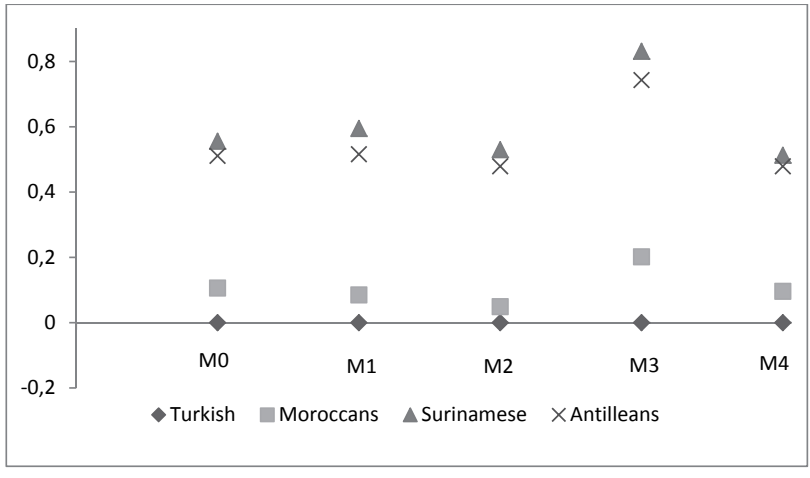


Figure 1: Relative positions on Gender Roles of the ethnic groups

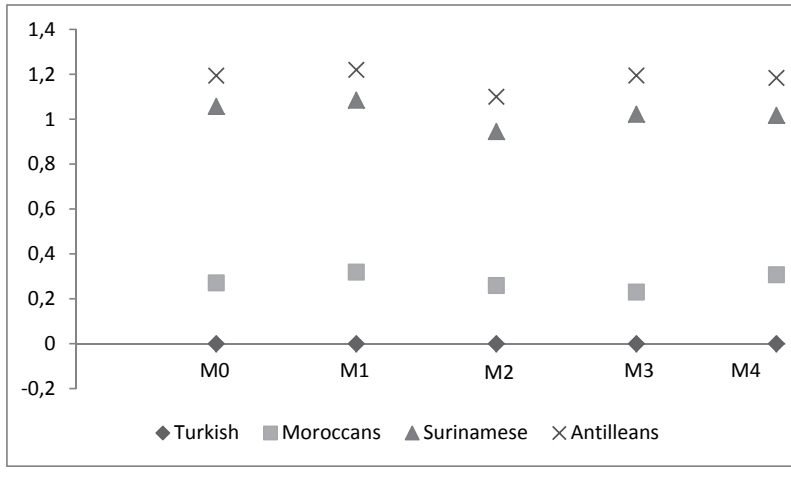


Figure 2: Relative positions on Family Ties of the ethnic groups

Model 2:
The impact of a (bilingual) interviewer with a shared ethnic background on the cross-cultural comparison in terms of nonresponse bias

The comparison of Model 2 (M2) with Model 1 (M1) shows the impact of a (bilingual) interviewer with a shared ethnic background on the cross-cultural comparison in terms of nonresponse bias. Model 2 also has a good fit according to the fit criteria (Table 3).

Compared to Model 1, the ethnic groups would have more similar attitudes if no provisions were made to accommodate for persons who do not speak Dutch or have a cultural specific etiquette when it comes to being asked to participate in an interview (see Figures 1 and 2). For attitudes towards *Gender Roles* only a significant difference between Turkish and Antilleans would remain and for *Family Ties* the observed difference between Turkish and Moroccans would no longer be significant (Tables 4 and 5).

Since the Tailor-Made Response Enhancing Measures (TMREM) mostly affected the Turkish and Moroccans, it can be said that the exclusion of potential respondents due to language problems and lack of cultural etiquette leads to less traditional attitudes of Turkish and Moroccans.

Model 3:
The effect of interview language, interviewer gender and gender match interaction, the presence of others on the cross-cultural comparison

Table 6 presents the results of the analysis with respect to the impact of *interview language*, *interviewer gender*, *gender match interaction* and *the presence of others* on attitudes towards *Gender Roles* and *Familiy Ties*. The complete results can be seen in appendix B. Model 3 (M3) shows an acceptable fit (Table 3).

The analysis results show that being interviewed in your native language by a bilingual interviewers with a shared ethnic background significantly affects the attitudes Turkish, Moroccan and Antillean respondents have towards *Familiy Ties*. In all cases more traditional views with respect to *Familiy Ties* are reported. Among the Surinamese there is no significant effect for interview language. This is mostly due to the fact that there are only very few Surinamese interviews conducted in another language.

The *Interviewer gender* only has an effect among Moroccans and only on attitudes towards *Gender Roles*. In this instance, Moroccan respondents report less traditional attitudes when the interview is conducted by a female interviewer.

There is an interaction effect for *Gender match* on attitudes towards *Gender Roles* among Turkish respondents. Turkish male respondents report more traditional attitudes when the interview is conducted by a male interviewer, while there is no significant effect in the case of Turkish female respondents.

Table 4: Overview of significant differences between ethnic groups for Gender Roles, separately for each model.

Gender Roles	M0	M1	M2	M3	M4
T vs. M					
T vs. S	*	*		*	
T vs. A	*	*	*	*	
M vs. S	*	*		*	
M vs. A	*			*	
S vs. A					

Note. * = Bonferroni corrected significance level (0.05/n of tests). T = Turkish, M = Moroccans, S = Surinamese and A = Antilleans

Table 5: Overview of significant differences between ethnic groups for Family Ties, separately for each model.

Family Ties	M0	M1	M2	M3	M4
T vs. M	*	*			
T vs. S	*	*	*	*	*
T vs. A	*	*	*	*	*
M vs. S	*	*	*	*	*
M vs. A	*	*	*	*	*
S vs. A					

Note. * = Bonferroni corrected significance level (0.05/n of tests). T = Turkish, M = Moroccans, S = Surinamese and A = Antilleans

Table 6: The impact of interview language, interviewer gender, gender match and the presence of others on Gender Roles (GR) and Family Ties (FT), separately for each ethnic group.

	Turkish		Moroccans		Surinamese		Antilleans	
	GR	FT	GR	FT	GR	FT	GR	FT
Interview language		*		*				*
Interviewer gender			*					
Gender match	*							
Others present					*	*		*

Note. * p = <0.05.

