Survey Research Methods (2017)
Vol. 11, No. 4, pp. 435-450
doi:10.18148/srm/2017.v11i4.7214

ISSN 1864-3361
http://www.surveymethods.org

Name-based measures of neighborhood composition: how telling are neighbors' names?

Hanno Kruse
University of Cologne
Cologne, Germany

Jörg Dollmann
Mannheim Centre for
European Social Research (MZES)
Mannheim, Germany

Name-based ethnicity classification is a common tool in the sampling of minority populations. In recent years, however, it has become a popular technique to construct measures of neighborhood composition if more objective data are unavailable. In this article, we test the accuracy of such name-based measures of neighborhood composition, relying on the example of German neighborhoods.

Drawing upon previous research, we assert that ethnic groups differ as to how well they are identifiable via name-based classification. Moreover, the ethnic mix in neighborhoods varies systematically, the ethnicities of minority members residing in majority-dominated neighborhoods differing from those residing in minority-dominated neighborhoods. Taken together, these two notions imply that a name-based classification bias should be neighborhood-specific. Results indicate a tendency to overestimate majority shares in minority-dominated neighborhoods and slightly underestimate them in majority-dominated neighborhoods. All analyses rely on data from the "Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU)" as well as on neighborhood compositional data from local statistics of two German cities. The article closes with a discussion of potential strategies to cope with the name-based classification bias.

Keywords: name-based classification; neighborhood composition; assimilation bias; onomastics; Germany

1 Introduction

Much empirical research on migration and integration in Western countries relies on a clear-cut distinction between two groups; those who are foreign-born or whose ancestors are foreign-born (from here on "minority members") and those who are not (from here on "majority members"). Name-based classification – an approach to identify the group membership of persons via the ethnic origin of their personal names – is in this regard becoming increasingly important (for an overview: Mateos, 2007).

Originally intended to improve the sampling process in minority surveys, name-based classification today is applied for various purposes. Among the more frequent applications are measures of context composition. A growing interest in contextual characteristics as determinants of social action (for a recent overview on neighborhood effects: Sharkey & Faber, 2014) led to a rise in demand for compositional information on very fine-grained spatial scales (i.e., small-level neighborhood data). Given that measures of majority shares on these lower spatial scales are often not readily available, proxies derived from the ethnic origin of inhabitants' personal names have become popular alternatives (see, for example, Drever, 2004; Kruse, 2017; Kruse, Smith, van Tubergen, & Maas, 2016; Sager, 2012)).

The question arises, whether name-based classification techniques are really adequate to serve these new purposes (Gramlich, 2015). Despite their steady advancement in recent years (Humpert & Schneiderheinze, 2000; Mateos, 2007, 2014; Schnell et al., 2013a, 2013b), they remain estimations and may thus be subject to systematic bias. When comparing name-based classifications to those resulting from persons' reported countries of birth – as a more objective measure, misspecifications become apparent: false negative (i.e., minority members wrongly classified as majority members) and false positive (i.e., majority members wrongly classified as minority members) classifications are thereby both a matter of concern.

In this article, we investigate the exact nature of the potential bias that name-based approaches can exert in the construction of measures of context composition, more specifically of majority shares in neighborhoods.[1] Relying on the

Contact information: Hanno Kruse, Greinstr. 2, DE-50939 Köln (E-mail: hanno.kruse@wiso.uni-koeln.de)

[1] We focus on majority shares in neighborhoods in terms of res-

example of adolescents in Germany, we test our arguments in an ethnically diverse setting where the use of name-based approaches is common (due to the lack of adequate official data).

Our point of departure is an argument already established by previous research (Schnell, Trappmann, & Gramlich, 2014). Whereas the names originating from some ethnic groups are clearly distinct from those of the majority population (e.g., Turkish versus German in Germany), this dividing line can be harder to trace for other ethnic groups (e.g., Polish versus German), potentially even more so in subsequent immigrant generations. We demonstrate that indeed the probability of true or false classifications of minority members in Germany depends on the specific ethnic origin of a person as well as on his/her generational status. The analyses rely on data from the "Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU)" (Kalter et al., 2016): a representative sample of ninth grade adolescents for whom we have both information on their "actual" group membership (i.e., their own/parents' country of birth) and their full names. This allows us to apply name-based classification and directly assess its validity.

