Huschka, Denis; Wagner, Gert G.

Working Paper

Statistical Problems and Solutions in Onomastic Research: Exemplified by a Comparison of Given Name Distributions in Germany throughout the 20th Century

SOEPpapers on Multidisciplinary Panel Data Research, No. 332

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Huschka, Denis; Wagner, Gert G. (2010) : Statistical Problems and Solutions in Onomastic Research: Exemplified by a Comparison of Given Name Distributions in Germany throughout the 20th Century, SOEPpapers on Multidisciplinary Panel Data Research, No. 332, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/150877

Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

www.diw.de

SOEP The German Socio-Economic Panel Study

SOEPpapers

on Multidisciplinary Panel Data Research

332

Denis Huschka • Gert G. Wagner

Statistical Problems and Solutions in Onomastic Research – Exemplified by a Comparison of Given Name Distributions in Germany throughout the 20th Century

Berlin, November 2010

SOEPpapers on Multidisciplinary Panel Data Research
at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | urahmann@diw.de

Statistical Problems and Solutions in Onomastic Research – Exemplified by a Comparison of Given Name Distributions in Germany throughout the 20th Century

by

Denis Huschka* and Gert G. Wagner**

November 2010

The authors would like to thank Joachim R. Frick, DIW Berlin, Jürgen Gerhards, Free University of Berlin, Gabriele Rodriguez, University of Leipzig, Ulrich Kohler, Social Science Center Berlin (WZB), and Anja Bruhn and Toby Carrodus (German Data Forum) for their help and suggestions. The views expressed here and all remaining errors are our own.

*) German Institute for Economic Research (DIW Berlin) and Institute of Social and Economic Research, Rhodes University / South Africa; contact: dhuschka@ratswd.de

**) German Institute for Economic Research (DIW Berlin), Berlin University of Technology (TUB), and Max Planck Institute for Human Development, Berlin; contact: gwagner@diw.de

Abstract

The German Socio Economic Panel Study (SOEP) offers the rare opportunity to look at patterns of given names amongst a representative sample of more than 50,000 people born since 1900.

In this paper, we first discuss the advantages and limitations of various data bases which have been widely used to study the distribution of given names. Second, we address the problem that name distributions are typically characterized by a "Large Number of Rare Events" (LNRE) zone. With regard to this, we focus our attention on the difficulties associated with comparing name distributions based on different sample surveys. Third, we apply some measures of the concentration of distributions from other lines of research (economics and computational linguistics). Finally, we stress the problem of the statistical significance of differences in name distributions based on samples.

Keywords: Given names, large number of rare events (LNRE), concentration of distributions, SOEP

JEL Classification: C49, C83, Y8

---

1 Introduction

Children's given names are widely unadulterated expressions of the personal taste of the name givers; usually, the parents. But parental taste and concrete name choices are influenced by societal and socio-structural settings.[1] Thus, it is of potential scientific interest to analyze the societal circumstances in which the social act of "name selection" takes place.[2] In most western societies there exist very little juridical restrictions limiting name choice;[3] naming is generally both *free* and *for free.*

In fact, researchers from various disciplines have recently stumbled upon given names as an area of interest.[4] Given names are attractive indicators for the analysis of societal structures and developments, not only for (socio)-onomasticians, but also for sociologists, historians, and economists.[5] For example, economists have come to analyze given names as variables in order to study the discriminating effects of particular names in the job markets.[6] Other scholars explicitly investigate the distribution patterns of names from a more "technical" perspective to learn about regularities and changes in name frequency-rank distributions.[7]

The most serious problems of empirical sociological and economic research on names are (1) a shortage of appropriate data and (2) the unequal numbers of observations in the samples used for comparative studies, which make valid comparisons not impossible but difficult. Addressing these methodological issues is the main focus of this paper.

---

[1] See Huschka, Gerhards & Wagner 2009, or Besnard & Desplaques 2001; Gerhards 2003, 2005; Lieberson 2000 and Wolffsohn & Brechenmacher 1999.

[2] From the great German sociologist Max Weber's point of view, the selection of a name is a social act. Of course, the act of selecting a name is a very personal decision undertaken by the parents, since there is no commercial "naming industry" that forces people to favor certain names.

[3] In Germany, parents are generally free to choose their children's names. Only the invention of *really* new names is more difficult than in other countries (e.g., the U.S.), especially if these names are not merely variations in the spelling of other existing names. However, the reservoir of possible names has always been huge. During the long shared history of Europe, tens of thousands of names have emerged.

[4] See recently Huschka et al. 2009 and earlier, for instance, Besnard et al. 2001; Gerhards 2003, 2005; Gerhards & Hackenbroch 1997, 2000 Lieberson 2000;Lieberson & Bell 1992. Also the studies of the historians Wolffsohn and Brechenmacher 1999 and the economist Galbi 2001, 2002 can be seen as contributing to the enlightenment in the field of a rather "sociological onomastic."

[5] From this point onwards, unless otherwise stated, the use of "names" refers exclusively to what are commonly known as "given names" or "first names."

[6] See e.g. Arai, Skogman & Thoursie 2004; Aura and Hess 2004; Bertrand & Mullainathan 2003; Fryer & Levitt 2003.

[7] See Eshel 2001; Hanks & Tucker 2000 and. Tucker 2001.

1. In section 2 of the paper we address the problems of available databases. Telephone books, for example, are not representative for a population in a certain area or country. Moreover, telephone books do not contain the contextual parental information necessary for analysis of processes such as, for example, societal modernization or transnationalization. Even for analyses of the distributions of names, the validity of results based on telephone book data is limited. What is crucial for making inferences and general conclusions about naming trends and naming innovations valid for a whole society, or for comparisons over time or between countries, are comparable files of names that represent the society under investigation. We will show that large representative sample surveys can be attractive alternatives to telephone directories and church registers. We exemplify this by analyzing German naming patterns based on data from the German Socio- Economic Panel Study (hereafter SOEP).
2. The main section of this article, section 3, is a reflection of methodological problems. First of all, we discuss the problems of the so called "Large Numbers of Rare Events" (hereafter LNRE) zone in name distributions; that is, that the name distribution in a society is usually characterized by the occurrence of only few very popular names and a very large number of (very) rare names. The occurrence of a LNRE zone in sample surveys leads to unstable mean frequencies, relative frequencies, and other parameters of distributions. We present a simple solution for this problem. These methodological discussions are extended by a discussion of sampling errors. Finally, we apply some indicators developed in economics and apply them to name distributions. The idea behind this is that economists deal with similar problems when they analyze the "market power" of firms in an economy.
3. In the last section of the paper, as an example, we present socio-onomastic results for Germany. These results are based on the representative data of the SOEP, allowing us to draw conclusions about the (changing) name patterns in Germany since the 1900s. Concluding remarks follow.

2 Data Bases

2.1 Telephone Directories, Church Registers, and Social Science Surveys

Generally, researchers in the field of names research have to battle an imperfect data situation. Most of the studies we know so far use telephone directories or church registers as their data basis.[8] However, all of these studies systematically lack validity since they are based on *non-representative* data which do not sufficiently take all the social and societal components of naming (e.g., naming differences along the social strata) into account. Further, some analysts (e.g., Eshel 2001) even compare non-representative sub-samples of very unequal size from telephone directories for several countries, and, on this basis, suggest conclusions.

Even outstanding studies, such as that of Tucker (2001), which was conducted on the basis of approximately 70 million US telephone book listings, suffer from the major problem associated with directories: they do not adequately cover societal reality. The basic problem of telephone book data is the unequal selection probability of the listed persons. This has several implications:

a) The most severe point is the under-representation of women: married women are less likely to be listed because many families are only found under the husband's entry.[9] Younger single women may indeed be represented, but a significant share may not be due to personal security protection issues.[10]

b) Since children are barely listed, studies based on telephone books have a "delay" of approximately 16 to 20 years in capturing recent developments in naming. In addition, telephone directory listings are susceptible to the age structure of a society: in most Western societies, for example, the elderly make up the largest share of society and so are listed in directories more frequently than younger people. Assertions made on such a basis therefore

[8] Nonetheless, we have recently become aware of some quite inventive strategies for names research when dealing with unorthodox data sources. Tucker (2004), in his recent research, uses data from the electoral roles of Ireland. Even lists of names of credit card holders have been made accessible to other authors.

[9] Tucker (2001:71) himself notes the problem of dealing with listings like "Mr. and Mrs. Frank Churchill." For a discussion on the limitations of using telephone directory data as a basis for names research see Hanks &Tucker (2000).

[10] Since telephone directory data usually over-represents men, it is only possible to draw conclusions on gender differences in naming based on directory data if at least a ratio adjustment by random sampling measures has been undertaken. This ensures that the sub-samples for both sexes are comparable; however, this procedure still does not make the telephone books representative.

overestimate those name types that were given to people decades ago, while newer, and the latest, trends are underestimated.

c) Moreover, the probability of being listed in a telephone directory is influenced by social structures. For example, elites and members of the upper class prefer not to be listed due to security concerns.

d) A last additional, and as of recent increasingly important point, is the huge amount of exclusive cellular phone users whose numbers are rarely found in directories.

In general, telephone book data are not representative for the population of a given country. Conclusions made on the basis of such data may shed some interesting and valuable insights about naming and name trends, but, on account of the name distributions, cannot be seen as representative of a society.

The hitherto most widely used alternatives to telephone directories are church registers and registry offices (see e.g., Galbi 2001, 2002; Gerhards 2003, 2005 or Simon 1989). These data sources contain, in addition to names, "explanatory" data such as year of birth. They are, however, not without limitations: most studies using such data are restricted to *regional* studies since central registers for nation-wide data are usually not available. By and large, single birth registers cover only children born within a certain district. A clear advantage of using data from single birth registers is the availability of additional variables. For more in-depth sociological analyses of names and naming within a whole society across the social strata, however, the value of single church registers and registry office data is limited too.