The *presence of others* during the interview significantly affects the attitudes of Surinamese for both *Gender Roles* and *Familiy Ties*, as well as Antilleans' attitudes towards *Family Ties*. In all instances the presence of others led to more traditional opinions. Interestingly enough this effect is not (significantly) present among Turkish and Moroccans. The number of interviews in which the interviewer found the presence of others to have a biasing effect varied between 5.6 percent of all interviews conducted among Antilleans and 7.2 percent of all interviews conducted among Surinamese (Turkish 5.8 % and Moroccans 6.4%).

With the exception of attitudes towards *Familiy Ties* among Antilleans, there is at least one significant source of method bias present that systematically affects the attitudes reported by the respondents. Furthermore, there is no source of method bias that has a consistent impact across ethnic groups for one or both latent constructs. As a result, the cross-cultural comparison of these attitudes is biased when comparing the ethnic groups. The actual size of the bias with respect to the cross-cultural comparison of latent means between ethnic groups depends on both the size of the effect and the number of respondents showing this effect.

Model 3 (M3) in Figures 1 and 2 shows the (estimated) relative positions of the latent means for each ethnic group in case adjustments are made for the impact of these sources of method bias. In this case, eight socio-demographic characteristics were also included as covariates to take into account the nonrandom allocation of these source of method bias. Model 3 (M3) in Tables 4 and 5 show how the adjustments impact the ethnic group comparison. In this instance, the adjustments resulted in the same significant differences as Model 0 (M0) with the exception of the significant difference between Turkish and Moroccans for *Family Ties*.

Model 4: The impact of the incomparability of samples on the cross-cultural comparison

A propensity score weighting method has been used to assess the impact of differences in socio-demographic sample composition between ethnic groups. A summary of the significant differences between the ethnic groups for eight socio-demographic variables is given in Table 7 (see Table 2 for a description of the socio-demographic variables included in this comparison and Appendix C for the actual results). For modeling reasons, the original variables – *municipality size* and *employment status* – have been condensed to dummies – Big city dweller (y/n) and Employed (y/n). 21 significant differences are observed between the ethnic groups if they are weighted to their respective population distributions.[8] Using the propensity weighting procedure described in section 2.2, only seven of these significant

8 Weighted to the respective population distribution for gender, household size, municipality size, immigration generation, age groups (12)

Table 7: Summary of the significant differences in socio-demographic characteristics between the ethnic groups

Variable (no. of categories)	Weighted to population distribution	Propensity score reweighted
Gender (2)		
Age group (6)	TS*; MS*; SA*	TS*; MS*; SA*
Immigration generation (2)	SA*	
Education level (4)	TS*; TA*; MS*; MA*	
Big city dweller (2)	TM*; TS*; SA*	
Employed (2)	TS*; TA*; MS*; MA*; SA*	
Children (2)	TA*;	
Partner (2)	TS*; TA*; MS*; MA*	TS*; TA*; MS*; MA*

Note: *significant p =<0.01; T = Turkish; M= Moroccans; S=Surinamese and A = Antilleans

differences remained, observed on two variables – *Age Group* and *Partner* – that were not included in the propensity score weighting model. The reason for their exclusion from the propensity score weighting model was that these socio-demographic variables did not have a significant impact on the indicators used to measure *Gender Roles* and *Family Ties* (see also Appendix C).

The comparison of Model 4 (M4) with Model 1 (M1) shows the impact of differences in sample composition for five socio-demographic variables (*Immigration generation, Educational level, Big city dweller, Employed and Children*, see Table 7) between ethnic, non-Western groups on the cross-cultural comparison of attitudes towards *Gender Roles* and *Family Ties.* Model 4 has a good to very good fit according to the criteria (Table 3).

The observed differences in attitudes towards *Gender Roles* between the ethnic groups are to some small degree the result of the differences in sample composition; the effect is even less noticeable for *Family Ties,* where differences in sample composition hardly affect the results at all (see Figures 1 and 2). With respect to *Gender Roles,* the attitudes are more alike when there is a correction for the incomparability of samples, as compared to Model 1, none of the significant differences observed between the ethnic groups persist (Table 4). This is not the case for *Family Ties,* where the correction only leads to a non-significant effect between Turkish and Moroccan compared to Model 1 (Table 5).

4 Conclusion and discussion

The present study investigated how interviewer effects, the use of an interviewer with a shared ethnic background, interview language, interviewer gender, gender matching, the presence of others during the interview and differences in socio-demographic sample composition of ethnic minority groups can affect the comparison of attitudes towards gender roles and family ties.

The data used in this study comes from a large scale face-to-face survey conducted between October 2010 and June 2011 for which Statistics Netherlands drew a random sample of named individuals from each of the four largest non-Western minority populations living in The Netherlands. The data contained not only answers to substantive questions, but also socio-demographic information on both respondent and interviewer characteristics, as well as interviewer observations regarding the interview.

As a first step, a multi group confirmatory factor analysis model approach was used to test for full scalar invariance of the two factor model (*Gender Roles* and *Familiy Ties*). The model showed an acceptable fit, which meant the latent factor means for both *Gender role* and *Family Ties* could be compared in a meaningful way across the four ethnic groups.

As for the first research question – "How do interviewer effects influence the cross-cultural comparison of attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups in the Netherlands?" – interviewer effects were added to this base model using the unique interviewer number as cluster variable. This reflected the data structure well and the results show that the addition of interviewer effects as cluster variable mostly lead to increased standard errors for all parameter estimates. The effect on the parameter estimates was marginal, which led to some minor changes in the estimated means of *Gender Roles* and *Family Ties*. As a result of the increased standard errors and a slight change in the relative position of Moroccans, it was shown that the observed cross-cultural difference on attitudes towards *Family Ties* between Moroccans and Antilleans was mostly the result of interviewer effects. This confirms our hypothesis that the observed differences between ethnic groups with respect to *Gender Roles* and *Familiy Ties* can be partly explained by interviewer effects.

The second research question – "How does the use of an interviewer with a shared ethnic background affect the cross-cultural comparison of attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups in the Netherlands?" – was addressed in terms of nonresponse, in which way does the increase in non-response due to language problems and cultural differences affect cross-cultural comparison between the ethnic groups? The estimated additional nonresponse as a result of not using bilingual interviewer was based on interview language and the interviewers assessment of the Dutch language proficiency level of the respond-

ent. The analysis showed that the increase in nonresponse had a significant impact on the cross-cultural comparison of *Gender Roles.* Without the use of bilingual interviewers with a shared ethnic background, the attitudes towards *Gender Roles* turned out to be a lot more similar across the ethnic groups. A specific group of respondents having a more traditional view would have been missed. This means that our hypothesis with respect to the second research question is also confirmed, at least with respect to nonresponse bias. The use of bilingual interviewers with a shared ethnic background resulted in more traditional views with respect to *Gender Roles* and *Familiy Ties.* The third research question – how does the language of the interview affect the comparison of attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups in the Netherlands – was assessed in combination with other potential sources of method bias. To find out how interview language affected cross-cultural comparison a dummy was made which, together with dummies indicating interviewer gender, gender match, the presence of others as well as eight important socio-demographic variables such as education, gender, age, etc., was regressed as covariate on the latent variables of *Gender Roles* and *Familiy Ties.* For this a multi group MIMIC (Multiple Indicators MultIple Causes) model was used. The inclusion of the socio-demographic variables on the respondents was done to correct as much as possible for the inherent confoundedness of these sources of method bias with respondent characteristics.

