




Perceived Biological and Social Characteristics of a Representative Set of German First Names

Tillmann Nett¹ , Angela Dorrough², Marc Jekel², and Andreas Glöckner²

¹Department of Psychology, FernUniversität in Hagen, Germany

²Faculty of Human Sciences, University of Cologne, Germany

Abstract: We provide ratings for a representative set of 2,000 German first names with regard to perceived sex, foreign origin (yes/no), and familiarity. In two studies participants ($N = 736$ and $N = 237$) estimated intelligence, education, attractiveness, religiousness, age, warmth, and competence of persons with the respective name. Descriptive results show strong stereotypes in society in that most of the top-rated names on intelligence, competence, and religiousness were male, whereas all top-rated names on attractiveness and warmth were female. The reliability of most ratings is satisfactory. We provide correlations between the rated dimensions to give an overview of the internal structure of the dataset. To enhance usage of the dataset, we provide an R-package, which allows querying subsets of names depending on experimental requirements.

Keywords: first names, word norms, social perception, stereotypes, German language

A common experimental manipulation in the area of social psychology is to present first names to signal group membership (e.g., gender: Brosi, Spörrle, Welp, & Heilman, 2016; Heyder & Kessels, 2015; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Steinpreis, Anders, & Ritzke, 1999; ethnic groups: Bertrand & Mullainathan, 2004; Lütkenhöner, 2011; different ages: Kuhlmann, Bayen, Meuser, & Kornadt, 2016), to manipulate perceived characteristics, such as intelligence or attractiveness (e.g., Gebauer, Leary, & Neberich, 2012; Greitemeyer & Kunz, 2013; see also Newman, Tan, Caldwell, Duff, & Winer, 2018) or to allow participants to follow a narrative in multiple vignettes (Newman et al., 2018). For a successful manipulation, name carriers must actually be perceived to belong to the intended group and names need to be indeed associated with the intended characteristics. Furthermore, other characteristics associated with the respective names must be comparable and thus unconfounded with group membership (e.g., Böhm, Schütz, Rentzsch, Körner, & Funke, 2010; Brosi et al., 2016; Heyder & Kessels, 2015; Moss-Racusin et al., 2012; Schulz, Rudolph, Tscharschiew, & Rudolph, 2013; Steinpreis et al., 1999; I. Winkler, Jonas, & Rudolph, 2008, all controlled for one or multiple perceived characteristics of the given names). Typically, researchers in previous studies using first names either generated small ad hoc

samples of names that were rated by a small number of participants in a pilot study (e.g., Bertrand & Mullainathan, 2004; Lütkenhöner, 2011; Stevens, Volstorf, Schooler, & Rieskamp, 2011), re-used names from previous studies with similar research questions (e.g., Gebauer et al., 2012; Moss-Racusin et al., 2012; Steinpreis et al., 1999), or referred to existing validated sets of first names (e.g., Böhm et al., 2010; Brosi et al., 2016; Greitemeyer & Kunz, 2013; Heyder & Kessels, 2015; Kuhlmann et al., 2016; Schulz et al., 2013; I. Winkler et al., 2008). For German names, the dataset provided by Rudolph, Böhm, and Lummer (2007; see also Rudolph & Spörrle, 1999) is most frequently used. This dataset includes the 60 most common German first names (30 male and 30 female), rated by 149 participants in terms of age, intelligence, attractiveness, and religiousness. For names from the United States a dataset is provided by Newman et al. (2018), which includes 400 names (200 male and 200 female), with the names rated by 497 participants in terms of age, competence, and warmth. In some of the existing datasets also other characteristics such as topicality or sex are included. Typically, these characteristics have been measured based on demographic statistics. For example, Rudolph et al. (2007) categorized German first names with regard to topicality (modern, old-fashioned, or ageless) and sex based on demographic statistics about the allocation of

these names to newborns in the years between 1960 and 2004. Similarly, Bertrand and Mullainathan (2004) used statistics about the ethnicity from birth certificates to determine the ethnic membership of a name. While those measures by definition reflect the true relation between name and name carriers (e.g., frequency of names for different ethnic groups), people may have incorrect beliefs about the true relations. Thus, those measures might not be a valid indicator for the potentially distorted perceived relations between names and characteristics. In addition, the definition of the characteristics underlying these measurements may be different from how people actually perceive these characteristics (e.g., problems with construct validity).

In addition to these potential limitations for the use of past validation studies, the number of names included in those study was typically low. Experiments on decision making, for example, often require a large number of trials (e.g., Dorrough, Glöckner, Betsch, & Wille, 2017; Stevens et al., 2011) for which a larger number of names would be desirable to avoid repeating the same names across different trials. Repeating the same name may introduce undesired effects such as increased liking due to mere exposure (Moreland & Zajonc, 1982). Furthermore, the experimental design may require different information to be conveyed to the participants in different trials. Another experimental constraint, which can increase the number of required names, is that some names may not be usable in some studies. For example, Steinpreis et al. (1999) investigated the impact of stereotypes among psychologists using different first names for otherwise equivalent CVs. To avoid confusion with any real existing psychologists and the ones given in the CVs, they ensured that the names used in the study did not appear in the APA membership directory. For small sets of names (e.g., Rudolph et al., 2007), removing names may be impossible or may result in a set of names that are not comparable with respect to other perceived characteristics. Similarly, Rentzsch, Schütz, and Schröder-Abé (2011) specifically mentioned that they did not use any names in their study, as the current existing norms for German first names did not allow them to identify enough names rated similarly in terms of intelligence and attractiveness. By providing a much larger set of names, researchers can remove a larger number of names, which may be problematic for their design, and still be left with a usable number of validated names.

Furthermore, but related to the previous issue, a name set representative of a certain reference set often needs to be selected for generating internally and externally valid results (Brunswick, 1955; see also Newman et al., 2018, for this argument specifically applied to names). Generating such a subset requires a representative set and ratings on relevant selection criteria to begin with. For names, such a set has not been provided so far. Ad hoc selections of names based on experimenters' intuition or mental simulation as well as

some kinds of piloting can artificially increase estimates of the true effect and should, therefore, be avoided (Fiedler, 2011). Even more so, previous studies investigating the names used in psychological experiments have shown a systematic tendency to use names, which can strongly bias the findings in the direction required by the researcher (Kasof, 1993). For example in research on stereotypes, male names were often associated with higher intellectual competence than the female names, with which they were contrasted.

Finally, although ratings of names in prior studies cover plenty of perceived characteristics, ratings on two fundamental dimensions of social perceptions, namely warmth and competence, are usually missing (i.e., Stereotype Content Model, SCM; Fiske, Cuddy, Glick, & Xu, 2002). In research on stereotypes, the SCM has been shown to be a valuable tool for predicting the attitude and behavior toward members of a group. According to this model, the attitude toward a member of a group is governed by the perceived warmth and competence of that group (Fiske, Cuddy, & Glick, 2007). To fill this gap, we also included items from a German questionnaire of warmth and competence (Asbrock, 2010). In previous research, these items have been only used to analyze the perception of social groups, such as immigrants, women, or homeless people but not for names. We will thus also analyze if we can measure these variables reliably and furthermore if we are able to identify the two factors of warmth and competence also for name ratings. In a similar approach, Newman et al. (2018) also included items for warmth and competence when collecting ratings of names for use in the United States. However, Newman et al. (2018) only used single items for collecting the competence and warmth ratings, whereas we used a set of six validated items (Asbrock, 2010).

In sum, we extend prior studies on name sets in three ways. First, we provide a large set of 2,000 representative German male and female names rated by 973 participants. Second, we provide direct ratings on the perceived topicality of names instead of using demographic statistics. Third, we provide additional ratings on name characteristics such as warmth and competence (Asbrock, 2010; Fiske et al., 2002, 2007) that have not been systematically assessed in most other name studies before.

Methods

The software for conducting this research can be retrieved from the Open Science Framework (OSF) at <https://osf.io/jepzp/>. The software is put under an open license (MIT open source license) such that it may be freely adapted and re-used for future research. While the current implementation is in German to generate a German name set, the software is written such that it may easily be

adapted for other countries or languages. For example, all questionnaires are defined in an easy to understand XML format that may be translated into other languages even with little programming knowledge. In addition, the format used to define the questionnaires provides a simple method to add or remove questions.

Participants

In the first study, we collected data from 736 participants who rated subsets of names. These participants were recruited using a participant pool of students from the FernUniversität in Hagen as well as through social media. Participants studying psychology at the FernUniversität in Hagen received partial course credit for taking part in the survey. Each participant rated $M = 54.08$ names on average ($SD = 29.06$) and thus we collected about 20 ratings per name ($M = 19.90$, $SD = 0.72$). With only 20 ratings per name, however, the distribution of the collected ratings may not sufficiently reflect the real distribution. Thus, for example, the mean ratings for each name may have a large measurement error (e.g., difference between estimated mean and true population mean), which may pose a problem when selecting names for other studies based on these ratings. Thus in a second study, we recruited 237 additional participants. These participants did not receive course credit but instead were paid 15 € for taking part in the survey. The names rated by these participants were a subset of the names used in the first study, such that we could collect particularly precise ratings for this subset of names. For all analyses, data from both studies were combined into a single dataset. All participants in both studies indicated that they were fluent in German. One hundred seven participants indicated they were not from Germany or did not provide any information about their origin. The mean age of the participants was 34.24 years ($SD = 10.69$; ages above 78 were imputed as 78 since no selection of birthdates before 1940 was possible). Of the analyzed participants, 73% participants were female, 27% participants were male, and 0.31% identified as neither male nor female. The majority of participants were students (62%) and/or employed (51%) with the majority of the students in the field of psychology (83%).