It is not our intention to peculate the merits of well-conducted studies using telephone directory data or registers. Certainly, they provide us with valuable knowledge about names, naming, and developments over time. But for the aforementioned reasons, these studies are simply not able to observe recent naming patterns and trends in a representative way either on a local or country-wide level. Accurate comparisons between regions, societal sub-groups, over time or even general name trend analyses, are hardly possible using such data.

The "gold standard" database for empirical sociological research on given names would be census data or official central register data. This type of data covers all individuals in a society

and enables researchers to establish links to the respective parents. But due to well-justified data protection reasons, the names and addresses of respondents are only assisting features that have to be separated from the actual census data all over the world.[11] As long as censuses do not explicitly survey the respondents' names, these data will never be available and, in fact, there is no good reason why censuses should survey names. Nonetheless, representative social surveys can be a powerful alternative – if first names are surveyed.

An important feature of household surveys is the opportunity they provide to link parental data to children's names. In the best case scenario, parental data is available for exactly that point in time at which the child was born, which can shed insight into the societal circumstances and influences on the naming act. But, even if the parental information does not precisely match the birth year of a child, survey analyses can still be fruitful. This is because several parental characteristics are time invariant or can be created retrospectively on the basis of approximation.

All in all, well-conducted representative surveys can cover not only the social, but also the naming reality of a society and their linkage, even if – due to the characteristics of a sample – they could never include all name types in a society. Representative surveys of sufficient size are of great importance, as smaller samples of, for example, only 1000 or 2000 respondents miss most name types found in the full population of a particular country.

2.2 The SOEP as an Example for Survey Data Based Socio-Onomastic Research

For our analyses of name frequency distribution patterns we use the SOEP. This is one of the largest household panel studies for a given country worldwide (see Wagner, Burkhauser & Behringer 1993; Wagner, Frick & Schupp 2007). It fully represents the current population in Germany. The SOEP was started in West Germany in 1984. Since the falling of the Berlin Wall, East Germany has been included, creating a very interesting perspective for comparative analysis within Germany. The SOEP additionally includes (and over-represents) the major five groups of labor migrants in Germany and therefore allows for valid analyses of the foreign population in Germany as well. For this article, however, we did not exploit data

[11] Tucker (2001) places great value on access to census data for onomastic research. However, privacy protection is not only technically difficult to assure but also a problem of research ethics, because in most countries the respondents have a right to their names and addresses not being linked to the actual data.

on these groups. Gerhards & Hans (2009), on the other hand, published an excellent article on adoption mechanisms of the name choices of migrants in Germany, using our data base.

From 1984 to 2002 (which is the last survey year we include for the purpose of this article), more than 56,000 persons were surveyed by the SOEP, with more than 4,000 of them each year for the above mentioned period. All of these persons were asked for their given names.[12] Survey-design related imbalances in the data can be suspended by using weights.[13] This is a standard procedure when analyzing survey data. Of course, due to selective mortality, data quality, in terms of perfect representativeness, decreases the older a birth cohort is. For the younger cohorts, where the social selectiveness of early deaths can be ignored, we can claim, in practical terms, total representativeness.

As already stressed, for reasons pertaining to research ethics as well as the strict German privacy protection regulations, first names are not collected from the respondent's street address, but rather explicitly asked for during the SOEP survey interview. Although these names are available in our data basis, special data security measures were implemented in order to avoid the disclosure of respondents' identities (full names and addresses are not included and completely unknown to us). The undertaken security measures include the publishing of results on an aggregate level.[14] Therefore, unique names and those with very

[12] The aim of our research agenda more generally is beyond the scope of this article, as we intend to couple parental characteristics (e.g., social strata, education, occupation, etc.) with name choices, as opposed to solely analyzing name distributions. We are able to create a link between parental information and that of their children for three different cases. The first and most convenient case is when a child is born into a family already surveyed by the SOEP. A second case arises if parents are surveyed for the first time after the birth of their child. Here, information can be obtained directly from the parents about their circumstances at the time of the child's birth. Finally, to obtain information about parents who were not surveyed by SOEP for reasons such as, for example, living elsewhere or having passed away, one may use retrospective questioning methods to interview (adult) children about their parent's situation at the time when the child was 16. Once this information has been collected, because naming is a singular event and changes of given names are not common place in Germany, we are able to recode the longitudinal data structure into one of a cross sectional nature. In doing this, we expanded the time span for naming analyses from today to the early 1900s. For the purposes of this article, however, there is no need to combine any additional information on parental characteristics, such as education, with children's names, since we only focus on name distributions.

[13] This is an important point, because weighting means that, based on the social structural background of the respondents, some individuals and hence names in the sample are given higher frequency values than their occurrence in the actual data base suggests. If, for instance, there are fewer upper class people in the sample than their actual share in a society is, this will have an impact on the outcomes if one studies milieu-typical name frequency distributions. But the influence of those people's choices on the general distribution is underestimated. For Germany, Gerhards (2003) (much like Besnard & Desplaques (2001) for France) found that especially the educated, upper class members of society are more likely to choose new, innovative names. If there are too few of these people in the sample, we would underestimate the "typical" name preferences of this group and therewith the actual occurrence and influence of these innovative names when looking at general distributions. As a result, top ten charts and the skewness of a distribution curve might change.

[14] Files of given names are not available within the standard data basis of the SOEP. They can be retrieved only within the DIW Berlin under special restrictions.

small frequencies in our sample are prohibited from being published.[15] This is, however, not a major problem, as most empirical studies on name frequency distributions either refer to general naming patterns (distributions) or they concentrate on the most common and thereby frequently used names anyway.

One major limitation of our data basis is that we only know the given name that the survey respondent *told* to the interviewer. If a person has more than one given name, which is quite common in Germany (or in the United States where people have "middle names"), we do not cover those additional names.[16]

One particular technical note is of utmost importance: many given names can be written in different ways. For instance, the German name "Klaus" can be written with a "K" or with a "C," but the pronunciation is completely the same. It is quite common in Germany to use several spelling variants of the *same* names (e.g., Claus – Klaus, Marlis – Marlies). Further, special characters, such as "é" in the name "René," are sometimes used, and sometimes they are not. Such artificial variance was suspended in our data base by standardization (see Huschka, Gerhards & Wagner 2005). The standardization procedure was performed by inspecting every single name that is covered by the SOEP.

Also worth noting is that the pronunciation of equally spelled names can differ. For instance, there is no variation in the spelling of "Kathleen," regardless of whether the parents meant the name to have a German or English pronunciation. It is hard to decide whether parents, upon naming their child, had the "same" names in mind. Are these intellectually motivated name variations truly different name types? It is impossible to answer that question with our data because the names in the SOEP are variables given in written format. We have no idea what the pronunciation actually is. This, however, is a problem common to all of the available relevant data sources, be it directories or church registers. Thus, we only can interpret names written in the same manner as being the same name type. As mentioned, for convenience, we have also standardized differences in spelling (in cases of alternative forms of the same

[15] In order to guarantee privacy protection, we do not publish names which are found for less than 10 respondents in the SOEP. We sometimes, for illustrations, will work with examples for rare names, but those names are drawn from the literature and they are not found at all in the SOEP and/or we do not use the original spelling of a SOEP name. We also do not disclose to which group (more than 10 cases, no case/misspelling) our examples of rare names belong to. For linguistic research questions, the real spellings, also of rare names, are very important. To satisfy such demands, it would be possible to implement additional codes for each name that could cover the phonetic features of every name.

[16] This can be a serious limitation if one wants to analyze the "heredity" of given names, which works through the channel of second and third names.

name). By standardizing spelling, as a side effect, we minimize data protection problems, which are most severe in cases of rare name variants.[17]

We faced further problems in dealing with hyphenated names containing two or more parts. Moreover, there are some names in the sample which contain two parts without a hyphen. We cannot be sure that everybody who has more than one name (with or without a hyphen) did fully report it to the interviewers. All in all, we only found a relatively small number of hyphenated name types in the SOEP, which lead us to the assumption that many people did not report all parts of their hyphenated names, but only the part of the name which is used in everyday life. To assure comparability and to reduce variance which cannot be classified as non-artificial, we decided to only use the first part of such names for our calculations. This is because there are good reasons to believe that most probably the first part of a name is the name by which one is usually called.

Table 1 gives an overview of the frequency of occurrence of name types in our database in their original (as collected) and their standardized spelling variants. About every third name type belongs to one of the above described groups of names and was in one way or another standardized.