Interview language had an effect on attitudes towards *Familiy Ties* among Turkish, Moroccans and Antilleans. When interviewed in their native language, they all give (significantly) more traditional opinions. As for Surinamese, no significant effect of interview language was found for either factor. This is not surprising, since only a handful of respondents completed the interview in another language. Also in this instance the hypothesis is confirmed. Interview language has a systematic effect on the measurement of *Gender Roles* and *Familiy Ties* and being interviewed in Dutch leads to less traditional views towards *Gender Roles* and *Familiy Ties.*

There are several remarks that need to be made in order to place this result of interview language in the right context. First of all, the effect of interview language is confounded with the effect of interviewer ethnicity. However, all Turkish and Moroccan respondents were interviewed by bilingual interviewers with a shared ethnic background, therefore no further disentanglement was possible. On the other hand, some of the interviewer ethnicity effect might already be captured by the modeling of interviewer effects.

Secondly, this effect might also partially be the result of systematic differences introduced by translation. However, the latter is unlikely, since the effect was not detected for just one ethnic group, but for three, one of which never benefitted from a translated questionnaire at all. In addition, the effect was measured on the factor, not on the indicators.

Thirdly, it is clear that the measured effect is confounded with potential non-response bias. The respondents that could not have participated if the possibility to have the survey in their native language did not exist did show a more traditional attitude.

Despite the alternative explanations for the effect of interview language, the fact remains that it had a systematic effect. This means there is a real trade-off between cross-cultural comparability and reducing nonresponse among some ethnic groups.

As for the fourth research question – "How does interviewer gender and gender match affect the cross-cultural comparison?" – the results showed a significant effect for interviewer gender among Turkish and gender match among Moroccans when it came to attitudes towards *Gender Roles.* Perhaps not surprisingly, female interviewers cause systematically less traditional attitudes towards *Gender Roles* than male interviewers among the Turkish. Also, Moroccan men have more traditional attitudes towards *Gender Roles* when they are interviewed by a male interviewer compared to the Moroccan men that were interviewed by a female interviewer. Moroccan women are not systematically affected in their attitudes by the gender of the interviewer. In this case the hypothesis is partly confirmed. Interviewer gender and gender matching did effect the cross-cultural comparability, but the effect of interviewer gender was only discernible among Turkish respondents and the effect of gender match was only present among Moroccan male respondents.

With respect to the fifth research question – "How does the presence of others during the interview affect the cross-cultural comparison of attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups in the Netherlands?" – the results show that respondents of Surinamese and Antillean origin offered more traditional views in the presence of others. Among Surinamese respondents, this systematic effect was present on both factors, whereas for the Antilleans this only occurred for *Familiy Ties.* Also in this instance the hypothesis is only partly confirmed. The presence of others during an interview resulted in more traditional views towards *Gender Roles* and *Familiy Ties,* but only among Surinamese and only with respect to *Familiy Ties* among Antilleans.

The modeling of the incomparability of samples was done using a propensity score reweighting procedure of the socio-demographic variables that showed both a significant difference in the distribution between at least two ethnic groups and a significant effect on the indicators designed to measure the latent constructs.

The results for the sixth and final research question – "How much of the observed differences in attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups can be attributed to differences in socio-demographic composition between non-Western populations in the Netherlands?" – showed that the incomparability of samples explains some of the observed cross-cultural differences on both

Gender Roles and *Familiy Ties*. In the case of *Gender Roles,* this effect was large enough to render all observed differences between ethnic groups non-significant. This result confirms our sixth and final hypothesis that part of the observed differences between the ethnic groups can be explained by differences in socio-demographic composition.

It is important to be aware of the fact that survey data can be affected by a manifold of factors. These can be unwanted spin-offs of survey design choices or uncontrollable disturbance factors. In this case, it is clear that tailor-made response enhancing measures and other, less controllable sources of method bias affect the cross-cultural comparison of non-Western minority ethnic groups, not only because they introduce a bias in estimates for an ethnic group, but, more importantly, because they impact the groups differently.

In the case of face-to-face surveys designed to compare ethnic groups or countries, these effects can lead to wrong conclusions about the relative positions of countries or groups. This can have serious consequences if the survey results contribute towards deciding whether or not a policy is effective in reducing an observed socio-economic or socio-cultural difference or if it informs the decision about the allocation of funds.

The comparability bias can be caused by differences in the size of the various sources of method bias that affects the groups or countries under investigation, by the differential impact of the same method bias between groups or by a combination thereof.

In the case of cross-cultural studies, it is important for the researchers to be aware of how the data were collected and how this can potentially bias survey estimates. This is especially important in the case of unexpected results based on data that used different data collection strategies among different ethnic groups.

With respect to data collected via face-to-face surveys it is recommended to take into account potential interviewer effects to avoid spurious effects, especially in the context of cross-cultural comparisons. In those cases when no information about the interviewer is available, one may consider using stricter criteria for significance testing, such as increasing the significance level to 0.01 instead of 0.05.

With respect to cross-cultural comparison, one also needs to consider how the research question is reflected by the results of the comparison. A substitution of observed differences between cultures with cultural differences is easily done, but that will mostly be confounded with differences in socio-demographic composition. For instance, observed differences in the *Gender Roles* between the Turkish and Surinamese group can be interpreted as the average Turkish person being more traditional than the average Surinamese person. However, the average Turkish person has a different set of socio-demographic characteristics than the average Surinamese person. When Turkish and Surinamese persons with the same set of characteristics are compared the conclusion might be different.

The present study has several limitations that make the interpretation of the results not entirely straightforward. First of all, a MGCFA approach was used that included a cluster variable to adjust for interviewer-dependent correlation between the answers of respondents that were interviewed by the same interviewer. Given this modelling approach, it was not possible to compare the competing non-nested models using AIC or BIC fit indices. Therefore, the relative fit of the competing models was evaluated using fit measures that are not designed for comparing non-nested models and no conclusions could be drawn as to which of the models best describes the data. However, given the observed effects of the different sources of method bias on the cross-cultural comparability, we believe that we have adequately demonstrated the potential threat to making valid cross-cultural comparisons when these sources are not taken into account.