Materials: Selection of the Initial Nameset

To select a set of names, we started with a large initial set of 8,173 names taken from a German name dictionary (Duden, 2007). This dictionary contains two tables of male

and female names, which we scanned and translated to text using optical character recognition (OCR). To check for errors during translations, we checked all names against a corpus of German words generated from newspaper articles from 2011 and 2012, which has been made available as part of the “Leipzig Wortschatz” project (Biemann, Heyer, Quasthoff, & Richter, 2007) and manually corrected where necessary. In cases where two names were very similar to each other, only the more common name (e.g., the one with the highest number of occurrences in the corpus) was kept. The similarity was determined using the following criteria: (1) The names differ only in terms of diacritics (äüößé). For example, “Jérôme” and “Jerome” were considered different forms of the same name and Jerome was kept because of the higher occurrence in the corpus (10 names were removed because they were considered to be the less common form of another name). (2) The names were similar in terms of sound *and* also similarly spelled. The sound similarity was calculated using the “Kölner Phonetik” (Postel, 1969). The Kölner Phonetik translates words or names to soundcodes, which correspond to the perceptual features of the name in German (e.g., by encoding guttural and plosive phones differently). For all names which were similar in terms of sound, similarity in spelling was checked using the Jaro-Winkler similarity (W. E. Winkler, 1990; see also W. Cohen, Ravikumar, & Fienberg, 2003) on character sequences ranging from 0 (= *the names are very dissimilar*) to 1 (= *the names are exactly the same*). Names were considered similar if the Jaro-Winkler similarity was above .8. The value of .8 was chosen, such that as many names as possible were removed, while still keeping at least 1,500 male and female names in the generated dataset (4,095 names were removed because they were too similar to another name).¹ In addition to removing similar names, we also removed any name that appeared both in the male and the female table of the name dictionary (210 names were removed due to sex ambiguity) and all names which appeared in neither corpus (2011 or 2012) due to being very uncommon in Germany (e.g., “Ermengard” or “Jodyokus”; 922 names were removed as uncommon). A full list of names and reasons for their removal is provided on OSF. The complete set of initial names, the number of occurrences from the two corpora and the generated soundcodes have also been uploaded to OSF. Furthermore, all python scripts, which were used to query the corpora, generate the soundcodes, and filter the names are provided on OSF for replicability.

This method left us with a set of 1,804 male and 1,524 female names. From these names, the 1,000 names for

¹ For example, the name “Adda” was removed because it was similar sounding and spelled similarly to the more frequent “Ada” (Jaro-Winkler similarity: .92); however, the name “Kimiko” was not removed although it had the same soundcodes as “Chinook,” “Kinga,” and “Ganga” because all three alternatives differed considerably in spelling (Jaro-Winkler similarity $\leq .60$).

Table 1. Rated dimensions and derived variables

Dimension	Type	Code	M (SD)	Reliability [95% CI]	N _{eff} ^e
Sex (weighted) ^{a,c}			0.30 (6.11)	.99 [0.98, 1.00]	55,955
Sex	Categorical	−1: Female, +1: Male	0.06 (1.00)	.99 [.98, .99]	55,955
Sex (certainty)	7-point Likert		5.85 (1.79)	.92 [.90, .93]	9,891
Topicality	Categorical	1-hot ^f			
Modern ^{a,c}			0.23 (0.42)	.82 [.79, .85]	2,717
Old ^{a,c}			0.40 (0.49)	.90 [.88, .92]	4,733
Ageless ^{a,c}			0.37 (0.48)	.75 [.71, .79]	3,067
Education ^c	7-point Likert		4.24 (1.22)	.72 [.67, .76]	6,812
Age ^c	Multiple Choice	≤ 20: 1, 20–30: 2, ..., ≥ 61: 6	3.39 (1.45)	.92 [.90, .94]	19,153
Attractiveness ^c	7-point Likert		4.13 (1.26)	.77 [.73, .80]	7,775
Intelligence ^c	7-point Likert		4.26 (1.18)	.66 [.61, .71]	5,680
Religiousness ^c	7-point Likert		3.81 (1.44)	.77 [.73, .80]	8,385
Competence (SCM) ^{a,c}			4.29 (1.06)	.58 [.51, .64]	3,606
Competent ^b	7-point Likert		4.25 (1.15)	.59 [.53, .65]	4,924
Competitive ^b	7-point Likert		4.18 (1.18)	.53 [.46, .60]	5,677
Independent ^b	7-point Likert		4.43 (1.23)	.48 [.40, .55]	4,410
Warmth (SCM) ^{a,c}			4.34 (1.07)	.58 [.51, .63]	3,343
Likable ^b	7-point Likert		4.37 (1.19)	.55 [.48, .61]	4,601
Warm ^b	7-point Likert		4.33 (1.17)	.56 [.49, .62]	4,718
Good natured ^b	7-point Likert		4.32 (1.17)	.51 [.44, .58]	4,324
Nationality ^c	Categorical	0: Foreign, 1: German	0.50 (0.50)	.95 [.94, .96]	2,280
Familiarity ^c	7-point Likert		3.02 (1.72)	.93 [.91, .94]	6,626
Associations ^d	Free text				

Notes. ^aDerived variable. ^bUsed in the factor analysis of warmth/competence (see Section "Ratings of Warmth and Competence"). ^cUsed during dimensionality reduction (see Section "Choosing Similar Names" and <https://osf.io/hcx2v/>). ^dProvided as is for future research but not included in any analysis here. ^eEffective sample size based on design effect correction using the intracluster correlation coefficient (ICC) with participants as clusters. ^f1-hot coding of categorical variable with three levels as three separate numerical variables; Modern: (1, 0, 0), Old: (0, 1, 0), Ageless: (0, 0, 1).

each sex according to the name dictionary (Duden, 2007), which had the highest total occurrence in the newspaper corpora for 2011 and 2012, were used as representative German names.

Procedure

Ratings were collected using an online survey, which was programmed in oTree (Chen, Schonger, & Wickens, 2016). A translated example of the full survey is provided at <https://osf.io/erykn/>. The original in German can be found on OSF at <https://osf.io/uwdt9/>. A full list of dimensions on which each name was rated is summarized in Table 1. To assess these dimensions, we asked participants to indicate the ratings for the average person with this name (e.g., whether the average person with this name is female or male; not at all vs. very educated/intelligent; etc.).

Participants were asked to agree with a statement of consent about data collection and usage before starting with the main part of the study and provided demographic data. They were then directed to the main survey, in which each participant was asked to rate a subset of the names. Each name

and the associated ratings were presented on a separate page. Most items were taken from the study by Rudolph et al. (2007), currently the most extensive existing validated name set for German first names. Furthermore, as outlined above, we also included questions about the perceived sex of the name and its topicality (modern, old, or ageless) and items to measure perceived warmth and competence (Asbrock, 2010). In addition, participants indicated how certain they were about the associated sex, whether they considered this name to be a German name and how common they believed this name to be in Germany. Finally, to also collect open-ended perceived characteristics with the names, we provided a text field in which participants could provide any association they had with that name. We do not analyze these open answers in the current article but they might be used in future research to extract potential stereotype dimensions for names (cf. Koch, Imhoff, Dotsch, Unkelbach, & Alves, 2016). All ratings except for sex, age, age category, origin, and the free written associations were collected using a 7-point Likert scale with labels only at the endpoints of the scale (e.g., *not intelligent at all* vs. *very intelligent*). The ratings for age were collected using a 6-point

scale with age ratings between 20 years and 60 years in intervals of 10 years (1 = *less than 20*, 2 = *20–30*, 3 = *31–40*, 4 = *41–50*, 5 = *51–60*, 6 = *more than 61*; Rudolph et al., 2007). The ratings for sex, age category, and origin were collected using drop-down lists, from which the participants could select the appropriate response.

To generate the stimulus material for participants in the first study, we constructed sets of 75 different names from all 2,000 names, such that each name was used exactly 15 times in each set (400 sets in total). These sets were then used in the first round of the survey. However, since some of the initial 400 participants did not finish the survey, the frequency of ratings for each name differed at this point. Therefore, after the first phase of data collection, we created novel sets of 75 names, in which the names that previously had received a lower number of ratings were included more often. As before, participants never rated the same name twice. This process was repeated until we had at least 15 ratings for each name. The order in which the names were presented was randomized during trial generation. In the first study, we were able to achieve about 20 ($M = 19.90$, $SD = 0.72$) ratings per name. To collect more ratings per name for some names, in the second study we selected 200 names which were rated by new participants. These 200 names included 45 names that were also included in the study by Rudolph et al. (2007; see Table 2 for a complete list). In addition, we included names based on the following procedure: First we assigned sex and topicality categories to all names, such that each name was assigned the sex and topicality category that was chosen most often by participants in the first study. Based on these sex and topicality categories, we split our dataset into six groups (3 Topicality Categories \times 2 Sex Categories). From each of the six groups we selected those names rated as most familiar on average in the first study, such that an approximately equal number of names was selected from each of the groups. Participants in the second phase were given random sets of 75 names sampled from these 200 names only. The participants included in the analysis (both studies) rated $M = 57.51$ names on average ($SD = 27.65$). Because we also included data from participants, who did not complete the survey, the number of names rated are less than 75 for some of the participants. Each of the 2,000 names was rated between 17 and 103 times in total ($M = 27.98$, $SD = 23.9$) for a total of 55,955 name ratings.

After rating the names, participants were thanked and redirected to a page, where they could collect the course credit for the survey and send an email to receive additional information about the goals of this study. The goals of this study were not disclosed before all data were collected.

Results

Analyses were conducted using the R programming language version 3.4.1 (R Core Team, 2017) with the tidyverse set of packages (Wickham, 2017) for data preparation and wrangling. The original document for this paper used knitr (Xie, 2018; see also Xie, 2014) to embed R code into the document to ensure reproducible research (De Leeuw, 2001) and to prevent transcription errors of the computed values (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Figures are generated using the ggplot2 package (Wickham et al., 2018). The complete set of scripts, seed values for the random number generator, and original raw data files used to compute the analyses is provided on the OSF.