[17] For special research questions, the original spelling can be analyzed as well. Furthermore, we developed a code system which consists of four codes for each name. The first of our codes refers to the regional localization of a name. What we tried to mirror in this code is the idea about the regional origin of names, which the name givers most probably had in mind. This is basically the country or world region where the name is "typically" used (at the respective point in time when parents actually chose the name). For instance, in our first code we coded "Andy" as "English/American" but Peter, Maximilian and Maria as "German" names even if they are not "German" but "Christian" names. These names have been very common in Germany for a long time. Thus, they are, from a German perspective, commonly seen as "German" names today and as "Christian" only in the second sense. The three other codes refer to the cultural-historical (so-to-say the "true") origin of a name as we find it in dictionaries. It is worth mentioning that we did not code the "meaning" of names, but rather only their cultural heritages. For instance, in our system, the name of one of the three Holy Kings, "Melchior," would be coded as "Hebrew" because this is the "cultural heritage", even if the "meaning" of the two name parts would be "Mäläk" – meaning King – and "-or" – meaning brightness or light. Altogether, we coded up to three variables to account for the fact that some names have several roots. The first of the three codes is the "most recent" cultural origin, followed by the historically "older" and then "oldest" cultural origins (for further details see Huschka et al. 2005). Our coding system is an extended and more detailed version of the classification system which Gerhards (2003, 2005) used. The codes (not the names) will be part of the SOEP data in the future and can thus be analyzed by all interested researchers who come to the DIW Berlin (more codes could be implemented to cover phonetic features of names; see Huschka et al. 2005). The whole SOEP data (without name codes) are available, also in a fully translated and labeled English version, to the scientific community almost for free (an insignificant handling fee and a signed contract are required). Contact: www.diw.de/en/soep

Table 1

Overview of name types in the SOEP – collected vs. standardized spelling

	All	Men (all)	Women (all)	Germans only
Total number of persons in the sample	56,217	27,958	28,259	37,362
Number of name types found, using the original spelling of names	6,658	3,273	3,494	2,685
Numbers of name types found, using the standardized spelling of names	4,762	2,249	2,630	1,724
Share of name types, of which original spelling was standardized: in percentage of original name types	28 %	31%	25%	35%
Number of persons, whose names' spelling was standardized	5,882	3,201	2,681	3,368
Share of name types (original spelling) with hyphens in percentage of original name types	10%	12%	7%	16%
Number of persons with hyphenated names (original spelling)	1,405	984	421	1,048
Number of persons with hyphenated names (standardized spelling)	0	0	0	0

The difference between the total number of name types found and the sum of the numbers of name types found for both women and men gives the number of name types which are used for both sexes: 109 name types when using original (as written down by the interviewers) spelling variants, 117 when using standardized spelling variants.

Source: German Socio Economic Panel Study (1984 – 2002), own calculations

The SOEP covers a huge amount of foreigners as well as people who were born in other countries where different languages are spoken but have become German citizens. Typically, these migrants have different name types that those born into the German populace (the relatively lower share of standardized spelling variants might be due to the interviewers asking for the right spelling of these unknown names). The total number of "countries of origin" in the SOEP is 110. For this paper, we only analyze Germans who have been born in Germany. Thus, our sample size shrinks from approximately 56,000 to 37,362 people.

3 Methodological Possibilities and Problems Pertaining to Analyses of the Frequency Distribution and Concentration of Given Names

3.1 Number of Name Types and the Large Number of Rare Events (LNRE)

If one wants to discuss and compare the (typical) distributions of names' corpora, for example, by calculating distribution curves, frequency spectra, and "characteristic numbers," one must take into account that all measures are sensitive to the size of the corpus, which is in most cases a sample of the population. The "sample size effect" occurs because distributions

of names always include LNRE zones. Thus, sample size issues cannot be ignored in comparative name studies.

Indicators (numbers) that characterize distributions have been (unsystematically) introduced by other name studies (e.g., Tucker 2001), but without reflection upon the sample size dependence. We discuss this problem and its solution not by mathematical considerations, but by the presentation of different empirical findings.

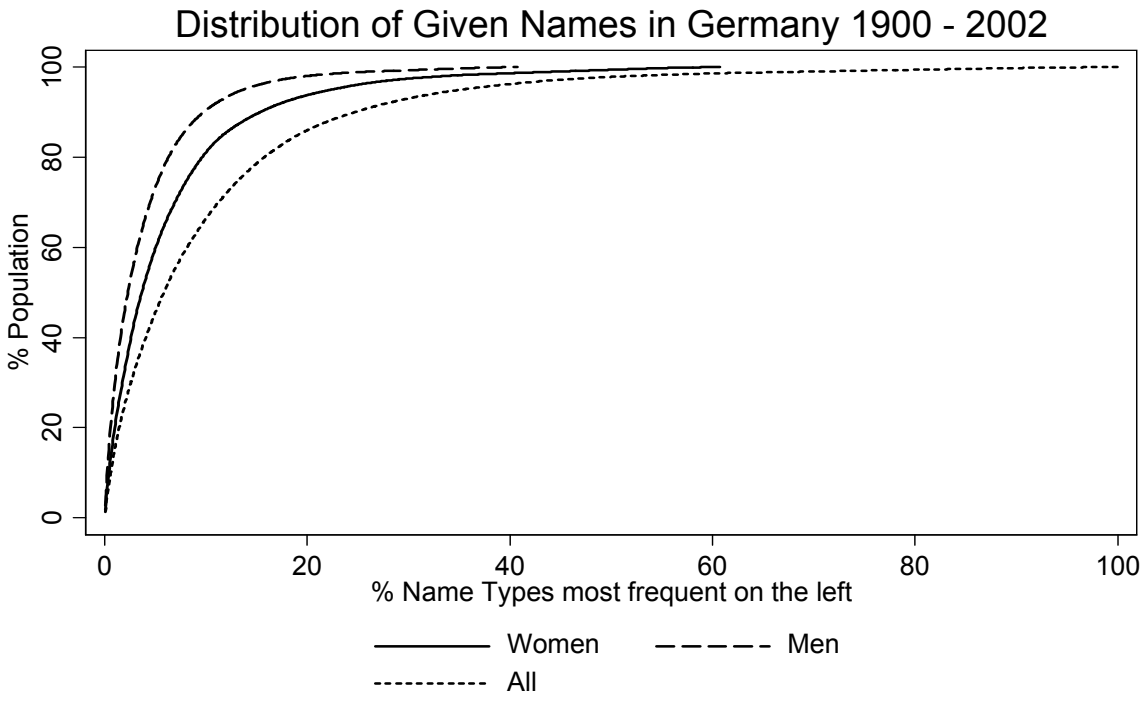
a) General naming patterns: gender specific analyses

To deal with our comparison groups in a statistically correct way, there is a need to use equally sized groups (e.g., men vs. woman) to assure comparability. We will explain later how sample size affects distributions. At this point, we will just mention that the numbers of male vs. female respondents in our data is practically equal, so that valid comparisons of numbers of name types and of distribution curves are possible.

Everybody knows about the phenomenon of very popular names; for example, the phenomenon that classmates or colleagues have the same name. On the other hand, we also often have to ask other people to spell their names if they are very rare and exotic (within a given society). Thus, it is no surprise at all that the general distribution curves (for both sexes and men and women separately) of the names of Germans born between 1900 and 2002, who are the subjects of our study, look as follows (compare Tucker (2001) for this kind of figure).[18]

[18] In economics, the same kind of "cumulative shares" are well known as the "Lorenz Curve." However, the Lorenz Curve is displayed in a different manner; namely, one in which the curves displayed in Figure 1 are "mirrored" along a 45° line. This means the Lorenz Curve does not start at the origin of the graph with the largest shares, but rather begins with the smallest shares.

Figure 1



Source: German Socio Economic Panel Study, own calculations

In Figure 1, on the horizontal axis, we entered the standardized name types in order of descending popularity (beginning with the highest rank at the far left) as percentages of found name types. The respective share of the covered population is shown on the vertical axis. Meeting our expectation to be very uneven, the distribution curve (for all Germans) starts near the zero point and rises rapidly to 70 (percent of population), which is covered by 10 (percent of names) and 90:22, and then slowly to 100:100. The most frequently occurring name in Germany between 1900 and 2002 is "Hans," followed by "Michael" and "Thomas." Statistically, about 1.4 percent of the entire German population included in the sample (including women) share the most popular name "Hans." However, "Hans" is only one name type of 1,724 in our sample, which equals 0.06 percent of the total included names. About 25 percent of all given name types (431 out of 1,724 types) are enough to supply nine out of ten Germans (about 33,625 out of 37,362) in the sample with a given name. The most popular female names are "Maria" (rank 7 of all names, including male names), "Ursula" (rank 10), and "Anna" (rank 15).[19]

[19] All calculations are done using the weighted SOEP data. We also tested the distribution of names using the un-weighted SOEP data, and only very little differences turned out. This shows that SOEP respondents have been drawn on a high quality level and weighting the data is performed only to adjust very little sample-design issues.

The separated distribution curves for men and women are much more rapidly rising than the joint curve, and they already end at 40 percent (for men) and 60 percent (for women) of all name types. This is due to the need to standardize the percentages of name types to the highest number of name types so as to assure the direct comparability of the gradient angles (derivatives). The curve for "All Germans" in the graph above covers all name types for both sexes. Thus, men and women separately cover their respective shares of name types, which in sum equal 100 percent.

If we compare the curves and numbers for men and women, we see that there are more different name types used for girls than for boys. Further, a second comparative result is that the distribution of male names is more concentrated than the distribution of female names. Men seem to be more likely to share names with more fellows, no matter whether the names are popular or rather unpopular name types. This seems to be already clear when looking at the smaller number of names which have been found for men compared to those for women. But one cannot necessarily draw conclusions from this observation. Theoretically, the curve for men could be flatter than the women's one, which would be the case if all male names would be more equally distributed over all male newborns (which is also true if there are fewer name types used for boys than for girls!) than the female names amongst the female newborns (even if there are more female than male name types found!). In that case the distribution curve for men would be (in the extreme case of an absolute equal distribution) a straight line from 0:0 to 40:40.

Computational linguists (see: Baayen, 2001) use a slightly different way of visualizing frequency distributions of words in texts. They introduced the concept of the "grouped frequency distribution" or "frequency spectrum." Expressed as a function, this becomes V(m;N), which represents the number of (name) types with a frequency *m* (number of persons who bear that name type) in a sample of *N* tokens (persons).

Even if these small differences suggest that weighting is not important, the opposite is in fact true. This is especially the case when using smaller sub-samples or if contrasting societal sub-groups which might not be represented in a sample with adequate shares. In contrast, telephone books, for example, cannot be as representative as even the un-weighted SOEP.