A second limitation concerns the quasi-experimental design used in this study. Data collected via this design does not allow for a complete disentanglement and entirely unbiased estimates of the different sources of identified method bias. Also, the data used in the present study did not allow for the complete disentanglement of the different ways (i.e., nonresponse, interview language and ethnicity) in which bilingual interviewers with a shared ethnic background can affect cross-cultural comparability.

A third limitation of the current study concerns the paradata. Several of the indicators measuring the existence of method bias are proxy estimates (i.e., interviewer assessments). A recommendation for further research could therefore be to include tape recordings of the interview in order to allow for more direct assessment of the effect of the interview language or of the extent to which others had an influence during (parts of) the interview.

As mentioned before, one can view the quasi-experimental design of this study as a drawback for this type of analysis. However, one should be aware of the fact that both the uncontrollable sources of method bias, such as the presence of others, as well as certain tailor-made response enhancing measures are always confounded with socio-demographic characteristics of respondents in cross-cultural surveys. Therefore, one may wonder if one should put effort in designing a fully randomized experimental design to capture these effects. Instead it may be more interesting to attempt building a body of evidence based on data collected via more realistic quasi experimental designs such as the present one, in order to gain a better understanding of the effect these inherently confounded sources of method bias can have on the comparability of cross-cultural surveys and of the extent to which they can compromise cross-cultural comparisons. It might be preferable to collect more and/or more direct paradata and to further develop models that are better suited to correcting or testing for the existence of these effects based on data collected via quasi-experimental designs.

References

AAPOR (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 7th edition. The American Association for Public Opinion Research. Retrieved from http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156 (last accessed March 2014).

Anderson, B. A., Silver, B. D., & Abramson, P. R. (1988). The effects of the race of the interviewer on race-related attitudes of black respondents in SRC/CPS national election studies. *Public Opinion Quarterly*, 52, 289-324.

Arends-Tóth, J. & Van de Vijver, F. J. (2008). Family relationships among immigrants and majority members in the Netherlands: The role of acculturation. *Applied Psychology*, 57, 466-487.

Baumgartner, H. & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38, 2, 143-156.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, 107, 2, 238-246.

Bia, M. & Mattei, A. (2008). A Stata package for the estimation of the dose-response function through adjustment for the generalized propensity score. *The Stata Journal*, 8, 354-373.

Billiet, J. B. & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research*, 36, 4, 542-562.

Billiet, J. B. & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural equation modeling: A Multidisciplinary Journal*, 7, 4, 608-628.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, 8, 4, 436-455.

Boehnke, K., Lietz, P., Schreier, M., & Wilhelm, A. (2011). Sampling: The selection of cases for culturally comparative psychological research. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 101-129). New York: Cambridge University Press.

Braun, M. & Johnson, T. P. (2010). An illustrative review of techniques for detecting inequivalences. In: J.A.Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler, B. Pennell, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 375-393). Wiley Online Library.

Campbell, B. A. (1981). Race-of-interviewer effects among southern adolescents. *Public Opinion Quarterly*, 45,2, 231-244.

Chen, C., Lee, S. y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, 6, 3, 170-175.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36,4, 462-494.

Cheung, G. W. & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, 31, 2, 187-212.

Cotter, P. R., Cohen, J., & Coulter, P. B. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, 46,2, 278-284.

Dagevos, J. & Gijsberts, M. (2009). Social-culture positie. In: M. Gijsberts & J. Dagevos (Eds.), *Jaarrapport Integratie 2009* (pp. 226-253). [In Dutch; Socio-Cultural position]. Den Haag: SCP.

Dagevos, J. & Schellingerhout, R. (2003). Sociaal-culturele integratie. Contacten, cultuur en orientatie op de eigen groep. In J.Dagevos, M. Gijsberts, & v. C. Praag (Eds). *Rapportage minderheden*. [In Dutch: Socio-Cultural integration. Contacts, culture and focus on the own ethnic group] Den Haag: SCP, pp. 317-362.

Dagevos, J., Schellingerhout, R., & Vervoort, M. (2007). Sociaal-culturele integratie en religie. In: J.Dagevos & M. Gijsberts (Eds.), *Jaarrapport Integratie 2007* (pp. 163-191). [In Dutch: Socio-Cultural integration and religion] Den Haag: SCP.

Davidov, E., Schmidt, P., & Billiet, J. (2011). *Cross-cultural analysis: Methods and applications*. London, England: Routledge.

Davis, D. W. (1997). The direction of race of interviewer effects among African-Americans: Donning the black mask. *American Journal of Political Science*, 41,1, 309-322.

Davis, R. E., Couper, M. P., Janz, N. K., Caldwell, C. H., & Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Education Research*, 25, 1, 14-26.

DiNardo, J. (2002). *Propensity score reweighting and changes in wage distributions*. Mimeo. http://www-personal.umich.edu/~jdinardo/bztalk5.pdf.

DiNardo, J., Fortin, N. M., & Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrics*, 64,5, 1001-1044.

Fernandez, A. L. & Marcopulos, B. A. (2008). A comparison of normative data for the Trail Making Test from several countries: Equivalence of norms and considerations for interpretation. *Scandinavian journal of psychology*, 49, 3, 239-246.

Feskens, R. C. W., Kappelhof, J., Dagevos, J., & Stoop, I. A. L. (2010). Minderheden in de mixed-mode? Een inventarisatie van voor- en nadelen van het inzetten van verschillende dataverzamelingsmethoden onder niet-westerse migranten. *SCP-special 57.* [In Dutch: Ethnic minorities in the mixed mode? An inventory of the advantages and disadvantages of employing different data collection methods among non-Western migrant] Den Haag: SCP.

Feskens, R., Hox, J., Lensvelt-Mulders, G., & Schmeets, H. (2006). Collecting data among ethnic minorities in an international perspective. *Field Methods*, 18, 3, 284-304.

Finkel, S. E., Guterbock, T. M., & Borg, M. J. (1991). Race-of-Interviewer Effects in a Preelection Poll Virginia 1989. *Public Opinion Quarterly*, 55,3, 313-330.

Groeneveld, S. & Weijers-Martens, Y. (2003). *Minderheden in beeld: SPVA-02*. [In Dutch: The focus on non-Western ethnic minorities: SPVA-02]. Rotterdam: Instituut voor Sociologisch-Economisch Onderzoek (ISEO).