Since our analyses aim to provide insights into the structure and quality of the collected dataset and are not meant to test any scientific research hypotheses, we present descriptive statistics, effect sizes, and the confidence intervals of these effect sizes. For all analyses, which were conducted multiple times (e.g., for the reliability of multiple ratings), we adjusted individual confidence intervals such that an aggregate confidence of 95% is assured (Bonferroni corrected confidence intervals).

Descriptive Results

Table 3 provides descriptive results for the 10 highest and lowest rated names for the dimensions intelligence (Table 3A), education (Table 3B), attractiveness (Table 3C), religiousness (Table 3D), familiarity (Table 3E), and age (Table 3F). The descriptive results show strong prevailing gender stereotypes in German society that are attributed to the average persons with male versus female names. Within the top rated names for intelligence and religiousness, there were almost exclusively male names, with the only exception of the name “Mitsuko” among names rated as most educated and the name “Aygül” among the names rated as most religious.² The female name rated as most intelligent was Viktoria with an average rating of $M = 5.3$ (rank 11). The reversed picture emerges for the dimension attractiveness in which the top names included only female names. For attractiveness, the male name with the highest rating was Raul with an average rating of $M = 5.25$ (rank 13). These observations can mainly be confirmed for the complete dataset. For all names the ratings for sex (weighted) (scale: $-1 = \text{certainly female}$ to $1 = \text{certainly male}$; sex rating weighted by confidence, details below) and intelligence correlated significantly even after controlling for

² The appearance of these names is most likely caused by stereotypes about the “efficient Asian” (Asbrock, 2010; Fiske et al., 2002) and “religious Muslim” (Koch et al., 2016), which may overrule the “women” stereotype for these two names.

Table 2. Demographic topicality from Rudolph et al. (2007) and topicality ratings from this study for all names used in both studies

Name	N	Topicality Rudolph et al. (2007)	Ratings (this study)		
			Ageless (%)	Modern (%)	Old-fashioned (%)
Alexander	100	Ageless	81	6	13
Andreas	101	Ageless	65	8	27
Christian	100	Ageless	80	8	12
Claudia	100	Ageless	49	9	42
Cornelia	99	Old	32	4	64
David	100	Modern	74	13	13
Dirk	98	Old	38	4	58
Felix	100	Modern	55	32	13
Florian	99	Modern	63	26	11
Frank	100	Old	33	8	59
Heike	100	Old	34	2	64
Heiko	101	Old	27	9	64
Holger	100	Old	17	3	80
Ines	98	Old	51	14	35
Jan	100	Modern	68	27	5
Jens	100	Old	55	15	30
Johanna	101	Modern	72	9	19
Jörg	98	Old	21	8	70
Katharina	99	Modern	79	7	14
Kerstin	100	Old	43	9	48
Laura	101	Modern	62	35	3
Lea	100	Modern	45	54	1
Lena	98	Modern	56	41	3
Leon	100	Modern	40	60	< 1
Leonie	99	Modern	33	62	5
Luca	100	Modern	28	69	3
Lukas	100	Modern	73	19	8
Manuela	99	Old	38	4	58
Maria	99	Ageless	70	3	27
Mario	99	Old	54	20	26
Matthias	100	Ageless	71	6	23
Maximilian	99	Modern	68	14	18
Michael	101	Ageless	75	4	21
Mike	100	Old	37	55	8
Olaf	100	Old	26	5	69
Paul	100	Modern	74	5	21
Peter	101	Old	43	3	54
Petra	99	Old	37	2	61
Sabine	99	Old	37	8	55
Sarah	100	Modern	82	15	3
Silke	99	Old	32	6	62
Susanne	102	Ageless	40	6	54
Thomas	99	Ageless	69	6	25
Tim	100	Modern	59	37	4
Uwe	100	Old	9	5	86

Table 3. Highest and lowest rated names concerning (A) intelligence, (B) education, (C) attractiveness, (D) religiousness, (E) familiarity, and (F) age (scale: 1 (= *not at all*) to 7 (= *very*))

Name	<i>M</i>	<i>SD</i>
(A) Intelligence		
Chen	5.65	0.88
Primus	5.63	1.01
Augustinus	5.60	1.19
Bartholomäus	5.45	1.43
Graham	5.42	1.12
Amadeus	5.35	1.09
Aristoteles	5.35	1.63
Cornelius	5.35	0.88
Fitzgerald	5.35	1.23
Justus	5.35	1.18
...
Igor	3.05	1.39
Chantal	3.05	0.83
Cindy	3.01	1.29
Mandy	2.91	1.20
Fifi	2.90	1.07
Dolly	2.89	1.15
Kevin	2.86	1.22
Cheyenne	2.79	0.92
Jacqueline	2.76	1.18
Candy	2.74	0.87
(B) Education		
Bartholomäus	5.75	1.48
Nathan	5.60	1.14
Primus	5.58	1.12
Amadeus	5.55	1.10
Augustinus	5.55	1.36
Laurentius	5.55	1.19
Graham	5.47	1.43
Cornelius	5.45	1.15
Mitsuko	5.42	1.26
Jacques	5.40	1.14
...
Aga	2.89	1.20
Cindy	2.84	1.29
Jacqueline	2.79	1.36
Mandy	2.78	1.37
Kevin	2.76	1.27
Destiny	2.70	1.34
Cheyenne	2.68	1.00
Dolly	2.58	1.17
Fifi	2.50	1.10
Candy	2.42	0.90
(C) Attractiveness		
Flora	5.60	1.27
Liz	5.55	1.10
Fleur	5.50	1.19

(Continued on the right)

Table 3. (Continued)

Name	<i>M</i>	<i>SD</i>
Grace	5.48	1.12
Aurora	5.40	1.35
Giulietta	5.40	1.39
Viktoria	5.40	0.99
Serena	5.35	1.27
Victoria	5.35	0.88
Laetitia	5.30	1.13
...
Fritz	2.80	1.15
Winifred	2.78	1.22
Ottmar	2.75	1.02
Hartmut	2.75	1.07
Ekkehard	2.75	1.16
Gottwald	2.70	1.17
Arnulf	2.70	1.17
Adolf	2.65	1.42
Igor	2.58	1.22
Ottfried	2.53	1.17
(D) Religiousness		
Evangelist	5.75	1.33
Hakan	5.75	1.16
Jesus	5.75	1.12
Aygül	5.70	1.13
Moses	5.70	1.26
Abraham	5.65	1.50
Franziskus	5.65	1.09
Josefa	5.65	1.31
Khalid	5.65	1.35
Paulus	5.60	1.23
...
Tilly	2.55	1.36
Roxy	2.55	1.19
Kelvin	2.55	1.19
Bibi	2.55	1.39
Kevin	2.55	1.31
Torben	2.47	1.22
Guy	2.35	1.04
Jacqueline	2.31	1.28
Dexter	2.28	1.41
Chanel	2.15	1.14
(E) Familiarity		
Michael	5.50	1.39
Christian	5.47	1.37
Stefan	5.39	1.26
Andreas	5.34	1.28
Alexander	5.27	1.44
Martin	5.27	1.32
Lisa	5.17	1.56
Daniel	5.16	1.42
Peter	5.16	1.32

(Continued on next page)

Table 3. (Continued)

Name	<i>M</i>	<i>SD</i>
Sabine	5.15	1.34
...
Eitel	1.19	0.60
Kraft	1.17	0.51
Andere	1.17	0.38
Quincy	1.15	0.49
Hai	1.15	0.49
Focke	1.15	0.37
Winnetou	1.14	0.48
Arpad	1.10	0.31
Solange	1.05	0.22
Guadalupe	1.05	0.22
(F) Age		
Klothilde	5.85	0.37
Edelgard	5.80	0.41
Gerhild	5.80	0.41
Sigismund	5.80	0.41
Friedewald	5.79	0.54
Brunhild	5.78	0.55
Irmhild	5.72	0.96
Gertrud	5.68	0.60
Adalbert	5.67	0.73
Ewald	5.65	0.49
...
Justin	1.71	0.76
Faith	1.70	0.92
Destiny	1.70	0.92
Janelle	1.67	0.77
Vanilla	1.65	0.81
Jara	1.65	0.93
Cinderella	1.65	0.81
Emily	1.62	1.02
Finn	1.62	0.83
Fia	1.55	1.23

Notes. (A) Correlation with sex (weighted): $r(1,998) = .10$ [.01, .18], $p < .01$;
 (B) Correlation with sex (weighted): $r(1,998) = .12$ [.03, .20], $p < .01$;
 (C) Correlation with sex (weighted): $r(1,998) = -.35$ [-.42, -.27], $p < .01$;
 (D) Correlation with sex (weighted): $r(1,998) = -.03$ [-.11, .05], $p = .20$;
 (E) Correlation with sex (weighted): $r(1,998) = -.01$ [-.09, .08], $p = .82$;
 (F) Correlation with sex (weighted): $r(1,998) = .14$ [.06, .22], $p < .01$.

multiple comparisons,³ confirming that males are assumed to be more intelligent than females. This effect was not moderated by the gender of the participants (see <https://osf.io/b6kfn/>). Similarly, there also was a significant negative

correlation between ratings of sex (weighted) and attractiveness, confirming that females are rated to be more attractive than males. Again, this effect was not moderated by the gender of the participants (see <https://osf.io/b6kfn/>). Only for religiousness the effect did not hold in an overall analysis but was only apparent in the extremes.

Reliability of Ratings

Since this study is based on a repeated measurement approach, ratings from the same participant may be correlated with each other, and thus the number of 55,955 collected ratings only insufficiently reflects the amount of collected information. Similarly, the number of ratings per name, and the number of names also only insufficiently reflect the amount of collected information. To estimate the actual information contained in the provided dataset, we calculated the effective sample sizes N_{eff} (Hox, Moerbeek, & Van de Schoot, 2017) for each measurement, which are reported in Table 1.⁴ The names included in the first study only have about 20 ratings for each attribute. Since low sample sizes in correlational studies are often linked to unreliable findings (Schönbrodt & Perugini, 2013) and inflated effect sizes (Gelman & Carlin, 2014), one might expect results from statistical analyses on these names to be unreliable. We circumvented this problem by conducting most analyses on the complete dataset either after calculation of the mean ratings, or without aggregation. While the low number of ratings for names only included in the first study may still cause individual means to deviate from the true population mean, this effect vanishes when all names are included in an analysis, and the effective sample sizes then give a much more realistic impression of the reliability of the results. Since all effective sample sizes are above 2,000, the chances that any of the results presented here are unreliable and would disappear with larger sample sizes (Schönbrodt & Perugini, 2013) can be considered negligible.