Subsequently, and as the main contribution of this paper, we demonstrate the potential consequences of such ethnic differences in classification accuracy for the construction of measures of context composition. Our argument accounts for the fact that some ethnic groups are more likely than others to live in majority-dominated neighborhoods. Areas with high majority shares attract different ethnic minority groups than areas with lower majority shares, thus yielding locally specific classification accuracies.

To substantiate our argument, we simulate the process of name-based classification in two German cities with sizable minority populations and compare the resulting majority shares in the neighborhoods to the "actual" compositions as reported by local statistics. Doing so allows us to infer that name-based approaches tend to overestimate the proportions of majority members in minority-dominated areas, while underestimating them in majority-dominated neighborhoods. Proceeding in this manner, we add to the present state of research by providing an encompassing view not only of the causes, but also especially of potentially problematic consequences of misspecification in name-based classifications for the construction of measures of context composition.

The structure of the remainder of the article is as follows: Section 2 provides a general overview of name-based classification approaches, followed by a discussion of why they are at risk of misclassification in section 3. Section 4 focusses on the consequences of possible misclassifications for the construction of measures of composition. Section 5 discusses the implications that follow from the considerations in the case of Germany. Section 6 lays out the analytical approach taken in the article, before introduction of the data and variables used in the analyses in section 7. The results are presented in section 8 and are summarised in section 9, together with a discussion of limitations and provision of practical guidance on how to cope with potential bias in name-based measures of composition.

2 A brief review of name-based classification approaches

Several recently developed techniques of name-based classification seek to determine a person's ethnicity based on information about the ethnic origin of his or her name (for an overview: Mateos, 2007). Whereas these techniques may differ in the number of targeted ethnic groups listed, in the size of the respective target groups, and in the number of unique fore-/surnames used in the reference lists, Mateos' overview of 13 studies reveals a general communality of all approaches. They classify persons in a target population as having a specific ethnic origin according to the ethnic origin of their names as reported in more or less exhaustive name reference lists. This so-called name-based classification is then validated by using a more objective measure of ethnicity, like self-reported ethnicity, country of birth, or nationality (Mateos, 2007, p. 249).

In contrast to such name-based classification procedures, Schnell et al. (2013a, 2013b) recently proposed another technique, which does not rely on complete names, but rather on substrings of consecutive characters in a name, so called n-grams. These n-grams are extracted from the target names, which themselves are then Bayes-classified according to the relative frequency of the n-grams within predefined lists of names from specific ethnic origins (Schnell et al., 2014). Therefore, and in contrast to the onomastic methods with more or less complete name lists, the n-gram method does not pretend that a specific name is German, Turkish, Italian, etc., but that some names (or better, some combinations of characters, i.e., the n-grams), are more frequent among persons with a specific nationality. Compared to previous approaches, the advantage of this method is that it is less prone to misspellings and variations of names in both sources given that it does not rely on complete names in either the target names or the reference list.

Despite these recent developments in the realm of ethnic classification, the prevailing and most popular method applied in the German context remains that based on complete names, the classification approach developed and continuously enhanced by Humpert and Schneiderheinze (from here

idents' (ancestors') country of birth, given the frequent use of this construct (Drever, 2004; Kruse, 2017; Kruse et al., 2016). This being said, things may look different for other ethnicity-related measures of context composition. For example, researchers explicitly interested in the share of residents with a German name in a neighborhood would not have to worry about name-based classification bias, at all.

on "HS approach"). While the studies reviewed by Mateos (2007) aimed at separating one or only a few ethnic groups from the rest of the underlying population, the HS approach uses much more comprehensive dictionaries comprising a large number of combinations of forenames and surnames and the respective probability that each empirically observed combination will have a specific ethnic origin. Using in total almost 2,200,000 surnames and 670,000 forenames results in almost 30,000,000 existing combinations of forenames and surnames together with their regional classification (for the general procedure see Humpert and Schneiderheinze, 2000; for recent developments see Humpert and Schneiderheinze, 2016).[2] Misspellings and alternative spellings of names are also part of the dictionary, aiming to reduce possible classification problems, and consequently contributing to the mere quantity of names listed in the dictionary. Due to this extensive database with name-group relationships, the HS approach has become the standard approach for name-based classifications in Germany (see Ersanilli & Koopmans, 2013; Kogan, 2012; Mammey & Sattig, 2002; Rother, 2005; Schenk et al., 2006). In the following, we will therefore concentrate exclusively on this approach.