Groves, R. M. & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T., Lyberg, L., Mohler, P. et al. (2010). *Survey methods in multinational, multiregional, and multicultural contexts.* Wiley: Hoboken, NJ.

He, J., & van de Vijver, F. (2012). Bias and Equivalence in Cross-Cultural Research. *Online Readings in Psychology and Culture*, 2(2). http://dx.doi.org/10.9707/2307-0919.1111

He, J. & Van de Vijver, F. J. (2013). A general response style factor: Evidence from a multiethnic study in the Netherlands. *Personality and Individual Differences*, 55,7, 794-800.

Hox, J. J., de Leeuw, E. D., & Brinkhuis, M.J.S. (2010). Analysis models for comparative surveys. In: Harkness, J., Braun, M., Edwards, B., Johnson, T., Lyberg, L., Mohler, P., Pennell, B.E., and Smith, T.W. (Eds.) *Survey Methods in multinational, multiregional, and multicultural contexts*. Hoboken, NJ: Wiley. Pp. 395-418.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6,1, 1-55.

Huang, I., Frangakis, C., Dominici, F., Diette, G. B., & Wu, A. W. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health services research*, 40, 1, 253-278.

Hui, C. H. & Triandis, H. C. (1985). Measurement in Cross-Cultural Psychology A Review and Comparison of Strategies. *Journal of Cross-Cultural Psychology*, 16, 2, 131-152.

Hui, C. H. & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20, 3, 296-309.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 3, 706-710.

Jann, B. (2008). The Blinder-Oaxaca decomposition for linear regression models. *The Stata Journal*, 8, 453-479.

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. *ZUMA-Nachrichten spezial*, 3, 1-40.

Johnson, T. P., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles evidence from 19 countries. *Journal of Cross-Cultural Psychology*, 36, 2, 264-277.

Johnson, T. P. & Van de Vijver, F. J. (2003). Social desirability in cross-cultural research. In J. Harness, F. J. van de Vijver, & Mohler, P. (Eds.), *Cross-cultural survey methods* (pp. 193–202). New York: Wiley.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36,2, 109-133.

Kankaras, M. & Moors, G. (2009). Measurement equivalence in solidarity attitudes in Europe insights from a multiple-group latent-class factor approach. *International Sociology*, 24, 4, 557-579.

Kappelhof, J. W. S. (Accepted). The effect of different survey designs on nonresponse in surveys among non-Western minorities in The Netherlands. *Survey Research Methods*, to appear in volume 9, 2, 2014.

Kemper, F. (1998). Gezocht: Marokkanen. Methodische problemen bij het verwerven en interviewen van allochtone respondenten. [In Dutch: Wanted: Moroccans. Methodological problems with obtaining response and interviewing respondents of foreign origin]. *Migrantenstudies*, 1, 43-57.

Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. *Sociological Methods & Research*, 33, 2, 188-229.

Lee, R. M. (1993). *Doing research on sensitive topics*. London, UK.: Sage.

Leung, K., Lau, S., & Lam, W. L. (1998). Parenting styles and academic achievement: A cross-cultural study. *Merrill-Palmer Quarterly* (1982-), 44, 2, 157-172.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within-and between-group differences and measurement invariance in the common factor model. *Intelligence*, 31, 6, 543-566.

Lubke, G. H. & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural equation modeling*, 11, 4, 514-534.

Lubke, G. H. & Muthén, B.O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, 10, 1, 21-39.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme Response Style and Acquiescence among Hispanics The Role of Acculturation and Education. *Journal of Cross-Cultural Psychology*, 23, 4, 498-509.

Martens, E. P. (1999). *Minderheden in beeld: SPVA-98*. [In Dutch: The focus on non-Western ethnic minorities: SPVA-98]. Rotterdam: NIWI.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13, 2, 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 4, 525-543.

Meredith, W. & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical care*, 44, 11, S69-S77.

Morren, M., Gelissen, J. P., & Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: a restricted latent class factor analysis approach. *Sociological Methodology*, 41, 1, 13-47.

Morren, M., Gelissen, J., & Vermunt, J. (2012a). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 8, 4, 159-170.

Morren, M., Gelissen, J. P., & Vermunt, J. K. (2012b). Response Strategies and Response Styles in Cross-Cultural Surveys. *Cross-Cultural Research*, 46, 3, 255-279.

Muthén, L. K. & Muthén, B. O. (2011). *Mplus User's Guide*. Sixth Edition. [Computer software]. Los Angeles, CA.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review*, 14, 3, 693-709.

Poortinga, Y. H. & Van de Vijver, F. J. (1987). Explaining Cross-Cultural Differences Bias Analysis and Beyond. *Journal of Cross-Cultural Psychology*, 18, 3, 259-282.

Rhodes, P. J. (1994). Race-of-interviewer effects: a brief comment. *Sociology*, 28, 547-558.

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

Ross, C. E. & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior*, 25,2, 189-197.

Schmeets, H. & van der Bie, R. (2005). *Enqueteonderzoek onder allochtonen. Problemen en oplossingen*. [In Dutch: survey research among minorities. Problems and solutions]. Voorburg/Heerlen: CBS.

Schuman, H. & Converse, J. M. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, 35, 1, 44-68.

Steiger, J. H. (1989). *EzPATH: Causal modeling*. Evanston, IL: SYSTAT.

Stoop, I. A. L. (2005). *The hunt for the last respondent: Nonresponse in sample surveys*. The Hague: The Netherlands institute for Social Research/SCP.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. Chichester, UK: John Wiley & Sons.

Sudman, S. & Bradburn, N. M. (1974). R*esponse effects in surveys: A review and synthesis*. Aldine Publishing Company Chicago, Ill.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, 133, 5, 859-833. Retrieved from American Psychological Association

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1, 1-10.

Van Heelsum, A. J. (1997). *De etnisch-culturele positie van de tweede generatie Surinamers*. Doctoral Dissertation. Amsterdam: Free University. http://hdl.handle.net/1871/13062

Van't Land, H. (2000). *Similar Questions: Different Meanings. Differences in the Meaning of Constructs for Dutch and Moroccan Respondents; Effects of the Ethnicity of the Interviewer and Language of the Interview among First and Second Generation Moroccan Respondents*. Vrije Universiteit Amsterdam, Amsterdam.

Van de Vijver, F. J. R. (2003). Bias and equivalence: Cross-cultural perspectives. In J. Harness, F. J. van de Vijver, & Mohler, P. (Eds.), *Cross-cultural survey methods* (pp. 143-155). New York: Wiley.

Van de Vijver, F. J.R. (2011). Capturing bias in structural equation modeling. In E.Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis. Methods and applications* (pp. 3-34). London, England: Routledge.