Table 2

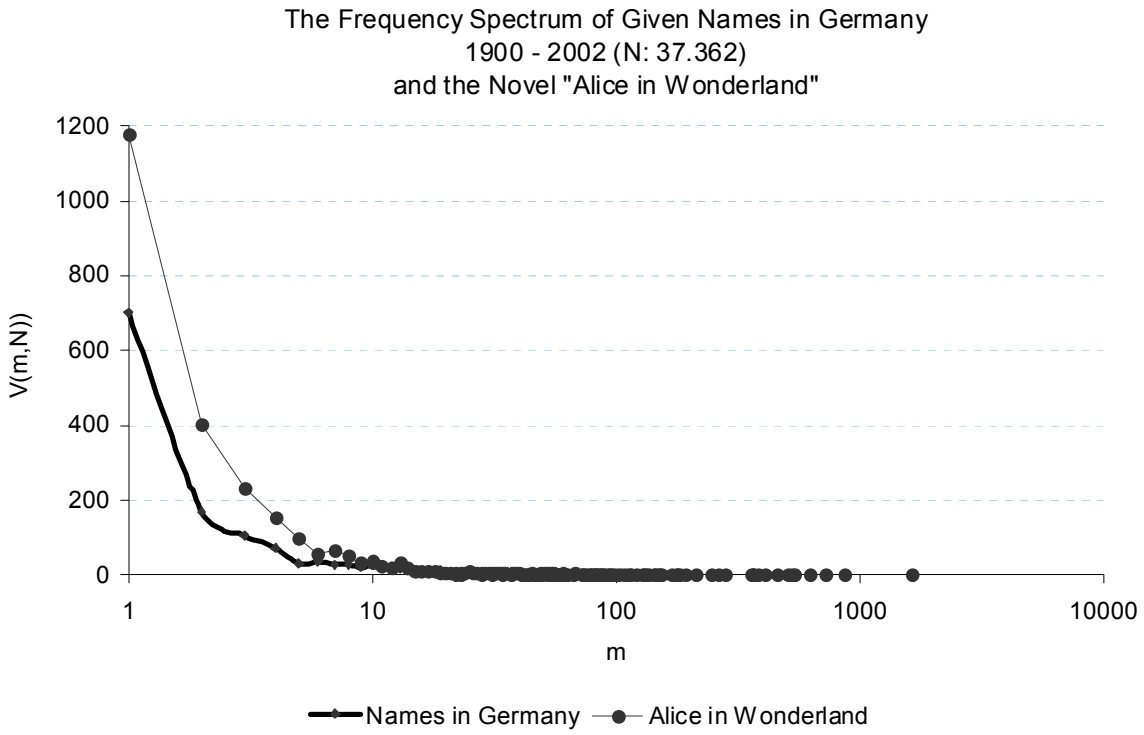
The frequency spectrum V(m;N) of Given Names in Germany 1900 – 2002 (N: 37,362)

M	V(m;N)	m	V(m;N)	m	V(m;N)	m	V(m;N)	m	V(m;N)	m	V(m;N)	m	V(m;N)	m	V(m;N)	m	V(m;N)
1	702	21	6	41	2	61	1	84	3	107	1	132	1	180	2	253	2
2	168	22	7	42	2	62	5	85	3	108	1	137	2	181	1	257	1
3	106	23	7	43	4	65	2	86	2	109	1	138	1	183	1	259	1
4	72	24	6	44	6	66	1	87	5	110	1	139	1	184	2	271	1
5	30	25	5	45	4	67	1	88	2	112	1	140	1	186	1	290	1
6	37	26	7	46	3	68	2	89	3	113	1	141	1	188	1	294	1
7	27	27	7	47	4	69	1	90	2	114	1	142	1	193	1	302	1
8	28	28	5	48	3	70	3	92	2	115	2	144	1	194	1	310	2
9	21	29	4	49	3	71	1	93	1	116	1	147	1	195	1	323	1
10	26	30	3	50	7	72	3	94	1	117	4	149	2	205	1	361	2
11	19	31	6	51	2	73	6	96	1	118	2	150	1	213	1	364	1
12	22	32	5	52	2	74	3	97	2	119	2	151	3	219	2	395	1
13	16	33	3	53	3	76	2	98	1	121	1	152	1	223	1	415	1
14	12	34	5	54	2	77	1	99	1	122	2	154	2	227	1	521	1
15	14	35	6	55	4	78	4	100	2	124	2	156	1	232	1		
16	11	36	7	56	1	79	2	101	2	125	2	158	2	234	2		
17	14	37	7	57	8	80	1	102	2	126	2	163	1	235	1		
18	12	38	4	58	2	81	1	104	1	128	1	164	1	238	1		
19	15	39	2	59	5	82	1	105	3	130	1	166	2	241	1		
20	9	40	4	60	4	83	3	106	3	131	3	179	1	245	1		

Source: German Socio Economic Panel Study (1984-2002), own calculations

Table 2 shows that the most popular name (Hans) occurs 521 times (the last number in the table above) in a sample of 37,362 individuals, whereas there are 702 names found only once in our sample (the very first number in the table above). In linguistic text analyses, such uniquely occurring words would be called “hapax legomena.” This term stems from the Greek words *hapax*, meaning “once,” and *legomenon*, meaning “read.” Plotted in a graph (with logarithmic scale on the horizontal axis), the frequency spectrum of names in Germany has a similar shape to the frequency spectrum of texts (e.g., the novel “Alice in Wonderland,” which was analyzed by Baayen, 2001: 9 –10) and looks as follows (Figure 2):

Figure 2



Source: German Socio Economic Panel Study (1984-2002), own calculations, numbers for "Alice in Wonderland" taken from: Baayen 2001, pp. 9f.

b) Problems with comparing name frequency distribution curves: number of name types and LNRE.

Our name frequency-rank distribution curves are similar to those drawn for other societies. Tucker (2001) and Eshel (2001) have published similar curves (without gender division), comparing the US and Canada (Tucker) and the US, Canada, and Israel (Eshel), using telephone directory data. But even if the overall patterns in all of these graphs look quite similar to ours, there are in fact huge differences: Tucker describes that for the US only about 1 percent of the most popular given name types in the sample accommodate for approximately 95 percent of the population (Tucker, 2001: 75). In our data, the respective share of people accommodated by 1 percent of the most popular names is about 16 percent. About 35 percent of the name types are necessary to supply 95 percent of the Germans in our sample with given names. Plotted in a graph, that means that Tucker's curve rises much

steeper than ours.[20] Does that mean that Americans are much more likely to be named with a very popular name, and on average significantly more persons (relative to population size) share names, no matter whether these names are very popular or not? The answer is "no" and the reasons for it are different sample sizes in Tucker's and our study.

As mentioned above, we would have to standardize the number of names describing each of the distribution curves to make accurate comparisons of the gradient angles possible. But this alone is not enough: not only the number of different names matters, but also the sample size (the number of persons in a sample) matters due to the different numbers of names – and therewith possible ranks – covered in differently sized samples. A sample of the size N persons covers a number of X names. X is always smaller than N (only in the extreme case of unique names for all persons X could be equal N). And we know that the N persons distribute highly unevenly over the names X. Further, the number of name types X does not linearly rise with a rise of the number of persons N. A sample of N=100 might contain X=80 names. This does not imply that a sample of N=1000 contains X=800, as this could contain, for example, X=500 name types (this is highly plausible since X is smaller than N and the distribution of the persons over name types is very uneven!). Thus, the larger a sample already is, the higher the increase in N would have to be in order to add one more name type X. And those additional people in larger samples do not evenly distribute over the different names; instead they have a higher probability of having one of the more popular names. As a result, the concentration of names gets larger with increasing N. Therefore, we have to keep in mind the fact that different sizes of samples produce both different numbers of name types and thus incomparable distributions of people over these names. Concentration indices and concentration curves rely on calculations that comprise both the overall number of names (ranks) and the respective numbers of people covered by those names, each as shares of the *respective* totals. If these totals are not comparable (equal), the curves or measures are not comparable. Hence, all conclusions made regarding names in the literature must be evaluated and discussed with respect to sample size. Comparisons of name distribution patterns are valid only if the samples are equally sized (and additionally assure an equal selection probability of the covered people).

Statisticians describe the distributional behavior of names as a phenomenon called "Large Numbers of Rare Events" (LNRE) (see: Baayen, 2001: 51). This is also observable in texts

[20] Tucker's curve for Canada is lower than his curve for the US, but would nonetheless still be more steadily rising than the German one.

when analyzing the frequency word distribution. This implies that if one wishes to compare texts or name corpora with different "lengths," which refer to different numbers of "events" (word types in text corpus analyses; in our case, names) in a statistically valid way, one must (1) compare equally sized corpora (texts, samples of persons) or must (2) correct estimated frequencies and sample characteristics by appropriate mathematical methods. The correction of characteristic numbers is a "science of its own," the relevant mathematical corrections required for which rely on assumptions (for an excellent overview see: Baayen 2001). Researchers in computational linguistics are, however, accustomed to using assumption-based corrections because the alternative – standardizing sample sizes – is difficult to apply to texts. Standardization, as mentioned in (1) above, means that one would draw sub-samples out of the larger texts under consideration (the shortest text determines the level of standardization, as text lengths must be of equal size). But drawing sub-samples of texts is problematic because texts consist of series of passages. Indeed, certain parts of a text have certain textual functions. With respect to these functions, the frequencies of certain words can be highly different. And one cannot just shorten a text, for instance, by omitting the introduction. Such a procedure would be comparable to, for example, deleting all people with a university degree from a representative social survey. This would certainly change results.