3 Causes for misclassification

One aim of the above-mentioned approaches is to identify correctly actual members of ethnic minorities as such (true positive classifications) and actual members of the majority population as such (true negatives). However, like any estimation-based procedure these approaches do not yield results that align perfectly with empirical reality, which is why some majority members will be wrongly coded as minority members (false positives), while some persons who actually have an ethnic minority background will be wrongly identified as majority members (false negatives).[3]

3.1 Minority misclassification

Turning first to the reasons for misclassifying actual minority members as majority members (false negatives), things are rather clear. Whereas the causes may be manifold, almost all of them relate to common, assimilation-driven mechanisms (for an elaborated overview of the following arguments, see Schnell et al., 2014, p. 234). For example, intermarriage between (usually better-assimilated) members of the minority and the majority population may lead to the subsequent adoption of the majority spouse's surname by the minority member. This may be one reason for the misclassification of ethnic minority members as majority members. Given that females are still more likely than males to adopt their spouse's name, intermarriage will lead to misclassifications especially among female minority members and their binational children (Waters, 1989). Secondly, minority members' names may also be adapted in the course of their naturalization process. For example, given that a minority member's original forename is "Piotr", he might adjust it to its German equivalent "Peter" upon naturalization. A third assimilation-driven reason for misspecifications can be minority parents' naming of their children, influenced perhaps by their degree of assimilation (Lieberson, 2000). Better-assimilated ethnic minorities are usually more likely to provide their children with first names that are more similar to the first names common among the majority population (Becker, 2009; Gerhards & Hans, 2008, 2009), thus leading to greater ambiguity about the child's actual group membership. Finally, beside reasons related to minorities' degree of assimilation, their ethnicity may equally play a role in determining how successfully an approach can identify them as such. If minority members stem from regions with languages similar to that of the receiving country, misclassifications will more likely occur. This also holds true for minority groups from regions where names are common that are similar to those among the majority population, for example due to historical idiosyncrasies linking the sending and the receiving country. In the case of minority members in Germany, one example would be former German emigrants to South America whose descendants return to Germany, but also Ethnic Germans from Eastern Europe who migrate to Germany.

3.2 Majority misclassification

Turning to the erroneous classification of actual majority members as minority members (i.e., false positives), less is known about the causes. This is perhaps because false positives are less problematic in name-based sampling approaches, given that they do not lead to an omission of minority subsamples but only to increased survey costs due to inflated sample sizes (Schnell et al., 2014). However, both false negatives and false positives may have an effect with regard to the construction of context measures.[4]

[2] Consistent combinations of foreign fore- and surnames (e.g., a usual Turkish forename and a usual Turkish surname) are coded as minority members. Inconsistent fore- and surnames (e.g., Turkish forename and German surname) suggest that the persons are offsprings of a binational relationship and are as such usually treated as minority members. However, things may be different if a forename is of foreign origin but frequently used among majority members. The HS approach accounts for these cases. Consequentially, "Kimberley Müller" and "Justin Meier" would always be coded as majority members.

[3] The terms "false-positive" or "false-negative" simply mean that a person has a name with a distinct origin that does not fit to the migration history of the person. It is not the name that is coded as false-positive or false-negative, as a specific name may have a distinct origin and keeps it, but rather a person with a specific migration background or nationality who does not fit to the origin of the name that is coded wrongly.

[4] As we will see in the next section, they may thereby either add up to substantial bias or they may partially cancel each other out,

One reason for the misclassification of members of the majority population is that their families may in fact look back on an ancient, long-forgotten immigration history. This ancient minority status may often still manifest itself in the family names, but can no longer be assessed based on more objective measures such as nationality, (self-reported) ethnic identity and/or (grand-)parents' country of birth. Furthermore, the practice among the majority population of providing their children with unusual forenames with a foreign connotation may also lead to false positive classifications. Finally, intermarriage as outlined above may lead not only to false negative classifications, but also to false positives, especially if a domestic spouse adopts the name of the minority partner.

Given the arguments above, it becomes obvious that misclassifications will not occur at random; they are to be expected, especially among specific demographic groups. Persons categorized as minority members according to name-based classifications will – besides some wrongly coded majority members – mainly comprise actual minority members, with recent immigrants (i.e., first generation immigrants) showing lower error rates. In contrast, persons classified as majority members according to name-based approaches will largely comprise actual majority members, but will also include minority members, especially those who are ancestors of immigrants (i.e., second generation immigrants). Furthermore, minorities' ethnic background plays a decisive role when it comes to probabilities of correct specification. In the case of Germany, Ethnic Germans from the Former Soviet Union (from here on "FSU") and Polish minority members will show higher error rates than will culturally more distant ethnic groups, such as Turkish minority members or those from the Former Yugoslav Republic (from here on "FYR").