In addition, to test how well ratings from different participants for the same name corresponded to each other, we calculated split-half reliabilities using the formula for the two-part alpha reliability coefficient ($r_{2\alpha}$) and the corresponding confidence intervals for perceived characteristics to assess in how far different participants perceive names in a consistent way (Charter, 2000; see also Kristof, 1969). For this, we split all individual ratings for the same name randomly into two sets and calculated the mean

³ We tested all possible correlations between the collected ratings (78 comparisons). Therefore, significance was tested at $\alpha = .00064$. See Table 3 for r and p -values.

⁴ The effective sample size for a repeated measurement design is an indication of the numbers of samples required to gather the same amount of information without repeated measures. While the effective sample size can also be used to determine the df of a statistical test, for most tests we instead chose a conservative approach by only basing the df on the number of names.

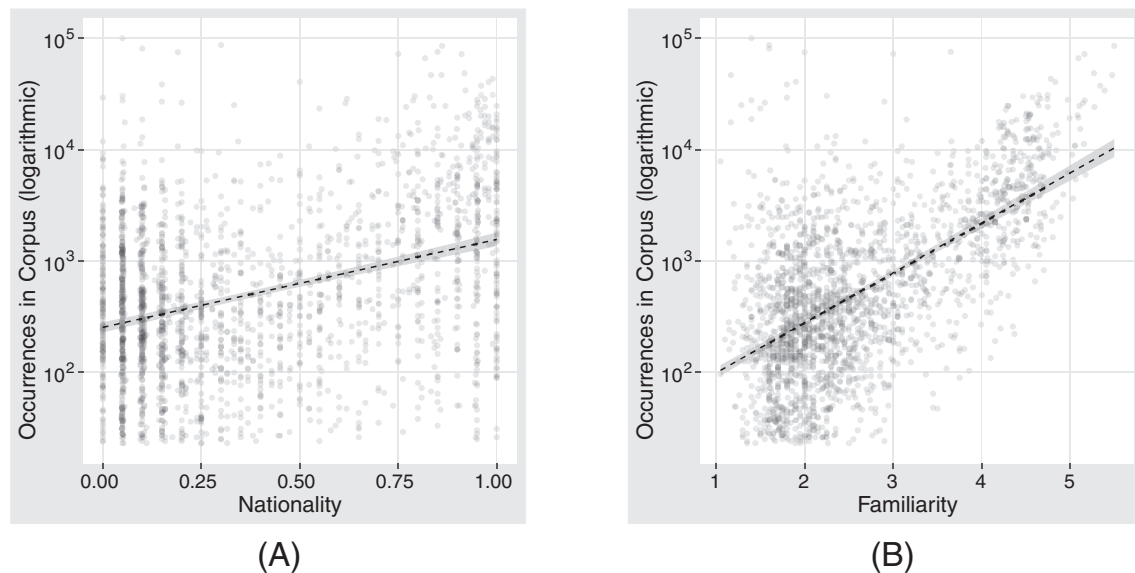


Figure 1. Correlation between the frequency of occurrences and nationality ratings (A) and familiarity ratings (B).

ratings for each name separately. The paired-sample correlations (r_{XY}) of the mean ratings for each name were then used to derive maximum likelihood estimates of the reliability coefficient alpha ($r_{2\alpha}$) using the formula provided by Charter (2000). We used maximum likelihood estimators for two-part alpha instead of Spearman-Brown corrections for split-half reliability since two-part alpha, in general, provides more reliable estimates and confidence intervals (Charter, 2000). We repeated this random splitting 1,000 times and averaged the resulting reliability scores and confidence intervals.⁵ For sex ratings, to also include the confidence participants had in their choice, we multiplied the recoded sex variable (-1 for female and $+1$ for male) by the numeric confidence variable (weighted sex ratings) to achieve ratings that can be interpreted as more or less certainly female/male. Higher positive values thus reflect more certain ratings of male sex while lower negative values indicate more certain ratings of female sex and values close to zero uncertain sex ratings. For the categorical topicality variable, we recoded all ratings using 1-hot coding.⁶ Overall, reliabilities differed largely between items (Table 1), ranging from $r_{2\alpha} = .48$ for the item independence to $r_{2\alpha} = .99$ for the weighted sex ratings. All variables, which described objective characteristics of a person (age, sex, and nationality) but also familiarity of the name

in Germany (which is directly related to nationality) showed excellent reliability scores ($r_{2\alpha} \geq .92$). Subjective ratings of attractiveness, intelligence, education, and religiousness were of moderate reliability ($.66 \leq r_{2\alpha} \leq .77$). The reliability of the warmth and competence variables (Asbrock, 2010) showed poor reliability ($.48 \leq r_{2\alpha} \leq .56$), indicating that different participants rated the same name very differently on these items. Also, variables for warmth and competence were similar to each other with regard to their reliability. Finally, topicality showed acceptable reliability ($.75 \leq r_{2\alpha} \leq .90$).

Validity of Ratings for Familiarity and Sex

To test the convergent validity of the ratings for familiarity and sex, we correlated those variables to external criteria that should be related. For sensible familiarity ratings, the frequency of occurrence of a name in a German text corpus should be correlated with participants' ratings of familiarity. Additionally, more frequent names in a German text corpus (Biemann et al., 2007) are more likely of German than of foreign origin. As predicted, a comparison between the ratings of nationality (German or foreign name) and familiarity of the names with the frequency of occurrences in the text corpus showed a medium correlation between

⁵ Separate confidence intervals were computed for each split using the formula provided by Charter (2000). The values from all these estimates were then averaged to reduce the influence of each single split. Since we only varied the random splits of the datasets while keeping all ratings, the method we used cannot be considered a bootstrap, and therefore the individual estimates may not be used when determining the confidence intervals.

⁶ 1-hot or one-of-many coding recodes a categorical variable with K categories into K separate variables. This is similar to dummy coding, which, however, only uses $K - 1$ variables and takes one of the categories as a reference category. We chose 1-hot coding instead of dummy coding, because it does not require choosing a reference category, to which all other categories are compared. Also 1-hot coding allows us to report data for all three topicality categories and does not require omission of the reference category.

the ratings for the nationality and the logarithm of the occurrence count ($r(1,998) = .36$ [.31, .42], $p < .01$; see Figure 1A) and a strong correlation between the familiarity ratings and the logarithm of the occurrence count ($r(1,998) = .60$ [.55, .64], $p < .01$; see Figure 1B).⁷

For sensible sex ratings, participants' classification of names regarding a name carrier's sex should correspond to some extent to the classification in the name dictionary (Duden, 2007). An independent sample t -test of the sex ratings with the names split according to the sex provided by the name dictionary and female coded as -1 , male coded as $+1$ showed a large difference of the mean ratings, $t(1,583) = 96.75$, $d = 4.33$ [4.14, 4.51], $p < .01$ ($M = -0.74$, $SD = 0.46$ for female names and $M = 0.88$, $SD = 0.26$ for male names). This shows that participants rated the names listed as male names in the name dictionary more often as male compared to female and the other way around for female names. Cohen's d and the confidence intervals for this and the next analysis were computed using the effsize package (Torchiano, 2018), with Bonferroni corrections on the confidence levels. Degrees of freedom are corrected using the Welch modification, as the variances in both groups may differ. In addition to the sex ratings, we also collected ratings of the confidence that participants had in their sex ratings. We expected some of the names to be more or less ambiguous than others. An independent sample t -test of the weighted sex ratings produced similar results, showing a large difference of the mean ratings, $t(1,741) = 94.39$, $d = 4.22$ [4.04, 4.41], $p < .01$ ($M = -4.55$, $SD = 2.73$ for female names and $M = 5.25$, $SD = 1.82$ for male names). This result shows that participants were also more certain with their sex rating if their rating corresponds to the Duden sex classification. In sum, the results demonstrate a large correspondence between our collected ratings and ratings provided from other sources. Nevertheless, a manual inspection of the names, which were most strongly assigned to a different sex compared to the source material, showed that some names were consistently rated as belonging to a different gender. However, this comparison indicated that the difference between the ratings and the source material can mostly be explained by errors in the source material or changes in the usage of the names since the source material was collected.

To conclude, the comparison of the datasets with other sources of the same or similar variables demonstrates a

reasonably high validity for ratings of demographic characteristics. This matches the analysis of the internal reliability from the previous section, which also found excellent reliability for all ratings of demographic characteristics. This demonstrates that ratings can be used to manipulate or control sex, nationality, or familiarity of a name in future studies.

Ratings of Warmth and Competence

In addition to the items used by Rudolph et al. (2007), we also included a German version of warmth and competence items, which can be used to predict the perception of a name, most importantly the attitude of a participant toward a carrier of that name, according to the SCM (Asbrock, 2010). To test if the included first names can be also located along the dimensions of warmth and competence as used for the stereotype content model, we first examined the number of meaningful factors that can be extracted from ratings using a principal component analysis (PCA). For this PCA we only used the six warmth and competence ratings (see Table 1 for details). For this analysis, we averaged all ratings for each name and scaled and centered the resulting variables, then we computed a PCA on these averaged ratings to identify the number of factors underlying the ratings of all names. An inspection of the scree plot (see <https://osf.io/v5fsy/>) showed that two principal components can capture a substantial portion of the variance of the ratings. Since a manual inspection of the scree plot is highly subjective and therefore open to debate, we also confirmed results of two components using a parallel analysis (Horn, 1965)⁸ and bootstrapping. Together, these two components were able to account for 90% of the total variance. We thereby confirm the hypothesis of Asbrock (2010) that these items can be organized along two separate dimensions.