Van de Vijver, F. & Leung, K. (1997). Methods and data analysis of comparative research.In: Berry, John W.; Poortinga, Ype H.; Pandey, Janak (Eds). *Handbook of cross-cultural psychology*, Vol. 1: Theory and method (2nd ed.). Handbook of cross-cultural psychology (2nd ed.)., (pp. 257-300). Needham Heights, MA, US: Allyn & Bacon, xxv, 406 pp.

Van de Vijver, F.J.R. & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Europeenne de Psychologie Appliquee/European Review of Applied Psychology*, 54, 2, 119-135.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3,1, 4-70.

Veenman, J. (2002). Interviewen in multicultureel Nederland. In: H. Houtkoop en Veenman, J. (Eds), *Interviewen in de multiculturele samenleving. Problemen en oplossingen*. [In Dutch: Interviewing in the multi-cultural Netherlands] Assen: Koninklijke Van Gorcum. pp. 1-19.

Wicherts, J. M. (2007). *Group Differences in Intelligence Test Performance*. Universiteit van Amsterdam, Amsterdam.

Williams Jr, J. A. (1964). Interviewer-respondent interaction: A study of bias in the information interview. *Sociometry*, 27, 338-352.

Van der Zouwen, J. (2006). De interviewer, hulp of hindernis? In: A.E.Bronner, P. Dekker, E. D. d. Leeuw, L. J. Paas, K. d. Ruyter, A. Smidts, & J. E. Wieringa (Eds.), *Ontwikkelingen in het Marktonderzoek*, Jaarboek 2006 (pp. 63-76). [In Dutch: The Interviewer: help or impediment?] haarlem: spaarenhout.

Appendix A:
Parameter estimates and standard errors of the five multi group models

Parameter estimates (se)	M0		M1		M2		M3		M4	
$\overline{Gr_M}$	0.106	(0.047)	0.085	(0.113)	0.049	(0.119)	0.202[a]	(0.098)	0.096	(0.140)
Gr_S	0.556	(0.056)	0.595	(0.152)	0.530	(0.151)	0.831[a]	(0.103)	0.513	(0.173)
Gr_A	0.511	(0.054)	0.516	(0.121)	0.479	(0.121)	0.743[a]	(0.091)	0.479	(0.148)
$\overline{Ft_M}$	0.271	(0.053)	0.319	(0.084)	0.259	(0.093)	0.230[a]	(0.066)	0.307	(0.084)
Ft_S	1.057	(0.066)	1.085	(0.094)	0.945	(0.097)	1.022[a]	(0.065)	1.017	(0.103)
Ft_A	1.194	(0.069)	1.220	(0.087)	1.100	(0.093)	1.195[a]	(0.055)	1.184	(0.101)
$Corr(Gr,FT)_T$	0.272	(0.029)	0.270	(0.039)	0.208	(0.038)	0.240	(0.041)	0.268	(0.027)
$Corr(Gr,FT)_M$	0.193	(0.029)	0.199	(0.041)	0.210	(0.046)	0.192	(0.047)	0.222	(0.048)
$Corr(Gr,FT)_S$	0.406	(0.047)	0.416	(0.103)	0.406	(0.102)	0.330	(0.080)	0.468	(0.155)
$Corr(Gr,FT)_A$	0.421	(0.045)	0.413	(0.073)	0.404	(0.075)	0.320	(0.057)	0.475	(0.082)
$\lambda^{Gr}_{Mangeld}$	1.000	(fixed)	1.000	(fixed)	1.000	(fixed)	1.000	(fixed)	1.000	(fixed)
$\lambda^{Gr}_{Inkjongs}$	0.951	(0.027)	0.949	(0.031)	0.949	(0.040)	0.956	(0.036)	0.949	(0.038)
$\lambda^{Gr}_{Vrwstopw}$	0.839	(0.025)	0.843	(0.036)	0.856	(0.042)	0.786	(0.042)	0.838	(0.037)
$\lambda^{Ft}_{Thuishuw}$	1.000	(fixed)	1.000	(fixed)	1.000	(fixed)	1.000	(fixed)	1.000	(fixed)

Appendix A continued

Parameter estimates (se)	M0		M1		M2		M3		M4	
$\lambda_{Vertrffama}^{Ft}$	0.608	(0.033)	0.574	(0.040)	0.608	(0.045)	0.576	(0.060)	0.558	(0.039)
$\lambda_{Kibezoud}^{Ft}$	0.668	(0.034)	0.667	(0.047)	0.705	(0.059)	0.655	(0.073)	0.644	(0.050)
$\tau_1^{Mangeld}$	-1.419	(0.062)	-1.457	(0.146)	-1.550	(0.145)	-1.755	(0.332)	-1.608	(0.147)
$\tau_2^{Mangeld}$	-0.543	(0.042)	-0.528	(0.102)	-0.638	(0.100)	-0.774	(0.302)	-0.670	(0.105)
$\tau_3^{Mangeld}$	-0.079	(0.039)	-0.085	(0.097)	-0.160	(0.096)	-0.270	(0.292)	-0.176	(0.112)
$\tau_4^{Mangeld}$	1.003	(0.052)	1.017	(0.137)	0.966	(0.137)	0.868	(0.283)	0.983	(0.172)
$\tau_1^{Inkjongs}$	-1.371	(0.057)	-1.415	(0.131)	-1.471	(0.133)	-1.554	(0.322)	-1.501	(0.123)
$\tau_2^{Inkjongs}$	-0.440	(0.039)	-0.434	(0.092)	-0.519	(0.093)	-0.516	(0.291)	-0.506	(0.099)
$\tau_3^{Inkjongs}$	-0.101	(0.038)	-0.120	(0.094)	-0.192	(0.092)	-0.148	(0.287)	-0.194	(0.107)
$\tau_4^{Inkjongs}$	0.981	(0.051)	1.006	(0.132)	0.965	(0.132)	1.031	(0.281)	0.965	(0.164)
$\tau_1^{Vrwstopw}$	-1.473	(0.059)	-1.457	(0.128)	-1.564	(0.127)	-1.451	(0.277)	-1.606	(0.123)
$\tau_2^{Vrwstopw}$	-0.573	(0.039)	-0.596	(0.081)	-0.714	(0.078)	-0.535	(0.251)	-0.715	(0.081)
$\tau_3^{Vrwstopw}$	-0.183	(0.035)	-0.208	(0.077)	-0.295	(0.076)	-0.131	(0.241)	-0.302	(0.084)
$\tau_4^{Vrwstopw}$	0.940	(0.047)	0.925	(0.126)	0.883	(0.130)	1.059	(0.242)	0.869	(0.145)
$\tau_1^{Thuishuw}$	-0.791	(0.051)	-0.722	(0.106)	-0.820	(0.125)	-0.692	(0.264)	-0.866	(0.099)
$\tau_2^{Thuishuw}$	0.315	(0.043)	0.335	(0.057)	0.201	(0.066)	0.416	(0.247)	0.235	(0.060)

