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.