To extract two factors from the six PCA components and to confirm that these dimensions indeed correspond to the concepts of warmth and competence, we performed a factor analysis by computing a PCA followed by dropping the four components with lowest variance explanation and a promax rotation of the retained two components (Asbrock, 2010) using the psych R-package (Revelle, 2019). The resulting loadings showed that the variables corresponding to competence loaded strongly and almost exclusively on a single factor with all other variables corresponding to warmth

⁷ We transformed the occurrence counts using a logarithmic scale since word occurrences tend to follow a Zipf distribution, which is essentially an exponential distribution in nature, and also their psycholinguistic properties tend to follow an exponential law (e.g., van Heuven, Mandera, Keuleers, & Brysbaert, 2014). Using untransformed occurrence counts, we found somewhat weaker but still reliable correlations ($r(1,998) = .23$ [.17, .28], $p < .01$ for nationality and $r(1,998) = .33$ [.28, .39], $p < .01$ for familiarity; confidence intervals were corrected to achieve simultaneous 95% confidence across all four correlations). Significance tests are done with $\alpha = .01250$ (four simultaneous tests).

⁸ To match the (unknown) distribution of the data, we used bootstrapping. To remove the correlations, we sampled all variables independently of each other. To retain the between-subject differences in the random datasets we separately bootstrapped the data for each participant, similar to methods commonly used for bootstrapping multi-level models (Van der Leeden, Busing, & Meijer, 1997). We performed 1,000 bootstraps.

Table 4. Factor loadings for the stereotype content model variables

Item	Competence	Warmth
Competent	.91	
Competitive	.99	
Independent	.94	
Likable	.21	.83
Warm		.99
Good natured		.98
Variance explained	46%	44%

Note. Factor loadings < .20 omitted.

loading on the other factor (see Table 4). The only exception was the item “Likable,” which was also somewhat correlated with the competence variables, albeit much lower than with the warmth variables. To include the factors competence and warmth from the SCM in the provided dataset, we then averaged the ratings for the three competence items to calculate a total competence score and the three warmth items to calculate a total warmth score. Furthermore, we checked whether averaging of the variables increased the overall low reliability of the SCM variables. The reliability, however, remained low ($r_{2\alpha} = .58$ [.51, .64] for competence and $r_{2\alpha} = .58$ [.51, .63] for warmth).

As before, we provide lists of the ten names rated the highest and lowest on these aggregate factors in Table 5. This table shows a similar prevailing gender stereotype as for the ratings of intelligence and attractiveness. Among the ten most highly rated names for competence, the only female name is again the name “Mitsuko.” In contrast, the ten names rated highest for warmth are exclusively female, with the name “Giovanni” as the highest rated male name for warmth ($M = 5.12$, rank 12). This is in line with other research on stereotype content, which frequently finds women to be rated as warmer but less competent compared to men (Asbrock, 2010; Fiske et al., 2002). Both competence as well as warmth correlated significantly with the sex (weighted) ratings.⁹ Neither the correlation between weighted sex ratings and competence nor the correlation between weighted sex ratings and warmth was moderated by the gender of the participants (see <https://osf.io/b6kfn/>).

Exploratory Analyses: Item Inter-Correlations

To identify relationships between the collected variables, we calculated pairwise correlations between all variables in an exploratory analysis. For this, we averaged the ratings for the same name from all participants who rated the name

Table 5. Highest and lowest rated names for the factors (A) Competence and (B) Warmth

Name	<i>M</i>	<i>SD</i>
(A) Competence		
Chen	5.53	0.83
Jacques	5.27	0.93
Cornelius	5.27	0.73
Erasmus	5.25	0.99
Primus	5.18	0.93
Neil	5.17	0.96
Mitsuko	5.16	1.08
Augustinus	5.15	0.95
Aristoteles	5.15	1.29
Clemens	5.14	1.01
...
Cindy	3.31	1.06
Mandy	3.29	1.07
Chantal	3.23	0.80
Cinderella	3.18	1.43
Destiny	3.13	1.11
Kevin	3.13	1.17
Cheyenne	3.09	1.08
Candy	3.05	0.90
Fifi	3.05	1.11
Jacqueline	3.03	1.17
(B) Warmth		
Gretchen	5.32	1.16
Bruni	5.23	0.96
Lisbeth	5.22	1.08
Betty	5.20	0.96
Jolanda	5.20	0.68
Maria	5.19	1.04
Rosalinde	5.15	1.16
Lucia	5.13	1.18
Anneli	5.13	0.94
Bea	5.13	1.24
...
Haider	3.33	1.12
Achmed	3.31	0.87
Zdenek	3.3	0.75
Erdogan	3.24	0.91
Igor	3.21	1.24
Hussein	3.21	1.01
Hassan	3.15	1.11
Arnulf	3.10	1.17
Etzel	3.05	1.24
Adolf	2.77	1.31

Notes. (A) Correlation with sex (weighted): $r(1,998) = .23$ [.15, .30], $p < .01$; (B) Correlation with sex (weighted): $r(1,998) = -.37$ [−.43, −.29], $p < .01$.

⁹ We tested all possible correlations between the collected ratings (78 comparisons). Therefore, significance was tested at $\alpha = .00064$.

Table 6. Correlations between the three topicality categories and all other dimensions

Rating	Modern		Ageless		Old-Fashioned	
	<i>r</i>	95% CI	<i>r</i>	95% CI	<i>r</i>	95% CI
Sex (weighted)	-.15	[-.23, -.07]*	< .01	[-.08, .08] <i>ns</i>	.10	[.02, .18]*
Education	-.27	[-.34, -.19]*	.12	[.04, .20]*	.10	[.02, .18]*
Age	-.74	[-.78, -.70]*	-.56	[-.61, -.50]*	.89	[.86, .91]*
Attractiveness	.42	[.35, .49]*	.53	[.47, .59]*	-.65	[-.70, -.60]*
Intelligence	-.25	[-.32, -.17]*	.12	[.04, .20]*	.09	[.00, .17]*
Religiousness	-.53	[-.58, -.46]*	-.15	[-.22, -.07]*	.46	[.39, .52]*
Competence (SCM)	-.25	[-.33, -.18]*	.19	[.11, .26]*	.04	[-.04, .12] <i>ns</i>
Warmth (SCM)	-.04	[-.12, .05] <i>ns</i>	.17	[.09, .25]*	-.10	[-.18, -.01]*
Nationality	-.40	[-.47, -.33]*	-.35	[-.42, -.28]*	.52	[.46, .58]*
Familiarity	-.31	[-.38, -.23]*	.16	[.08, .24]*	.10	[.02, .18]*

Note. All *df* = 1,998. **p* < .00064 (equivalent to *p* < .05 after Bonferroni correction for 78 simultaneous tests).

Table 7. Correlations between all variables with Bonferroni corrected confidence intervals

	Age	Attractiveness	Education	Familiarity	Intelligence	Warmth (SCM)
Attractiveness	-.58 [-.64, -.52]		.32 [.24, .39]		.35 [.27, .42]	.50 [.43, .56]
Competence (SCM)		.34 [.26, .41]	.85 [.83, .88]		.89 [.86, .91]	.37 [.30, .44]
Intelligence		.35 [.27, .42]	.92 [.90, .94]			.43 [.36, .49]
Nationality	.54 [.48, .59]	-.32 [-.39, -.25]		.67 [.62, .72]		
Religiousness	.41 [.34, .48]					
Warmth (SCM)		.50 [.43, .56]	.38 [.31, .45]	.31 [.24, .39]	.43 [.36, .49]	

prior to calculating the correlations. Since the categorical variable “topicality” was coded as three separate variables (1-hot coding), we performed individual correlations for each of the three topicality categories. The aggregated topicality variables measure the proportion of the participants who rated the name in each category. The correlations between the three topicality categories and all other ratings are given in Table 6. The correlations between sex (weighted) and all variables are reported in Tables 3 and 5. Other correlations that were at least of moderate size ($|r| > .3$; J. Cohen, 1988) can be found in Table 7. All correlations in Table 7 are also significant (all *ps* ≤ .00064; Bonferroni correction for 78 simultaneous tests). To keep all tests conservative, the degrees of freedom of the test statistics were estimated based on the number of names in the study. The number of names was below the estimated effective sample size for all characteristics (see Table 1). In addition, we put a strong focus on correct positive results, by only providing correlations of at least moderate size in Table 7 instead of providing all statistically significant correlations. In line with Rudolph et al. (2007)¹⁰ we found a significant negative correlation between both the topicality “modern” as well as “ageless” with age ratings, showing that the perceived age of name carriers decreases, the more frequently their names were rated as modern or ageless

names. In contrast, for the topicality “old” we found a significant positive correlation, showing that name carriers were rated as older, the more frequently a name is rated as old-fashioned. For attractiveness, the results differed somewhat from the pattern found by Rudolph et al. (2007). The strongest correlations between attractiveness and topicality were found for the topicality “ageless,” with somewhat reduced correlations for the topicality “modern,” showing name carriers were rated as more attractive the more often their names are rated as ageless or modern. In contrast, we found that the “old” topicality was negatively correlated with attractiveness, such that names were rated as less attractive the more often they were also rated as old-fashioned. For intelligence, we could not confirm the results found by Rudolph et al. (2007). Other than in the previous study, modern names were not rated as more intelligent, but rather as less intelligent, whereas ageless names were rated as more intelligent. Also, intelligence ratings were generally higher the more often a name was rated as old-fashioned, whereas Rudolph et al. (2007) found old-fashioned names to be rated as less intelligent. Similarly, the results for religiousness from Rudolph et al. (2007) could not be replicated. Instead of the modern and ageless names being rated as more religious, we found that names were rated as more religious, the more often they were also rated as

¹⁰ Since Rudolph et al. (2007) used demographic statics to define topicality variables instead of including these variables in their ratings, we can only conceptually replicate the statistical tests. Instead of using an ANOVA, we will perform correlations with each coded topicality.

old-fashioned. For the relationships between the other ratings also tested by Rudolph et al. (2007) we replicated the negative correlation between age and attractiveness ($r(1,998) = -.58 [-.64, -.52], p < .01$) and the positive correlation between attractiveness and intelligence ($r(1,998) = .35 [.27, .42], p < .01$). The latter of these can most likely be attributed to some kind of halo effect (Nisbett & Wilson, 1977). However, for our dataset, the correlation between age and intelligence was reversed ($r(1,998) = .18 [.10, .25], p < .01$) showing that older name carriers were rated as more intelligent and not as less intelligent. In addition to these results presented for comparison with the results by Rudolph et al. (2007), we also found the correlations between gender and intelligence, attractiveness, warmth, and competence, which we already reported in Descriptive Results and Ratings of Warmth and Competence sections. In addition, names rated as warmer on average were also rated as more attractive, better educated, more intelligent, and more competent. The same was true for competence, which also showed a correlation with attractiveness, education, and intelligence. The correlation between warmth and competence found for this dataset was atypical, as other studies on the stereotype content model found these two scales to be mostly uncorrelated (e.g., Asbrock, 2010; but see also Koch et al., 2016). Finally, the nationality ratings correlated negatively with attractiveness and positively with age, showing that carriers of German names were rated as less attractive and older than those with foreign names.