However, drawing *real* random sub-samples of persons (and names) from surveys is no problem at all, because the probability of each person (and therewith of each name type) to be drawn is equal. Social sciences are based on representative samples of populations, and empirical social scientists are experienced in dealing with the limitations of samples and their usefulness in representing a society. Therefore, for onomastic research to make valid comparisons of name distributions, we propose that random draws be taken out of all those samples larger than the smallest sample, so that all samples that are to be compared are made to be of equal size.

To illustrate, let us (again) provide a simple example. If we want to analyze the degrees of naming individualization in West and East Germany (during the time of separation), we would compare East and West German name distributions. There are some reasons to believe that the degree of individualization in West Germany was higher, also in naming, than it was in East Germany, due to East Germany's uniformity-oriented official ideology (see: Huschka et al., 2009).

A first hint for different degrees of individualization in naming is given by a look at the numbers of name types found for the two parts of Germany. All in all, we have 906 different names given to 11,615 persons born between 1950 and 1989 in the "old" Federal Republic of Germany (West Germany), and 678 different name types given to 5,678 persons born in the German Democratic Republic (GDR – East Germany) in our data. Exclusively looking at the numbers of found name types suggests that our assumption of a higher degree of individualization in West Germany was right. But if we calculate the ratio of population divided by the number of name types as a proxy (as a very simple "characteristic number") to describe the distribution of names, the result is that on average 8.4 East Germans share a name and 12.8 West Germans have a common name. Statistically seen, this means that more often newborns share the same names in our West German sample and the degree of individualization there was actually lower. But West Germany (and also our sample for West Germany) is much bigger than East Germany (and the East German sample). Taking these facts into consideration, an East German newborn statistically shares its name with 0.14 percent of the East Germans, whereas a West German newborn shares its name with only 0.11 percent of its fellow West Germans.

This little exercise demonstrates how misleading numbers can be when comparing samples of different sizes. As mentioned above, an effective way to handle the problem of incomparability in a statistically correct manner that does not rely on non-testable assumptions is to reduce the sample size of West Germany to that of the East Germany. This is done by random sampling and calculating the measures again. In order to improve the statistical robustness of our results, we not only draw one but 30 random "West sub-samples." The differences between the 30 sub-samples give us a hint about the precision of our calculations.[21]

What are the results? The mean number of names types found for 30 random West German samples (of the East German size) is 687 and therewith only slightly higher than in the East German sample (678). Also, the ratio of population over name types for West Germany, using the reduced West sample, is nearly exactly the same as that of East Germany. The original West German population-to-name type ratio was calculated using starkly different sample sizes for East and West Germany. Thus, it had to be corrected from 12.8 (see above) to 8.3, which is now as high as the East German one (8.4). A difference of 0.1 is insignificant. The

[21] To adequately ensure statistical correctness, we drew 1,000 random sub-samples for some calculations. Even for powerful modern computers, this is a very time consuming procedure. As it turns out, the 30 sub-samples produce statistically robust results.

same procedure of random sampling has to be applied when examining the distribution curves of names in East and West Germany in order to compare their gradient angles (see: Huschka et al., 2009).

Unfortunately, we have found incorrect graphical presentations in the "naming literature" in which neither the numbers of found names have been standardized nor have the sample sizes been, at least approximately, equal, (e.g. in Tucker, 2001, and Eshel, 2001). Figure 3, which compares the distribution curves for names given to East vs. West German newborns between 1950 and 1990, demonstrates how results based on such biased statistics can mislead researchers to drawing wrong conclusions. The lower line for East Germany suggests that naming was more individualized in East Germany. We cannot, however, draw conclusions about the numbers of found name types in the two parts since the horizontal axis is labeled as "percentages of name types" (a common method of scaling in onomastics).

Utilizing exactly the same data used in Figure 3, but making sample sizes equal via random sampling, the results in Figure 3a turn out to be quite different: there are no observable variations in the distribution as the lines are overlapping. This indicates that names were similarly distributed amongst East and West Germans (we do not, however, claim that East and West Germans used the same name types; for this see: Huschka et al., 2009). The sample for West Germany was randomly downsized by using the mean of the lower and upper bounds of 30 random draws. After equalizing the sample sizes, as mentioned above, the numbers of names found between East and West no longer differ. Thus, the ending of both lines at the 100 percent mark is appropriate.

Figure 3:

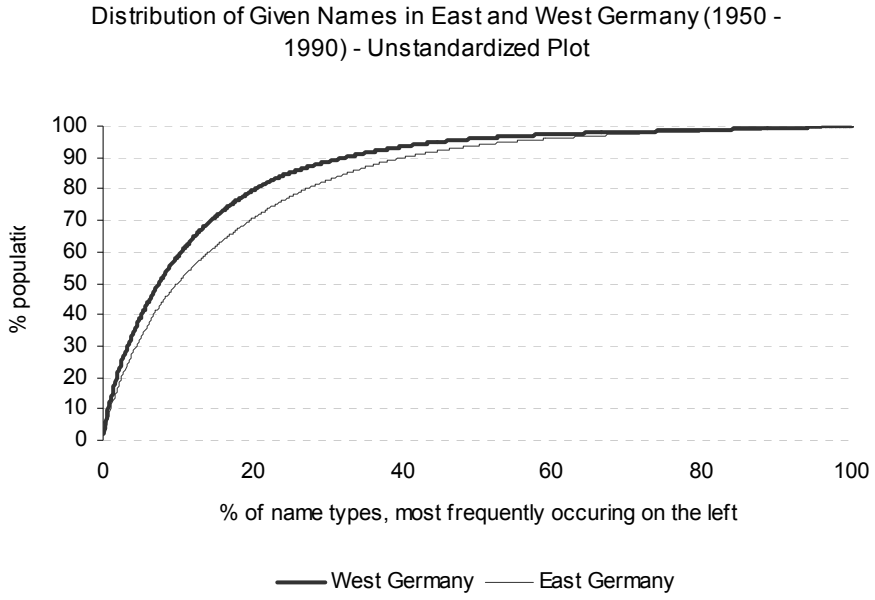
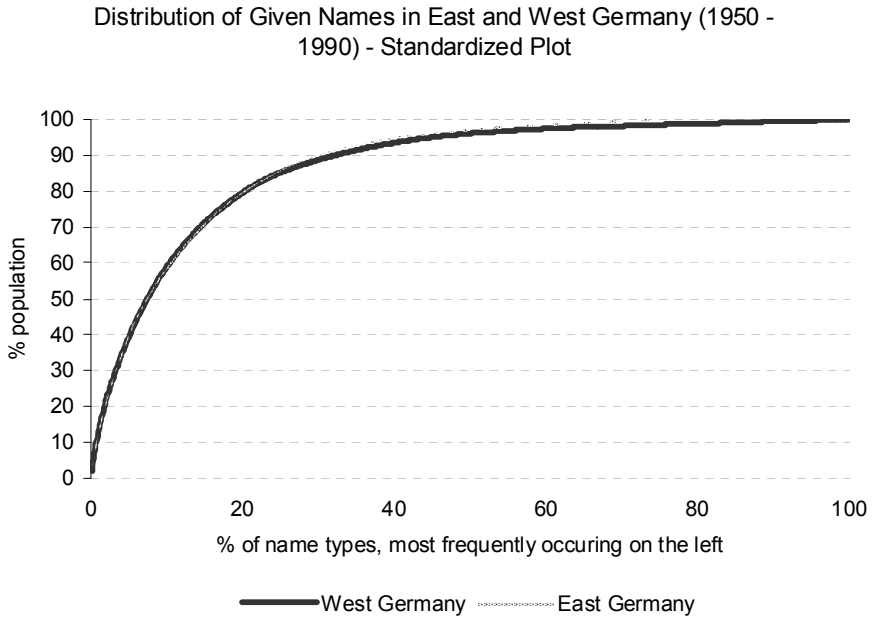


Figure 3a:



Source: German Socio Economic Panel Study, own calculations

We can conclude, in line with the literature on text corpora,[22] that names and their distribution in samples are not easy to compare in a meaningful way. This is due to the fact that differences in sample sizes between comparison groups produce wrong results if one is not

[22] See Baayen (2001: 2) for a general discussion about text corpus.

careful. That does not mean, however, that inference is impossible on the basis of samples. What it does mean is that comparisons of frequency distributions are only valid if the samples are of equal size.[23] We will come back to this point again when we analyze different birth cohorts for Germany.

3.2 A Note on Statistical Significance

For further analysis, we compared the name distributions of 10 birth cohorts in Germany. Of course, we computed the cohorts in a way that assures comparability, taking into account the sample size problem.[24] As expected, we found clearly visible *real* differences between the numbers of names and their distribution patterns over time. From a statistical point of view, this is, however, not the end of the story, as sample-based calculations generally do not deliver sharp point estimates. The difficulty associated with blunt statistical inferences of *all* estimators (for example Top 10 shares, etc.) due to sampling error is a methodological problem that always occurs when using samples, even large ones.