Appendix A continued

Parameter estimates (se)	M0		M1		M2		M3		M4	
$\tau_3^{Thuishuw}$	0.652	(0.047)	0.673	(0.058)	0.544	(0.066)	0.788	(0.249)	0.569	(0.066)
$\tau_4^{Thuishuw}$	1.825	(0.078)	1.827	(0.103)	1.697	(0.112)	1.995	(0.281)	1.838	(0.120)
$\tau_1^{Vertrfama}$	-0.606	(0.043)	-0.483	(0.076)	-0.545	(0.086)	-0.202	(0.187)	-0.646	(0.069)
$\tau_2^{Vertrfama}$	0.736	(0.040)	0.767	(0.046)	0.711	(0.054)	0.987	(0.208)	0.701	(0.051)
$\tau_3^{Vertrfama}$	1.432	(0.058)	1.408	(0.059)	1.370	(0.077)	1.625	(0.235)	1.411	(0.060)
$\tau_4^{Vertrfama}$	2.490	(0.098)	2.394	(0.107)	2.419	(0.142)	2.530	(0.301)	2.544	(0.099)
$\tau_1^{Kibezoud}$	-0.367	(0.039)	-0.297	(0.075)	-0.329	(0.080)	0.030	(0.206)	-0.375	(0.070)
$\tau_2^{Kibezoud}$	0.881	(0.044)	0.880	(0.066)	0.818	(0.072)	1.203	(0.233)	0.780	(0.067)
$\tau_3^{Kibezoud}$	1.286	(0.057)	1.266	(0.082)	1.200	(0.089)	1.600	(0.256)	1.165	(0.086)
$\tau_4^{Kibezoud}$	2.195	(0.090)	2.164	(0.139)	2.120	(0.152)	2.490	(0.325)	2.107	(0.149)
χ^2	552.900		302.735		285.621		475.207		273.996	
Df	92		92		92		348		92	

Note. GR= Gender Roles and FT= Family Ties; T= Turkish; M= Moroccans; S=Surinamese; A= Antilleans; $GR_{Turkish}$ and $FT_{Turkish}$ are both set to zero. λ^{factor} = factorloading of the indicator; τ_x = threshold value of the indicator. a = adjusted for the (different) impact of the presence of others, own language, interviewer gender and gender match interaction between ethnic groups.

Appendix B:
Multiple causes results for Model 3 for Gender Roles (GR) and Family Ties (FT), separately for each ethnic group

Parameter estimates (se)	Turkish (N=812)				Moroccans (N=805)				Surinamese (N=779)				Antilleans (N=852)			
	GR		FT		GR		FT		GR		FT		GR		FT	
Intercept	0.000	(0.000)	0.000	(0.000)	0.566	(0.371)	0.583	(0.432)	0.404	(0.485)	1.272	(0.388)*	0.611	(0.362)	1.395	(0.348)*
Big City Dweller	-0.340	(0.183)	-0.069	(0.098)	-0.198	(0.128)	-0.154	(0.141)	0.045	(0.141)	-0.113	(0.100)	-0.219	(0.092)*	-0.273	(0.120)*
Employed	0.229	(0.075)*	0.018	(0.073)	0.179	(0.066)*	0.019	(0.177)	0.182	(0.109)	0.101	(0.084)	0.044	(0.068)	0.059	(0.079)
Has Child(ren)	-0.180	(0.171)	-0.388	(0.114)*	0.110	(0.105)	0.019	(0.177)	0.080	(0.093)	-0.071	(0.097)	0.008	(0.111)	-0.225	(0.096)*
Has a partner	0.050	(0.093)	-0.096	(0.089)	-0.120	(0.092)	-0.260	(0.140)	0.088	(0.067)	0.085	(0.073)	0.064	(0.071)	0.077	(0.073)
Educational level	0.101	(0.043)*	0.180	(0.046)*	0.082	(0.036)*	0.104	(0.047)*	0.171	(0.065)*	0.088	(0.048)	0.187	(0.047)*	0.272	(0.052)*
Male	-0.232	(0.090)*	0.176	(0.101)	-0.579	(0.081)*	-0.217	(0.113)	-0.671	(0.145)*	-0.057	(0.074)	-0.604	(0.099)*	-0.080	(0.099)
First generation immigrant	0.032	(0.154)	0.013	(0.121)	0.093	(0.125)	-0.192	(0.122)	-0.223	(0.093)*	-0.426	(0.096)*	-0.286	(0.094)*	-0.264	(0.113)*
Age group (ref group is 15-24)																
25 – 34 year	0.046	(0.149)	0.443	(0.167)*	0.126	(0.104)	0.288	(0.162)	0.004	(0.139)	0.168	(0.130)	0.004	(0.105)	-0.090	(0.119)
35 – 44 year	0.162	(0.174)	0.558	(0.178)*	-0.013	(0.140)	0.353	(0.288)	-0.153	(0.131)	0.221	(0.151)	0.017	(0.116)	-0.007	(0.150)
45 – 54 year	0.016	(0.189)	0.554	(0.191)*	0.004	(0.145)	0.495	(0.211)*	-0.115	(0.149)	0.106	(0.137)	0.075	(0.140)	0.130	(0.146)
55 – 64 year	0.069	(0.139)	0.510	(0.179)*	-0.045	(0.182)	0.513	(0.286)	-0.066	(0.154)	0.133	(0.152)	0.001	(0.127)	-0.219	(0.159)

Appendix B continued

Parameter estimates (se)	Turkish (N=812)				Moroccans (N=805)				Surinamese (N=779)				Antilleans (N=852)			
	GR		FT		GR		FT		GR		FT		GR		FT	
65 + year	-0.075	(0.221)	0.344	(0.232)	-0.115	(0.183)	0.331	(0.262)	-0.147	(0.183)	-0.061	(0.161)	-0.135	(0.182)	-0.060	(0.229)
Others were present	-0.249	(0.160)	-0.109	(0.170)	-0.012	(0.159)	-0.298	(0.184)	-0.689	(0.202)*	-0.405	(0.162)*	-0.062	(0.119)	-0.259	(0.100)*
Interviewed in native language	-0.142	(0.100)	-0.364	(0.131)*	-0.105	(0.114)	-0.356	(0.140)*	-0.414	(0.856)	-0.013	(0.438)	-0.169	(0.132)	-0.241	(0.113)*
Gender match interaction	-0.294	(0.117)*	-0.048	(0.162)	0.133	(0.184)	0.015	(0.208)	0.168	(0.182)	0.123	(0.117)	0.006	(0.126)	-0.079	(0.145)
Gender interviewer	-0.022	(0.157)	-0.043	(0.156)	-0.339	(0.159)*	-0.070	(0.195)	0.031	(0.197)	-0.050	(0.120)	-0.054	(0.102)	-0.094	(0.128)

Note. * = $p < 0.05$

Appendix C:
Observed differences on socio-demographic variables between ethnic groups after weighting for population distribution (Table C1) and after propensity score weighting (Table C2).