Comparison of Our Data With Rudolph et al. (2007)

We observed statistically significant correlations that differed in sign in comparison to correlations reported by Rudolph et al. (2007). Differences in the methodology that may explain these discrepancies are discussed below. First, instead of using demographics to determine the topicality, in our study participants rated names in terms of perceived topicality. Therefore, the variables representing the topicalities in our analysis could differ from the ones used by Rudolph et al. (2007). Second, the names we used in our study come from a much larger set, including many less popular and unusual names. For example, our dataset included some modern names that follow short lived trends and are mostly associated with lower social and educational class, such as “Destiny” or “Cheyenne” (see also the lower part of Table 3A). Similarly, we included many less popular old fashioned but highly religious names, such as “Moses” or “Abraham.” Since these names were not included by Rudolph et al. (2007), the inclusion in our dataset may have caused the differences.

To test these two possible explanations, we specifically focused on the subset of names also used by Rudolph

et al. (2007). Due to our method of selecting the names from the original dataset, only 45 of the 60 names used by Rudolph et al. (2007) were included in our dataset. Since we selected the names for our second study such that all of these 45 names were included in both studies reported here, we combined the data from both our studies for this analysis. Thus, all analyses are based on about 100 ratings per name ($M = 99.73, SD = 0.89$). To investigate whether differences between the topicality attributes in our study and the study by Rudolph et al. (2007) may explain the different results, we analyzed how strongly the topicality ratings by our participants differed from the topicality categories that were assigned to names by Rudolph et al. (2007) based on demographic statistics (demographic topicalities). Figure 2A shows the aggregated percentages each topicality category was chosen by our participants split by the demographic topicalities (see Table 2 and Figure 2B for a direct comparison). This comparison shows that the perception of the topicalities does not coincide well with the demographic topicalities. This effect was particular strong for names classified as modern based on demographics, with only around 31% of our participants also rating these names as modern and most participants rating these names as ageless (61%). Similarly, names classified as old-fashioned based on demographics were only rated as old-fashioned by 55% of our participants and rated as ageless by 35%. Thus, either participants have incorrect beliefs about the true demographics of these names, or the definitions of the topicalities used by Rudolph et al. (2007) do not reflect how our participants interpreted these terms. In addition Figure 2A also shows the total percentage each topicality category was chosen by our participants for all 45 names used by Rudolph et al. (2007) and us (“Total”) as well as for all 2,000 names used in our study (“Total”). A direct comparison shows, that the name set used by Rudolph et al. (2007) contains a disproportionally large amount of names our participants perceived as ageless, whereas both old fashioned and modern names were underrepresented. Since topicality was used as a variable in most analyses performed by Rudolph et al. (2007) and us, the differences in the variables may explain the different findings.

To further analyze the differences between our findings and the ones presented by Rudolph et al. (2007), we repeated the analyses restricted to the 45 names contained in both datasets. Since we only calculated the correlations corresponding to the analyses by Rudolph et al. (2007) instead of correlating all variables, we only performed Bonferroni corrections for 15 simultaneous tests (e.g., $\alpha = .003$). In addition, since we found a large difference between the topicality ratings from our participants and the topicalities assigned by Rudolph et al. (2007), we also replicated their analyses on our dataset using an analysis of variance (ANOVA) with the topicality categories derived

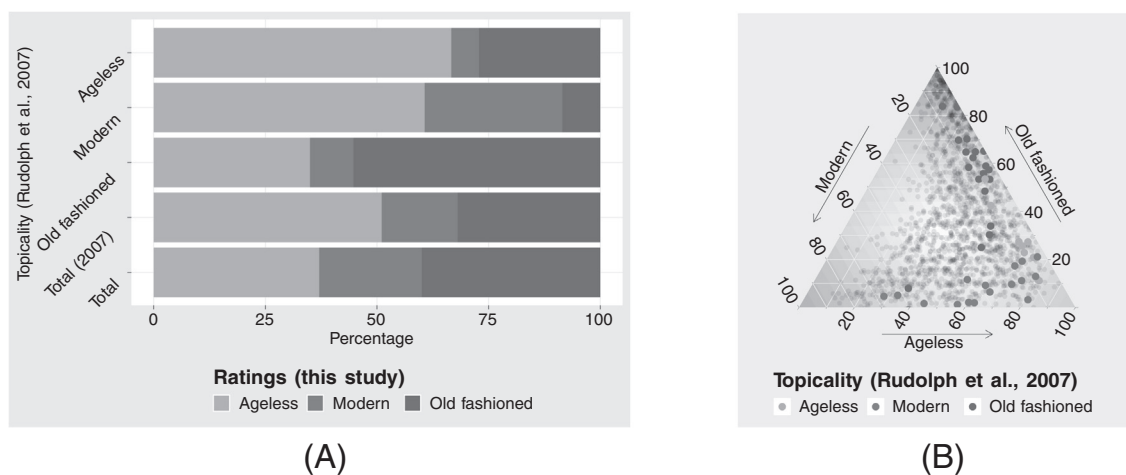


Figure 2. Distribution of the topicality ratings for the names used by Rudolph et al. (2007) aggregated by the topicality assigned in their study (A) and for individual names (B). Axes in (B) indicate the percentage each topicality category was chosen for a name, that is, the sum of all three values for each point is equal to 100% (ternary plot). Gray dots in (B) indicate names not used by Rudolph et al. (2007).

from demographics (demographic topicality) as independent variables. In line with Rudolph et al. (2007), the correlations between the age ratings and the modern and old-fashioned topicality ratings remained statistically significant for the reduced dataset, whereas the ageless topicality ratings were not statistically significant after Bonferroni corrections ($r(43) = -.83 [-.93, -.62]$, $p < .01$ for “modern,” $r(43) = -.41 [-.71, .03]$, $p = .01$ for “ageless,” and $r(43) = .90 [.77, .96]$, $p < .01$ for “old”). Similarly, the age ratings also differed for demographic topicality, $F(2, 42) = 58.83$, $p < .01$, $\eta^2 = .74$. For the correlations between attractiveness and the topicality ratings on the reduced dataset, we also found the same pattern as before, with names being rated as more attractive the more often they were rated as “ageless” or “modern” and names being rated as less attractive the more often they were rated as “old.” Again, the correlation between “ageless” ratings and attractiveness was stronger than between “modern” ratings and attractiveness ($r(43) = .54 [.14, .79]$, $p < .01$ for “modern,” $r(43) = .66 [.32, .85]$, $p < .01$ for “ageless,” and $r(43) = -.89 [-.96, -.74]$, $p < .01$ for “old”), thus showing the same discrepancy between the findings by Rudolph et al. (2007), and our findings. An ANOVA using the demographic topicality also showed statistically significant differences between the three topicality variables, $F(2, 42) = 51.73$, $p < .01$, $\eta^2 = .71$. Most importantly, using the demographic topicality, we found the same pattern reported by Rudolph et al. (2007) ($M = 4.40$, $SD = 0.26$ for “ageless”; $M = 4.68$, $SD = 0.25$ for “modern”; $M = 3.82$, $SD = 0.26$ for “old”). The correlations between topicality ratings and intelligence ratings for the reduced dataset was neither in line with the correlations on the complete dataset nor the ones reported by Rudolph et al. (2007). Just as on the complete dataset, we found a positive correlation between “ageless” ratings and intelligence ratings, showing that names were rated

as more intelligent, the more often they were also rated as ageless ($r(43) = .72 [.42, .88]$, $p < .01$). For “old” ratings and intelligence, the direction was now reversed compared to the previous analysis. Thus, on the reduced dataset, names were descriptively rated as less intelligent the more often they were rated as old-fashioned ($r(43) = -.56 [-.80, -.17]$, $p < .01$). This was more in line with findings by Rudolph et al. (2007), who also found old names being rated as less intelligent. The correlation between “modern” ratings and intelligence was not statistically significant anymore on the reduced dataset ($r(43) = .01 [-.42, .44]$, $p = .94$). Again, an ANOVA with the demographic topicality also showed statistically significant differences between the three categories, $F(2, 42) = 20.12$, $p < .01$, $\eta^2 = .49$. Most importantly, using the demographic topicality we found the same pattern as Rudolph et al. (2007), with the old-fashioned names being rated as least intelligent ($M = 4.61$, $SD = 0.26$ for “ageless”; $M = 4.67$, $SD = 0.21$ for “modern”; $M = 4.22$, $SD = 0.22$ for “old”). For religiousness, we neither could replicate the pattern on the complete dataset nor the findings by Rudolph et al. (2007). Of the original correlations, only the one between modern topicality ratings and religiousness remained significance on the reduced dataset ($r(43) = -.45 [-.74, -.02]$, $p < .01$ for “modern”; $r(43) = .31 [-.13, .66]$, $p = .04$ for “ageless”; and $r(43) = .08 [-.36, .49]$, $p = .61$ for “old”). An ANOVA with the demographic topicality also showed no statistically significant differences in religiousness ratings, $F(2, 42) = 2.41$, $p = .10$, $\eta^2 = .10$. Furthermore, the negative correlation between age and attractiveness remained on the reduced dataset ($r(43) = -.84 [-.94, -.65]$, $p < .01$). The same was true for the positive correlation between attractiveness and intelligence ($r(43) = .74 [.45, .89]$, $p < .01$). More importantly, on the reduced dataset the relationship between age and intelligence differed from the complete

dataset, thus older names were rated as less intelligent ($r(43) = -.45 [-.74, -.03]$, $p < .01$) in line with the findings by Rudolph et al. (2007).