4 Consequences of misclassification for measures of context composition

To explore the consequences of the outlined misclassification for the construction of measures of context composition, we proceed in two steps. First, we inspect the consequences under the simplifying assumption that both majority and minority members are ethnically homogeneneous groups. Consequentially, their error rates are homogeneous, as well. In a second step, we then relax this assumption and introduce heterogeneity in minorities' ethnic background and immigrant generational status, thus allowing for heterogenous error rates.

4.1 Assuming homogeneous groups

Imagine a city inhabited by majority and minority members. The city consists of an arbitrary number of neighborhoods whose actual majority shares vary between 0 and 100%. Further, assume that we want to estimate each neighborhood's composition by means of name-based classification. Depending on the error rates with which we falsely identify actual majority members as minority members (i.e., false positive rate p_{fp}) and actual minority members as majority members (i.e., false negative rate p_{fn}), our estimation might either closely match the actual composition of the neighborhoods or differ from it substantially.

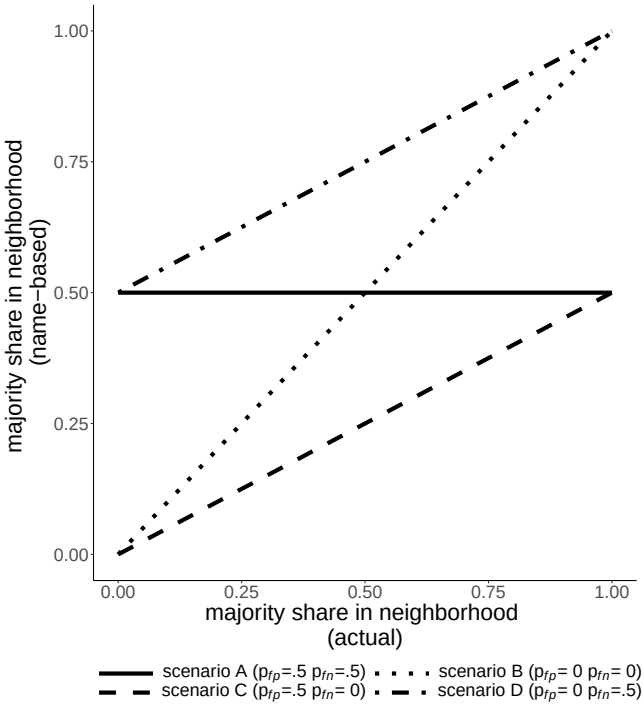

Figure 1. Hypothetical relations between actual and name-based majority shares in neighborhoods (with different error rates)

In this first step, we assume that both groups are ethnically homogeneous. As such, all members of a group face the same risk of misclassification. In other words, false positive and false negative rates are homogeneous, as well. Figure 1 visualizes the relation between actual (x-axis) and name-based neighborhood compositions (y-axis) based on four *scenarios A-D* with different, but homogeneous error rates (as indicated by the four different lines).

First, consider *scenario A* where name-based classifications are purely random. This holds if both false positive and false negative rates equal .5. What consequences would this have for the name-based neighborhood compositions? The answer is simple. Looking at a neighborhood exclusively inhabited by minority members, 50% of them would be classified correctly, whereas the other half would be misspecified as majority members. Similarly, in an all-majority neighborhood, 50% of all residents would be correctly identified as majority members, while the other half would be

such that the resulting bias is rather small.

misspecified as minority members. The same holds true for all mixed neighborhoods. Regardless of the actual neighborhood composition a name-based classification following scenario A would always yield an estimated majority share of 50% (see solid line).[5] Needless to say, a context compositional proxy based on name-based classification with these error rates would be worthless.

Second, turn to the other extreme, *scenario B*, where both the false positive rate and the false negative rate is zero ($p_{\text{fp}} = p_{\text{fn}} = 0$). In other words, there is no error in our classification whatsoever, such that all minority members and all majority members in all neighborhoods are correctly classified. Actual and name-based neighborhood compositions therefore align perfectly along the bisecting line (see dotted line).