Table C1: Observed differences on socio-demographic variables between ethnic groups after weighting for population distribution

				Significant differences between ethnic groups (bonferonni adjusted)		
Variable	Ethnic group	estimate	se	Turkish	Moroccans	Surinamese
Men (proportion)	Turkish	0.517	0.019			
	Moroccans	0.506	0.018			
	Surinamese	0.464	0.018			
	Antilleans	0.494	0.018			
Age Group (mean)	Turkish	2.750	0.052			
	Moroccans	2.739	0.053			
	Surinamese	3.079	0.054	*	*	
	Antilleans	2.710	0.052			*
First generation immigrant (proportion)	Turkish	0.693	0.018			
	Moroccans	0.664	0.017			
	Surinamese	0.646	0.017			
	Antilleans	0.721	0.016			*
Educational level (mean)	Turkish	2.074	0.039			
	Moroccans	2.005	0.038			
	Surinamese	2.607	0.037	*	*	
	Antilleans	2.533	0.035	*	*	
Big City Dweller (proportion)	Turkish	0.228	0.016			
	Moroccans	0.299	0.016	*		
	Surinamese	0.360	0.018	*		
	Antilleans	0.254	0.016			*

Table C1 continued

				Significant differences between ethnic groups (bonferonni adjusted)		
Variable	Ethnic group	estimate	se	Turkish	Moroccans	Surinamese
Employed (proportion)	Turkish	0.489	0.019			
	Moroccans	0.488	0.018			
	Surinamese	0.674	0.017	*	*	
	Antilleans	0.601	0.018	*	*	*
Has child(ren) (proportion)	Turkish	0.632	0.019			
	Moroccans	0.591	0.018			
	Surinamese	0.615	0.018			
	Antilleans	0.548	0.018	*		
Has partner (proportion)	Turkish	0.579	0.019			
	Moroccans	0.573	0.018			
	Surinamese	0.506	0.018	*	*	
	Antilleans	0.458	0.017	*	*	

Note. * p<0.05/no. of pairwise comparisons. Variables included in the population weights: gender, household size, municipality size, immigration generation, age groups (12)

Table C2: Observed differences on socio-demographic variables between ethnic groups after propensity score weighting

				Significant differences between ethnic groups (bonferonni adjusted)		
Variable	Ethnic group	estimate	se	Turkish	Moroccans	Surinamese
Men (proportion)	Turkish	0.523	0.025			
	Moroccans	0.523	0.020			
	Surinamese	0.494	0.017			
	Antilleans	0.515	0.019			
Age Group (mean)	Turkish	2.522	0.049			
	Moroccans	2.562	0.045			
	Surinamese	3.141	0.056	*	*	
	Antilleans	2.757	0.051			*
First generation immigrant (proportion)	Turkish	0.621	0.024			
	Moroccans	0.617	0.021			
	Surinamese	0.641	0.017			
	Antilleans	0.639	0.018			
Educational level (mean)	Turkish	2.628	0.048			
	Moroccans	2.640	0.045			
	Surinamese	2.597	0.037			
	Antilleans	2.621	0.036			
Big City Dweller (proportion)	Turkish	0.352	0.025			
	Moroccans	0.339	0.021			
	Surinamese	0.327	0.017			
	Antilleans	0.326	0.019			
Employed (proportion)	Turkish	0.687	0.018			
	Moroccans	0.673	0.018			
	Surinamese	0.662	0.017			
	Antilleans	0.671	0.016			

Table C1 continued

				Significant differences between ethnic groups (bonferonni adjusted)		
Variable	Ethnic group	estimate	se	Turkish	Moroccans	Surinamese
Has child(ren) (proportion)	Turkish	0.585	0.024			
	Moroccans	0.583	0.021			
	Surinamese	0.617	0.017			
	Antilleans	0.575	0.018			
Has partner (proportion)	Turkish	0.632	0.022			
	Moroccans	0.589	0.021			
	Surinamese	0.511	0.018	*	*	
	Antilleans	0.499	0.018	*	*	

Note. * p<0.05/no. of pairwise comparisons. Variables included in the propensity score reweighting: Immigration generation, Educational level, Big city dweller, Employed and Children

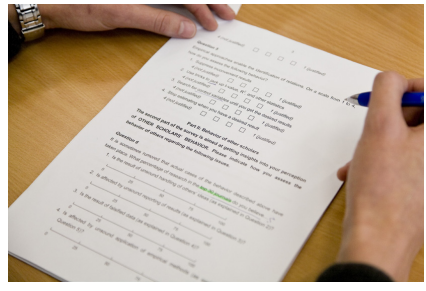
gesis
Leibniz Institute for the Social Sciences

GESIS Pretest Lab

Why pretest a questionnaire?

Did you know that pretesting helps to....

- identify questions that are misinterpreted by (some) respondents? (Sudman et al., 1996)
- reduce systematic measurement error, e.g. by identifying context effects? (Groves et al., 2004)
- reduce item non-response? (Forsyth et al., 2004)



The GESIS Pretest Lab supports researchers in optimizing their survey questions prior to data collection to improve the quality of the data obtained by the survey.

What we offer:

- Conducting cognitive pretests
- Conducting eye movement analyses (eye tracking) in combination with cognitive pretests
- Consulting on choosing and independent usage of various pretesting methods

For more informationen about our services and the costs and duration of questionnaire pretests please visit:
http://www.gesis.org/en/services/data-collection/pretest-lab/

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript by e-mail to mda(at)GESIS(dot)org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
- Please also send us an abstract of your paper (approx. 300 words), a brief biographical note (no longer than 250 words), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - Tiff
  - Jpeg (uncompressed, high quality)
  - pdf
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.

Please follow the APA guidelines when preparing in-text references and the list of references.

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).