In conclusion, this more direct comparison shows that both the larger set of names, which also included more uncommon names, as well as the different methodological approach to determine topicality caused the differences between our results and the ones reported by Rudolph et al. (2007). When we reduced the dataset to the names also used by Rudolph et al. (2007) the differences partially disappeared. Most importantly, the correlation between age and intelligence switched signs and was now in line with previous findings, although it was not statistically significant anymore. For the topicality ratings, the discrepancies also partially disappeared. In addition, when we switched from topicality ratings to demographic topicality, the pattern was much more in line with previous findings. The differences in our findings when using ratings versus when using demographics in combination with the initial comparison between these two sources supports our initial notions that demographics may sometimes differ strongly from participants' beliefs about these demographics.

Guidelines for Using the Provided Dataset

In this section, we provide guidelines on how to select names from our dataset, methodological pitfalls that may arise, and how to circumvent those. We also describe an R-package that may assist researchers in the process.

Choosing Similar Names

In a study on sex stereotypes in job interviews, a researcher might want present information on a job candidate who is either male or female and either competent or warm in an experimental design. Using our dataset, what is the most efficient method to select male or female names that differ most on the independent variables "competence" and "warmth" and that match on the many other variables that may relate to the dependent variable (e.g., perceived intelligence)? High dimensionality datasets often suffer from an effect referred to as the "curse of dimensionality" (Aggarwal, Hinneburg, & Keim, 2001; Beyer, Goldstein, Ramakrishnan, & Shaft, 1999). Without going into much detail, this term refers to a number of unexpected properties of high dimensionality spaces. Most importantly for the research presented here, in such a dataset the most similar (best match) and most dissimilar (worst match) to any given query (e.g., another name in the dataset) show only minor differences in terms of their similarity. Hence,

in "such a case, the nearest neighbor problem becomes ill defined, since the contrast between the distances to different data points does not exist. In such cases, even the concept of proximity may not be meaningful from a qualitative perspective" (Aggarwal et al., 2001, p. 421). Thus, the high dimensional nature of the dataset makes a search for similar names to any name ill defined. However, the curse of dimensionality can be avoided in case the variables show high correlations and the underlying dimensionality of the dataset is much lower (Beyer et al., 1999). In this case, the matching should be performed on a dataset of lower dimensionality, which approximates the original dataset. We constructed and tested such a dataset (details and quality metrics are given in <https://osf.io/hcx2v/>), which reduces the dimensionality to five dimension. The lower dimensionality variables are given as PC1 to PC5 in the dataset. Researchers who need to calculate the similarity of one or more names to each other are strongly advised to use these variables instead of the original variables.

R-Package for Name Selection

To give researchers a simple method for selecting names for their studies, we provide an open source R-package that allows to define criteria for the selection of names. The package can be downloaded at <https://github.com/agglueckner/GerNameR>. This section shortly sketches the main features of the package, interested readers should refer to the documentation included with the package for detailed examples. This package can either directly extract subsets of names based on the percentiles, for example, the 10% most familiar names, or the names which are, for example, both above the median in competence and intelligence. In addition, this package allows creating matched pairs of names from two different groups (e.g., male and female) based on their difference in ratings. The matching is based on the reduced dimensionality variables, but can also be tailored to include other ratings, to ensure that the names are both generally similar but more similar on a given dimension such as competence or warmth. To include any other characteristic, the weight with which this characteristic should be used can be set by the researcher. To match the names, the distance between all pairs is calculated with the given weighting, and then the names are paired such that the total distance between all pairs is minimized. The minimal weighted matching is identified using the Hungarian algorithm for bipartite matching (Hornik, 2018; see also Munkres, 1957).

In addition to creating a set of pairs of matching names, we also allow extracting of a set of names, with the same number of members from two groups. Again this set is created such that the overall difference between all names (not only between the two groups) is minimized, with the

additional possibility to give more weight to some characteristics if required for the experimental design. To find such minimal distance sets, a genetic algorithm is used with the distance used as the fitness function (Scrucca, 2019; see also Holland, 1975).

Using the Collected Variables as Control Variables

Variables may be used as control variables, for example, in a regression model to account for differences on dimensions for single names in a study. Including many or all variables that we report in this article may result in a failure of fitting regression weights due to high multicollinearity up to the point of exact multicollinearity if all variables are used. This multicollinearity reflects the fact, that the variables contain less information than one would expect given the number of variables (Goldberger, 1991). This again indicates that the actual number of meaningful dimensions we collected is much lower than the number of originally collected variables. The solution to problems of multicollinearity is therefore exactly the same as before, instead of using the original variables, researchers are advised to use the variables labeled PC1 to PC5 as control variables in any regression analysis.

Conclusion

We provide ratings on perceived demographic and social characteristics (e.g., sex, origin, familiarity, education, and intelligence) for a large set of 2,000 representative German names. The split-half reliability indicates that the reliability of these ratings ranges from very high values for more objective characteristics (sex, origin, familiarity, and age) to lower values for more subjective ratings such as warmth and competence. In addition, the correlation with similar ratings provided by other sources for sex and origin show that these ratings relate to external criteria in a meaningful way. Furthermore, a factor analysis on a subset of the ratings taken from a questionnaire about warmth and competence could show that these ratings collected for German names have a similar factor structure as the one that was found in previous studies using the same items for ratings of social groups (e.g., Asbrock, 2010; Fiske et al., 2007).

Considering the high number of names tested and the time-constraints of an online study, the number of items per name was limited. In the study we therefore focused on measures that we think are useful for research on stereotypes (Fiske et al., 2007). To give some more insight on which kind of association people typically have when presented with a name, we nonetheless included an open ended question. The answers to this question can further

inform researchers to plan which items to include in a possible follow-up study.

Due to collecting a large number of different names we were only able to collect relatively few ratings for most of the names. This may lead to the estimated means differing substantially from the population means for these names. In addition, low sample sizes are associated with inflated effect sizes (Gelman & Carlin, 2014) and false-positive results of hypothesis tests (Schönbrodt & Perugini, 2013). However, concerning the tests performed in this study, the small number of ratings per name is less problematic, because most of these tests were done on averaged ratings for each name. Thus, the degrees of freedom for these tests should be based on the number of names, not on the number of ratings per name. In addition, since averaging serves to remove noise, each value entered in the analysis carries less error than a single rating, thus leading to even higher true degrees of freedom. In fact, Table 1 shows that the effective sample sizes of all variables are much higher than the number of names. Therefore, by estimating the *dfs* for our tests based on the number of names, we could achieve conservative testing.

To our knowledge, the provided set of ratings is the most extensive to date. Therefore, this set of names may not only be used for studies where only a few names are given (e.g., Bertrand & Mullainathan, 2004; Moss-Racusin et al., 2012; Steinpreis et al., 1999), but also in studies that require a large number of trials with different names (Dorrough et al., 2017; Stevens et al., 2011). The representative total set of names furthermore allows generating representative subsets by random sampling with or without constraints (e.g., only names that are similar with respect to some dimensions).

Furthermore, since associated characteristics with names are subject to change (e.g., due to celebrities or other prominent figures carrying those names), it is important to repeatedly re-validate ratings of first names (Newman et al., 2018). By providing the complete source code to our survey software as well as all analysis scripts, we hope to provide an easy starting point for other researchers who are interested in replicating or extending our results.

Since we only collected ratings for German names by German native speakers, the collected dataset is specific to Germany and so its use for studies in other countries may be limited. Nevertheless, the methods and the software we use and provide as part of this research is created such that it may easily be adapted to other countries or languages as well.

References

- Aggarwal, C. C., Hinneburg, A., & Keim, D. (2001). On the surprising behavior of distance metrics in high dimensional space. In J. Van den Bussche & V. Vianu (Eds.), *Proceedings of the*