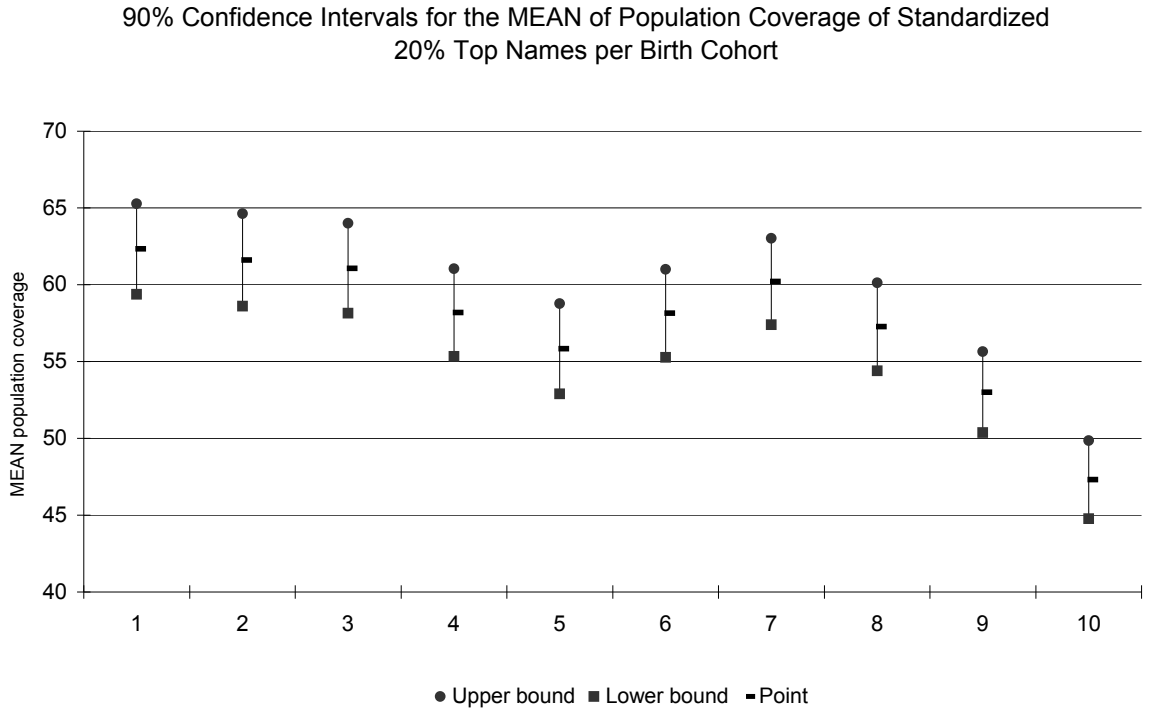
"Sampling error" means that, due to the process that generates a sample of a larger "universe" (e.g., *all* citizens in a country), a sample cannot be an exact mirror of this universe. This problem can be demonstrated with a simple experiment. Suppose there exist 1,000 balls, of which 500 are white and 500 are red. If drawing ten balls ten times, one will get 50 white and 50 red balls *on average*. But there will still be slight differences in the results between each of the ten draws. By increasing sample size to, for example, ten times 100 balls, one would get better results (in the sense that the 50-50 split reflecting the probability of drawing a red or white ball is more evident), but this is only proven if one would indeed draw all 1,000 balls. 1,000 balls could be seen as analogous to a census, which covers a whole population. Thus, when relying on samples (and even telephone books are nothing else than samples), one has to take the sampling error into account to produce accurate results, especially when comparing sub-groups of samples.

[23] This methodological problem is of great importance when comparing countries, birth cohorts or different points in time. One can try to overcome the problem of sampling by simulation measures (see Baayen, 2001, chaps. 3-6). In the case of names, making sample sizes equal by randomized sampling is deemed a robust method. But even then, one should take care to consider issues of equal sampling probabilities of the people in the samples. In this regard, a *representative* survey has advantages again.

[24] See section 4 of this article for a description of how we computed the birth cohorts.

In our analysis, it will be revealed that point estimates cannot always be separated clearly from each other (such a point estimate is for instance the share of population covered by a certain percentage of name types). The robustness of such point estimates, on the other hand, can be tested by calculating so called "confidence intervals" or "confidence bands" for each of the point estimates. Confidence bands are, said simply, the "corridors" of point estimates (defined by lower and upper bounds), which cover the *real* (*real* means *true* – not only true for the sample, but true for the whole population the sample was drawn from) point estimates in 90 percent of all imaginary confidence band calculations. These calculations are based on thousands of sub-samples taken from the sample. If neighboring confidence bands overlap for different cohorts, there is no strong statistical evidence for differentiation between the calculated point estimates.

Figure 4



Confidence Intervals with a 90 % significance level.

Source: German Socio-Economic Panel Study, own calculations

Figure 4 offers the point estimates and confidence bands[25] for the average shares of people covered by the 20 percent of the most popular names in each birth cohort (percentages on the vertical axis), using equally sized birth cohorts. As we can see, there are some differences between the upper and lower bounds of the point estimates within cohorts, mainly in a range between 5 percent and 6 percent. More important, though, is that the lower bounds of the older cohorts (1-7) are clearly crossing each others' edges with the upper bounds of some of the neighboring cohorts. Thus, we cannot claim for some neighboring cohorts that our point estimates *really* statistically differ from each other over time.

Only if a cohort's lower bound is clearly higher than the next cohort's upper bound is there solid statistical evidence that the differences between the point estimates are fully robust. For the younger cohorts, these differences are significant with a 10 percent error level. Moreover, comparing, for example, cohort 10 with cohort 7, there are the greatest chances that there are clear differences between point estimates. In sociological terms, this means there is a real change in naming behavior amongst the parents of these cohorts.

3.3 More "Characteristic" Measures

When reviewing the literature in search of meaningful indices or numbers to describe (and compare) name distributions more easily (for example, using a single number), we found that not only linguists, but also economists, provide useful indicators when analyzing market shares of firms. Thus, we ought to adopt several of these indices that help to characterize the distribution of a variable.[26]

It is absolutely noteworthy to mention that sample size *always* matters. There is no index or measure which is completely unaffected by sample size differences, especially of distributions with large LNRE zones. When dealing with names, even more caution is needed when drawing conclusions. Even when comparing equally sized comparison groups (e.g., 1,000

[25] Our confidence bands are calculated using the textbook formula for random samples. Because the SOEP's sample is not a plain random, but a stratified and clustered sample, for statistically perfect confidence band calculations we would have to use the "bootstrap" approach or at least the 8 so called "random groups" of the SOEP data. These procedures are difficult and time consuming (in terms of computer time) to apply, especially when dealing with the very special distribution behavior of names within a sample. Nonetheless, it is our intention to apply this method in further research. The much easier text book confidence interval approach is perhaps not perfect, but there is evidence from experience that these confidence bands are very similar to the ones that would prevail if more sophisticated approaches were employed.

[26] See Baayen (2001); and for examples from economics, e.g., Scherer & Ross (1990).

men vs. 1,000 women, or birth cohorts), we are quite likely to find unequal numbers of name types within these comparison groups. Of course, more name types (e.g., over time) can be explained as a real difference, indicating individualization in a society. But as brought up earlier, more name types can also be found in less individualized distributions if the biggest share of the names is only marginally used and the majority of people stick to a few very popular names. In other words, only looking at the numbers of name types measures the extent to which a "value" of a variable represents individualization, but this says nothing about the distribution of these values (names).

In Tucker's 2001 study, he provided readers with some typical shares to describe his names distribution. These included, for example, the share of people, covered by a certain number of name types (ordered by their popularity). Similar "shares" are used by economists for the analysis of the concentration of markets. In the case of name distributions, we certainly deal with the "market shares" of names and their "market power," which could be used to describe their popularity. In Table 3, we present "name shares" (percentages of people covered by a certain number of top name types) in a manner analogous to the way economic literature portrays the market power of the "largest four," "largest eight," "largest twenty," and "largest fifty" firms in a market (in our case, names in a society). We also display the Hirschman-Herfindahl Index, which will be explained shortly.

Table 3

Concentration Ratios for Names (and, for comparison, for representative industries)

	Number of name types / firms	Largest four	Largest eight	Largest twenty	Largest fifty	Herfindahl Hirschman Index
All Germans	1,724	4.6	8.3	17.1	32.4	35
German Men	703	9.5	16.8	33.1	56.7	89
German Women	1,047	6.1	11.0	22.5	43.2	53
For comparison:						
Screw machine products*	1,744	8.0	11.0			30
Semiconductors*	685	40.0	57.0			597
Paints and allied products*	1,170	24.0	36.0			222

* Data for the firms from the US Bureau of the Census, 1982 Census of Manufacturers, "Concentration Ratios in Manufacturing" MC82-S-7, taken from: Table 3.6: In: Scherer/Ross 1990, p. 77. Unfortunately no numbers for the columns "Largest twenty" and "Largest fifty" are given.

Source: German Socio Economic Panel Study (1984-2002), own calculations.

From Table 3, we can see that when looking at "All Germans" (men and women), the Top 4 name types cover 4.6 percent of the population (given our sample size), the Top 8 names 8.3 percent, the Top 20 about 17 percent and the Top 50 names almost every third German. This means that only 50 out of 1,724 name types are enough to supply about 12,000 out of 38,000 people in the sample with a name.

Comparing men and women, we find that the Top 4, 8, 20, and 50 male name types always cover a larger share of men than the respective female name types do. As shown above, the number of name types found in a sample depends on the number of observations (sample size). In our case, the sample sizes for the sub-samples of men and women are almost equal. Thus, we can argue that, based on our representative sample (which ensures approximately equal selection probabilities), there are more female names in use than male names and, on average, women share their names with fewer other women in Germany than men do. In sociological terms, women's names are more individualized in terms of both a higher number of different value types (name types) and a more equal distribution of these value types.

It is also known from the economics literature that the numerical values of such characteristic indices depend heavily on the numbers of firms which are operating in a particular market (in our case, the number of names in the sample).
For the sake of illustration, if we compare our results for names with the results for market power, we must try to find markets where the numbers of firms is approximately equal to the numbers of names in our sample. In the lower rows of Table 3 we thus added some market segments of the US economy that consist of a similar number of firms to the number of name types in our data. What we learn from these comparisons is that names are more evenly distributed than market shares in comparable economic markets.

In order to summarize the full distribution of market shares with only one "characteristic" number, the so called Herfindahl-Hirschman Index (Scherer & Ross, 1990) is popular amongst economists. This index summarizes the degree of concentration of firms in a market (and in our case, of names) in one number. The index values are reported in the right column

of Table 3. The HHI[27] is defined as the sum of all squared (market) shares[28]:

$$\sum_{i=1}^{K} S_i^2$$

The maximum value of the HHI is 10,000 (the square of 100 percent), indicating a monopoly, an undesirable degree of market power (in the case of names it would be undesirable as well because one name would not distinguish one person from another).

Our results for the HHI applied to names given to German newborns point in the same direction as the market share results: the distribution of male name types is more concentrated than that of women's names, and almost all economic markets that we use as illustrations are more concentrated than the "market for given names."[29]

At this point, we want to mention two more, quite old, "characteristic numbers" that have been suggested as "text constants" by linguists. Yule's K (Yule, 1944) and Simpson's D (Simpson, 1949) can be seen as "weighted average probabilities" or the "repeat rate" of words/names (see Baayen, 2001: 25).[30]

$$K = 10000 \frac{\sum_m m^2 V(m,N) - N}{N^2} \qquad D = \sum_m V(m,N) \frac{m}{N} \cdot \frac{m-1}{N-1}$$

The measures K and D are both based on the same approach as the HHI and thus, their results should conform (some features in the formulas are just adjustments of scales). Also, K can range from 0 (equipartition) to 10,000 (maximum concentration: monopoly), whereas D ranges from 0 to 1.

For our name distribution, the values for Yule's K are 35.2 (all), 53.4 (for women) and 88.9 (for men). For comparison, the distribution of the word types in "Alice in Wonderland" has a K of about 102. What we confirm with these results is that male names are more concentrated

[27] For an extended description see Hirschman (1964).
[28] Again, the number of firms and name types matters. Adelman (1969) states that the HHI falls "monotonically but nonlinearly with an increasing number of firms" (cited in: Scherer & Ross 1990: 73). For a more recent discussion of the HHI see also Kelly (1981).
[29] In fact, the market for "screw machine products" has a lower HHI value despite having higher market share values for the "largest four" and "largest eight." This means that in this market, the largest firms are large in comparison to all others but beyond the largest firms, the market is quite balanced. Numerous other similar concentration measures have also been proposed (see below). They are all more or less sensitive to the LNRE zone. The market shares only concentrate on the largest firms/top names, while the HHI gives more weight to the distribution of the smaller firms/more rare names.
[30] There are a few more proposed characteristic numbers, all with certain advantages and/or limitations. See: Baayen (2001, ch. 1.4).

than female names, and that names are less concentrated than the word types in Lewis Caroll's novel.[31]

4 An Empirical Example: the Distribution of Names in Germany 1900 – 2002

We now turn to a more in-depth analysis of the distribution of given names in Germany. How does the picture look if one analyzes the population differentiated by sex and over time? How do sociological and socio-onomastical conclusions look like?

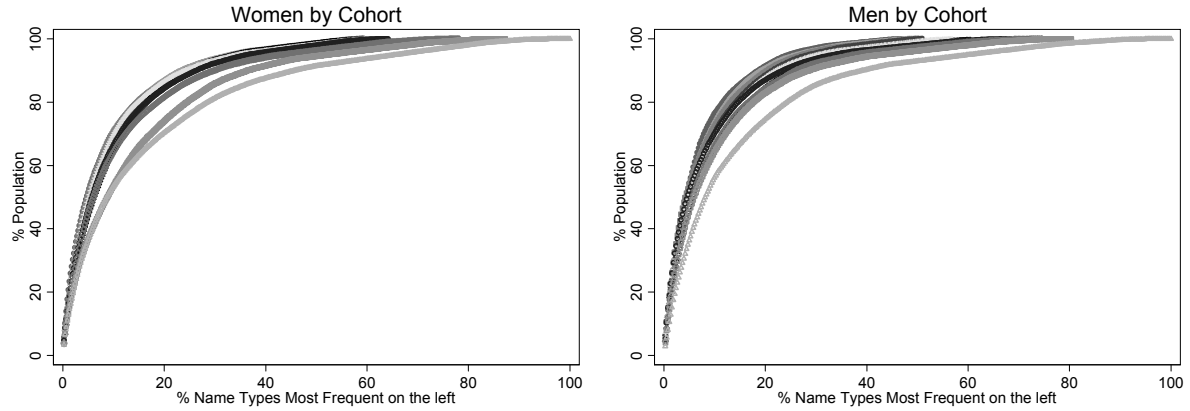
There are real differences in the naming patterns between men and women in Germany. We learned from analyzing our data that there is a greater variety of different name types used for females; a fact that can be explained by a greater degree of "individualization" (and innovativeness) when it comes to naming a girl. Figure 5 highlights graphically the distribution of names amongst German men and women in 10 birth cohorts to make visible the changes in naming over time.

The problem is, again, that birth cohorts can have different sample sizes and therefore do not correctly cover real naming differences. We are going to demonstrate the problem of unequal sample sizes once more. As we learned above, the curve's gradients are influenced by the sizes of the samples as well as the numbers of found name types. Thus, we have to standardize the sample size of each cohort to learn about any change in the concentration of name distributions over time. There are two ways to adjust the sample size in cohorts. The first would be to draw random sub-samples of each birth cohort (which usually covers 10 years) on the basis of the lowest sample size of all cohorts. But in doing so, we lose data and hence weaken statistical efficiency. The other option is to simply use the percentiles (as quasi-cohorts) of the whole sample, which is ordered by the birth years. By doing this, we get 10 equally sized sub-samples that allow for comparison. As an inconvenient byproduct of this form of standardization, we unfortunately lose the nice and even 10-yearly steps that usually define birth cohorts, and some respondents with same birth-years are even randomly pushed into different birth cohorts (for example, 194 persons with birth year 1927 are pushed into the lowest birth-percentile, and 102 persons with the same birth year are put into the second

[31] From a pure statistical point of view, Yule's K, Simpson's D or the HHI cannot be calculated for the entire German population because men in Germany cannot be given female names and vice versa. Thus, there is no access to about half of the names for about half of the individuals.

percentile).[32] In our standardized equal-sized sub-samples (birth cohorts), the covered range of calendar years now spans between 6 and 12 years, depending on the birth years of the respondent.[33]

Figure 5: Name distribution curves of 10 birth cohorts with equal sample sizes



Source: German Socio-Economic Panel Study, own calculations

What we can see in Figure 5 is that over time there are different distribution patterns. As indicated by the early ending of some of the lines, we have on the one hand different numbers of name types found for each of the cohorts covering the same number of people. On the other hand, the gradient angles are different. From a sociological point of view (given the archetypal skewness of name distribution curves), a flatter curve also means more individualization in naming. This is because a flatter curve stands for a less concentrated distribution of the names used to name the same number of children. A very steeply rising curve indicates a high concentration of names, which implies that a few names have been extremely popular amongst the largest shares of newborns, whereas a considerable number of name types has been given to very few children only. The opposite (equipartition) is a straight line – for example, if everybody in a sample would have a unique name.

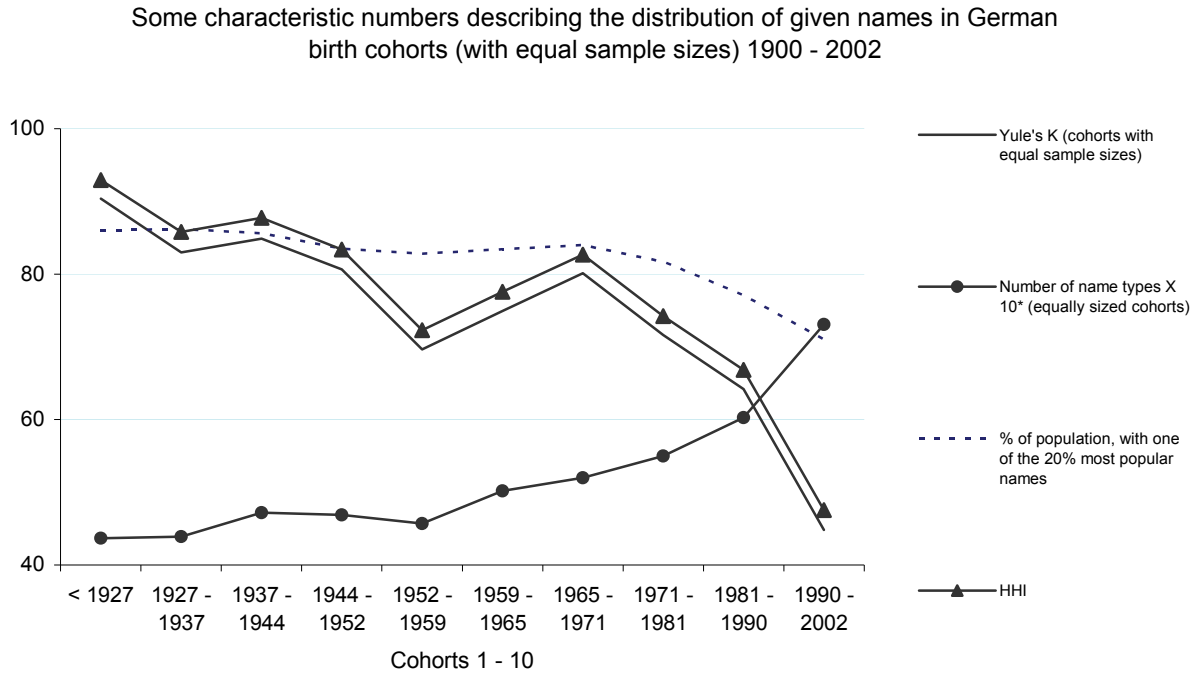
We found it quite hard to "translate" the distributional behavior and comparisons of names into sociological evidence since one always has to consider two things; namely, (a) the

[32] In proper sociological analysis, both the division of people born in the same year into two birth cohorts and too greater differences in the span of years covered by each cohort would be inappropriate. These artificial cohorts cannot effectively cover differing social realities in different cohorts in a meaningful manner. Nonetheless, the main point here is that one has to compare equal samples of equal size.