- 8th International Conference on Database Theory (pp. 420–434). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-44503-X_27
- Asbrock, F. (2010). Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology*, 41, 76–81. <https://doi.org/10.1027/1864-9335/a000011>
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94, 991–1013. <https://doi.org/10.1257/0002828042002561>
- Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is “nearest neighbor” meaningful? In C. Beeri & P. Buneman (Eds.), *Proceedings of the 7th International Conference on Database Theory* (pp. 217–235). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-49257-7_15
- Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig Corpora Collection – monolingual corpora of standard size. In M. Davies, P. Rayson, S. Hunston, & P. Danielsson (Eds.), *Proceedings of Corpus Linguistic 2007*. Birmingham, UK. Retrieved from https://www.researchgate.net/publication/228345579_The_Leipzig_corpora_collection-monolingual_corpora_of_standard_size
- Böhmer, R., Schütz, A., Rentzsch, K., Körner, A., & Funke, F. (2010). Are we looking for positivity or similarity in a partner's outlook on life? Similarity predicts perceptions of social attractiveness and relationship quality. *The Journal of Positive Psychology*, 5, 431–438. <https://doi.org/10.1080/17439760.2010.534105>
- Brosi, P., Spörkle, M., Welp, I. M., & Heilman, M. E. (2016). Expressing pride: Effects on perceived agency, communality, and stereotype-based gender disparities. *Journal of Applied Psychology*, 101, 1319.
- Brunswick, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217. <https://doi.org/10.1037/h0047470>
- Charter, R. A. (2000). Confidence interval formulas for split-half reliability coefficients. *Psychological Reports*, 86, 1168–1170. <https://doi.org/10.2466/pr0.2000.86.3c.1168>
- Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree – An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97. <https://doi.org/10.1016/j.jbef.2015.12.001>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd rev. ed.). Hillsdale, NJ: Taylor & Francis.
- Cohen, W., Ravikumar, P., & Fienberg, S. (2003). A comparison of string metrics for matching names and records. In AAAI (Ed.), *Proceedings of the workshop on Data Cleaning and Object Consolidation at the International Conference on Knowledge Discovery and Data Mining (KDD)*. AAAI. Retrieved from <https://www.cs.cmu.edu/afs/cs/Web/People/wcohen/postscript/kdd-2003-match-ws.pdf>
- De Leeuw, J. (2001). *Reproducible research. The bottom line* (Technical Report No. 2001031101). Los Angeles, CA: University of California. Retrieved from <https://escholarship.org/uc/item/9050x4r4>
- Dorrough, A. R., Glöckner, A., Betsch, T., & Wille, A. (2017). When knowledge activated from memory intrudes on probabilistic inferences from description – the case of stereotypes. *Acta Psychologica*, 180, 64–78. <https://doi.org/10.1016/j.actpsy.2017.08.006>
- Duden. (2007). *Das große Vornamenlexikon* [The big first name encyclopedia]. Mannheim, Germany: Brockhaus AG.
- Fiedler, K. (2011). Voodoo correlations are everywhere – not only in neuroscience. *Perspectives on Psychological Science*, 6, 163–171. <https://doi.org/10.1177/1745691611400237>
- Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11, 77–83. <https://doi.org/10.1016/j.tics.2006.11.005>
- Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82, 878–902. <https://doi.org/10.1037/0022-3514.82.6.878>
- Gebauer, J. E., Leary, M. R., & Neberich, W. (2012). Unfortunate first names: Effects of name-based relational devaluation and interpersonal neglect. *Social Psychological and Personality Science*, 3, 590–596. <https://doi.org/10.1177/1948550611431644>
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9, 641–651. <https://doi.org/10.1177/1745691614551642>
- Goldberger, A. S. (1991). *A course in econometrics*. Cambridge, MA: Harvard University Press.
- Greitemeyer, T., & Kunz, I. (2013). Name-valence and physical attractiveness in Facebook: Their compensatory effects on friendship acceptance. *The Journal of Social Psychology*, 153, 257–260. <https://doi.org/10.1080/00224545.2012.741629>
- Heyder, A., & Kessels, U. (2015). Do teachers equate male and masculine with lower academic engagement? How students' gender enactment triggers gender stereotypes at school. *Social Psychology of Education*, 18, 467–485. <https://doi.org/10.1007/s11218-015-9303-0>
- Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Oxford, UK: University of Michigan Press.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. <https://doi.org/10.1007/BF02289447>
- Hornik, K. (2018). *Clue: Cluster ensembles* [Computer software]. (R package version 0.3-56). Retrieved from <https://CRAN.R-project.org/package=clue>
- Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.
- Kasof, J. (1993). Sex bias in the naming of stimulus persons. *Psychological Bulletin*, 113, 140–163. <https://doi.org/10.1037/0033-2909.113.1.140>
- Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110, 675–709. <https://doi.org/10.1037/pspa0000046>
- Kristof, W. (1969). On the sampling theory of reliability estimation. *ETS Research Report Series*, 1969, i–10. <https://doi.org/10.1002/j.2333-8504.1969.tb00577.x>
- Kuhlmann, B. G., Bayen, U. J., Meuser, K., & Kornadt, A. E. (2016). The impact of age stereotypes on source monitoring in younger and older adults. *Psychology and Aging*, 31, 875–889. <https://doi.org/10.1037/pag0000140>
- Lütkenhöner, L. (2011). *Hat Julia aufgrund ihres Vornamens Wettbewerbsvorteile gegenüber Ayse und Chantal? Ein Experiment auf dem Beziehungs-, Nachhilfe- und Wohnungsmarkt* [Does Julia have competitive advantages due to her first name compared to Ayse and Chantal? An experiment in the relationship, tutoring, and housing market]. Diskussionspapiere des Instituts für Organisationsökonomik. Münster, Germany: Institut für Organisationsökonomik. Retrieved from <https://www.wiwi.uni-muenster.de/io/de/forschen/diskussionspapiere>
- Moreland, R. L., & Zajonc, R. B. (1982). Exposure effects in person perception: Familiarity, similarity, and attraction. *Journal of Experimental Social Psychology*, 18, 395–415. [https://doi.org/10.1016/0022-1031\(82\)90062-2](https://doi.org/10.1016/0022-1031(82)90062-2)
- Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of*

- Sciences* 109, 16474–16479. <https://doi.org/10.1073/pnas.1211286109>
- Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5, 32–38. <https://doi.org/10.1137/0105003>
- Newman, L. S., Tan, M., Caldwell, T. L., Duff, K. J., & Winer, E. S. (2018). Name norms: A guide to casting your next experiment. *Personality and Social Psychology Bulletin*. Advance online publication. <https://doi.org/10.1177/0146167218769858>
- Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35, 250. <https://doi.org/10.1037/0022-3514.35.4.250>
- Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48, 1205–1226. <https://doi.org/10.3758/s13428-015-0664-2>
- Postel, H. J. (1969). Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse [The Kölner Phonetik. A method for the identification of personal names based on Gestalt analysis]. *IBM-Nachrichten*, 19, 925–931.
- R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Rentsch, K., Schütz, A., & Schröder-Abé, M. (2011). Being labeled nerd: Factors that influence the social acceptance of high-achieving students. *The Journal of Experimental Education*, 79, 143–168. <https://doi.org/10.1080/00220970903292900>
- Revelle, W. (2019). *Psych: Procedures for psychological, psychometric, and personality research* [Computer software] (R package version 1.8.12). Retrieved from <https://CRAN.R-project.org/package=psych>
- Rudolph, U., Böhm, R., & Lummer, M. (2007). Ein Vorname sagt mehr als 1000 Worte [A name says more than a thousand words]. *Zeitschrift für Sozialpsychologie*, 38, 17–31. <https://doi.org/10.1024/0044-3514.38.1.17>
- Rudolph, U., & Spörrle, M. (1999). Alter, Attraktivität und Intelligenz von Vornamen: Wortnormen für Vornamen im Deutschen [Perceived age, attractiveness and intelligence as a function of first names: Word norms for first names in German]. *Experimental Psychology*, 46, 115–128. <https://doi.org/10.1026/0949-3964.46.2.115>
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47, 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- Schulz, K., Rudolph, A., Tscharaktschew, N., & Rudolph, U. (2013). Daniel has fallen into a muddy puddle—schadenfreude or sympathy? *The British Journal of Developmental Psychology*, 31, 363–378. <https://doi.org/10.1111/bjdp.12013>
- Scrucca, L. (2019). *Ga: Genetic algorithms* [Computer software] (R package version 3.2). Retrieved from <https://CRAN.R-project.org/package=GA>
- Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 41, 509–528. <https://doi.org/10.1023/A:1018839203698>
- Stevens, J. R., Volstorff, J., Schooler, L. J., & Rieskamp, J. (2011). Forgetting constrains the emergence of cooperative decision strategies. *Frontiers in Psychology*, 1, 1–12. <https://doi.org/10.3389/fpsyg.2010.00235>
- Torchiano, M. (2018). *Effsize: Efficient effect size computation* [Computer software] (R package version 0.7.4). Retrieved from <https://CRAN.R-project.org/package=effsize>
- Van der Leeden, R., Busing, F., & Meijer, E. (1997, April). *Applications of bootstrap methods for two-level models*. Paper presented at the Multilevel Conference, Amsterdam, The Netherlands.
- van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEXUK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67, 1176–1190. <https://doi.org/10.1080/17470218.2013.850521>
- Wickham, H. (2017). *Tidyverse: Easily install and load the “tidyverse”* [Computer software] (R package version 1.2.1). Retrieved from <https://CRAN.R-project.org/package=tidyverse>
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., & Woo, K. (2018). *Ggplot2: Create elegant data visualisations using the grammar of graphics* [Computer software] (R package version 3.1.0). Retrieved from <https://CRAN.Rproject.org/package=ggplot2>
- Winkler, I., Jonas, K., & Rudolph, U. (2008). On the usefulness of memory skills in social interactions: Modifying the iterated prisoner's dilemma. *Journal of Conflict Resolution*, 52, 375–384. <https://doi.org/10.1177/0022002707312606>
- Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In ASA. (Eds.), *Proceedings of the section on survey research methods* (pp. 354–359). Washington, DC: American Statistical Association. Retrieved from <https://www.researchgate.net/publication/243772975>
- Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible research. The R series* (pp. 3–30). Boca Raton, FL: Chapman and Hall/CRC.
- Xie, Y. (2018). *Knitr: A general-purpose package for dynamic report generation in R* [Computer software] (R package version 1.21). Retrieved from <https://CRAN.Rproject.org/package=knitr>

History

Received June 27, 2018

Revision received January 23, 2019

Accepted January 23, 2019

Published online August 20, 2019

Authorship

All authors were involved in all parts of the research. Tillmann Nett, Angela Dorrough, and Marc Jekel designed the study and collected the data. Tillmann Nett performed the data analysis and implementation of the R-scripts.

ORCID

Tillmann Nett

 <https://orcid.org/0000-0001-7816-8944>

Open Data

All materials used in the survey, Tests for gender differences in stereotype effects, a scree plot for 3 warmth and 3 competence items, as well as a construction of a lower dimensionality approximation of the original variables can be found in <https://osf.io/jepzp/>.

Tillmann Nett

FernUniversität in Hagen

Office C011

Universitätsstr. 33

58084 Hagen

Germany

tillmann.nett@fernuni-hagen.de