[33] The SOEP is a high quality data base. Thus, the different reproduction rates in Germany are more or less mirrored in the different numbers of persons born in different years.

number of names and (b) the distribution of these names amongst the population. It is not enough to simply look at either the number of names or their distributions in isolation. However, even offering a caption for all lines in Figure 5 would not make comparative conclusions based on the distributional line's end and gradient angles any easier. Thus, we find that the use of characteristic numbers, such as market shares, the HHI or Yule's K, has an advantage. With one, easy to handle number, these measures characterize a name distribution that considers the whole distribution and not only the most popular names.

Figure 6:



* The number of found name types was divided by ten for reasons of visualization.

Source: German Socio-Economic Panel Study, own calculations

Figure 6 displays concentration measures which have been calculated for ten birth cohorts of Germans (each with the same number of newborns) as well as the numbers of name types found within these cohorts (divided by 10 for graphical reasons). Additionally, we added the share of newborns with one of the top 20 percent of the most popular names (serving as a market share measure). What we learn from this is that, over time, the naming behavior changed not only because people use more different name types nowadays, but also because a

more equal distribution of these names becomes observable (as illustrated by the K line in Figure 6). Notably, the HHI line follows exactly the same pattern as the K line (the mathematical approach is similar).

The development of our concentration measures over time is complemented by the rising numbers of found names in each cohort. But these two developments in naming did not always happen simultaneously: the increase in the numbers of found name types is by and large linear. In contrast, aside from the oldest birth cohort, the concentration of names has two peaks: in the third cohort (which covers the time of the Hitler regime) and during the mid-1960s. The concentration of names has been at its lowest level since the 1980s, at which point one notices a strong tendency towards diffusion (even if the distribution is far away from an equipartition). If one would only look at the percentages of people covered by a certain percentage of names (in Figure 6 we use the most popular 20 percent of names), one can see a similar general development but with smaller differences between the cohorts. At any rate, the deflections of this line are relatively small and the share of babies with very popular names seems to be quite stable. Notably, the most popular 20 percent of names encompass greater variety over time because the overall number of found names has risen.
We advise against taking measures that mainly concentrate on a certain percentage of top names to characterize certain distributions, since such measures are not sharp. It is important to study the developments of all segments of a distribution curve, not only for sociological analysis of individualization or the diffusion of name distributions. This point also highlights the usefulness of concentration measures like Yule's K, Simpson's D, and the HH-Index; they can reduce an entire concentration curve to a single number. In contrast, just by looking at the shares of children who received one of the top names, one is restricted to considering only certain points of a distribution, which nobody can legitimately claim as being "the right" or "most meaningful" points to be taken and compared.

5 Conclusions

On a general methodological level, we demonstrated the crucial importance of carefully considering sample quality and size in empirical onomastic research. Additionally, we showed that large representative sample surveys can be attractive alternatives to telephone directories and church registers.

In our methodological section, we applied methods commonly used in empirical onomastic research to our data (frequency distributions and shares of individuals with popular names). Additionally, we applied methods borrowed from the disciplines of economics and computational linguistics, namely the HH-Index, the frequency spectrum and the K and D indices. More importantly, however, we demonstrated the systematic effect that sample sizes have on all the indices calculated. Thus, one always has to consider sample size when comparing the name distributions of countries, societal sub-groups or of populations over time. We briefly summarized the literature on the "Large Numbers of Rare Events" (LNRE) problem, which has a similar effect on name distributions. We proposed a standardization of sample sizes as an easy-to-apply solution in order to overcome sample size variations in onomastics. Moreover, we provided an example of the importance of calculating confidence intervals when analyzing samples, as opposed to full populations.

Using data from the German Socio-Economic Panel Study (SOEP), we showed that given name distributions in Germany typically follow patterns akin to those presented by other authors for other countries and cultures. Moreover, we found differences in the name distributions between female and male given names. Name choices for male babies have been (relatively) more oriented towards conformity; men are more likely to have one of the very popular names than women. In addition, the total number of names used for male babies is lower than the number of female name types.

We also showed that there exist changes in typical name distribution patterns over time. Today, the likelihood of being named with one of the very popular names is relatively lower than it was at the beginning of the 20th century. For both sexes, there are a significantly greater numbers of names in use today compared to former times. Economists state a reason for this as being the growth of the amount of information disseminated by modern information technology. A byproduct of this is that people are exposed to more appealing new name types.

References

Adelman, Morris Albert (1969). Comment on the "H" concentration measure as a numbers-equivalent. Review of Economics and Statistics 51(1): 99-101.

Arai, Mahmood, Skogman, Peter & Thoursie, Anna (2004). Changing family names: Discrimination or assimilation. Unpublished paper presented at the Harvard University Inequality Summer Institute.

Aura, Saku & Hess, Gregory Dawson (2004). What's in a name? *CESifo Working Paper*: 1190.

Baayen, Harald R. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Bertrand, Marianne & Mullainathan, Sendhil (2003). Are Emily and Greg more employable than Latisha and Jamal? A filed experiment evidence on labour market discrimination. *NBER Working Paper:* 9873.

Besnard, Philippe & Desplaques, Guy (2001). Temporal stratification of taste. *Revue Francaise de Sociologie* 42(supplement): 65-77.

Eshel, Amram (2001). On the frequency distribution of first names. *Names* 49(1): 55-60.

Fryer, Roland Gerhard & Levitt, Steven David Jr. (2004). The causes and consequences of distinctively black names. *Quarterly Journal of Economics* CXIX(3): 767-803.

Galbi, Douglas A. (2001). A new account of personalization and effective communication. http://www.galbithink.org/pers.pdf (as of Nov 2003, 2010/03/10)

Galbi, Douglas A. (2002). Long-term trends in personal given name frequencies in the UK. http://www.galbithink.org/names.htm (as of 2010/03/10)

Gerhards, Jürgen (2003). *Die Moderne und ihre Vornamen*. Opladen: Westdeutscher Verlag.

Gerhards, Jürgen (2005). *The name game: Cultural modernization and first names*. Edison: Transaction Publishers.

Gerhards, Jürgens & Hans, Silke (2009). From Hasan to Herbert. name giving patterns of immigrant parents between acculturation and ethnic maintenance. *American Journal of Sociology* 114(4): 1102-1128.

Hanks, Patrick Wyndham & Tucker, D. Kenneth (2000). A diagnostic database of American personal names. *Names* 48(1): 59-69.

Hirschman, Albert Otto (1964): The paternity of an index. *American Economic Review*: 54(5): 761.

Huschka, Denis, Gerhards, Jürgen & Wagner, Gert G. (2005). Messung und Analyse des sozialen Wandel anhand der Vergabe von Vornamen: Aufbereitung und Auswertung des SOEP. Dokumentation der Datenbasis und der Vercodung. http://www.polsoz.fu-berlin.de/soziologie/arbeitsbereiche/makrosoziologie/projekte/dateien/projektdoku_vornamen.pdf

Huschka, Denis, Gerhards, Jürgen & Wagner, Gert G. (2009). Naming differences in divided Germany. *Names* 57(4): 208–228.

Kelly, William A. Jr. (1981). A generalized interpretation of the Herfindahl Index. *Southern Economic Journal* 48(1): 50-57.

Lieberson, Stanley (2000). *A matter of taste. How names, fashions, and culture change*. New Haven: Yale University Press.

Lieberson, Stanley & Bell, Eleanor O. (1992). Children's first names: An empirical study of social taste. *American Journal of Sociology*. 98(3): 511-554.

Scherer, Frederic Michael & Ross, David (1990). *Industrial market structure and economic performance*. Boston: Houghton Mifflin Company Boston.

Shannon, Claude Elwood & Weaver, Warren (1999). *The Mathematical Theory of Communication*. Chicago: University of Illinois Press.

Simon, Michael (1989). *Vornamen wozu? – Taufe, Patenwahl und Namensgebung in Westfalen vom 17. Jahrhundert bis zum 20. Jahrhundert*. Münster: F. Coppenrath Verlag.

Simpson, Edward H. (1949): Measurement of diversity. *Nature* 163: 688.

Tucker, D. Kenneth (2001). Distribution of forenames, surnames, and forename-surename pairs in the United States. *Names* 49(2): 69-96.

Tucker, D. Kenneth (2004). The forenames and surnames from the GB 1998 electoral roll compared with those from the UK 1881 census. *Nomina* 27: 5-40.

Wagner, Gert G., Burkhauser, Richard V. & Behringer, Friederike (1993). The English Language Public Use File of the German Socio-Economic Panel. *The Journal of Human Resources* 28(2): 429-433.

Wagner, Gert G., Frick, Joachim R. & Schupp, Jürgen (2007). The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements. *Schmollers Jahrbuch* 127(1): 139-169.

Wolffsohn, Michael & Brechenmacher, Thomas (1999): *Die Deutschen und ihre Vornamen*. München: Diana Verlag.

Yule, George Udny (1944): *The Statistical Study of Literary Vocabulary*. London: Cambridge University Press.