In the third *scenario C*, false positive and false negative rates differ. Minority members are always correctly classified ($p_{\text{fn}} = 0$), whereas every second majority member is classified incorrectly ($p_{\text{fp}} = 0.5$). An all-minority neighborhood therefore would be correctly identified as such. The composition of an all-majority neighborhood, however, would be misspecified, given that 50% of its inhabitants would be falsely identified as minority members. The same holds true for mixed neighborhoods: considering again an evenly mixed neighborhood (i.e., 50% actual majority share) all minority members would be correctly classified, while half of the majority group would be mistakenly identified as minority members, thus leading to an underestimation of the majority share by 25 percentage points. In this third scenario the higher the actual majority share in a neighborhood, the stronger the underestimation of the majority share based on name-based classification (see dashed line).

Respectively, the exact opposite would be true in the final *scenario D* where majority members are always correctly classified ($p_{\text{fp}} = 0$) and every second minority member is classified incorrectly ($p_{\text{fn}} = 0.5$). The relation would then be as follows: the higher the actual majority share in a neighborhood, the better the estimation of the majority share based on name-based classification (dashed-dotted line).

To summarize, in contrast to scenario A both scenarios C and D would create estimations of context composition that clearly correlate with the actual context compositions. However, analyses based on such rather crude estimations could still lead to serious bias.[6]

4.2 Assuming heterogenous groups

So far, we have assumed that both groups, majority and minority, are ethnically homogeneous, all group members thus having the same error rates. Empirically, however, this is almost never the case. Turning to the example of German cities, the minority group actually consists of several ethnic groups, the largest of them being of Turkish, Polish, FSU, and Italian backgrounds. Concerning the various dimensions of integration these groups differ strongly (e.g., Kalter & Schroedter, 2010; Kristen & Granato, 2007). Therefore, and as argued in section 3, we expect error rates to be ethnically specific. It would thus be oversimplifying matters to assume that the false negative rate is the same across all minority members in German cities.

Relaxing this assumption leads to new open questions that need to be addressed in order to learn about how actual and estimated compositions of context relate: First, we need to specify the ethnic composition in our hypothetical city. Second, we need to make explicit where the different ethnic groups live, that is, whether an all-minority neighborhood entails the same ethnic mix of minority members as a neighborhood with only a modest share of minority members.

Assuming equal distribution of ethnic groups across neighborhoods, things would be rather simple: we could derive an overall false negative rate for all minority members from an (ethnic-group-size) weighted average of the ethnically specific error rates and proceed as outlined above. For example, assuming that half of all actual minority members in our hypothetical city are Turks, identifiable at an error rate of 5%, and the other half are Polish, at an error rate of 30%, the overall false negative rate would be $0.5 \cdot 0.05 + 0.5 \cdot 0.3 = 17.5\%$.

However, in real-world situations it seems rather unlikely that an all-minority neighborhood would entail the same ethnic mix as a neighborhood with only a modest share of minority members. Some ethnic groups – especially those of lower socioeconomic level – are more likely than other ethnic groups to live in minority-dominated neighborhoods (Janssen & Schroedter, 2007). From this perspective, we would need to specify not only the overall ethnic composition of the hypothetical city, but also the ethnic mix within each of the city's neighborhoods. When adding this further complexity to our hypothetical example, it becomes much less

[5] A neighborhood with an actual majority share of 50% would therefore be correctly identified, providing an example where false positive and false negative rates balance out each other.

[6] For example, imagine we want to know to what extent the majority share among actors' close friendships reflects the majority share in their local environments. Let us assume that in reality the majority share among actors' friends is equal to that in their actual neighborhood compositions. An analysis based on an estimated context composition as outlined in scenario D (dashed-dotted line) would come to different conclusions. Based on the biased measure, we would infer that actors located in all-majority contexts have friendships whose compositions closely match those of their local surroundings, whereas actors living in minority-dominated areas maintain more friendships with majority members than majority members are relatively present in their environment. This could lead, for example, to erroneous conclusions about ethnically specific friendship preferences of the latter. From this perspective, it seems important to know if and how exactly name-based context compositions deviate from actual compositions.

straightforward and more case-specific to derive the resulting relation between actual and name-based majority shares in the neighborhoods.

To see this take the following example: Assume that local statistics provide information about the respective composition of two neighborhoods A and B. Each neighborhood comprises 100 residents, among them 60 majority members. The actual majority share in neighborhoods A and B – as reported by local statistics – would thus be 60%. Neighborhood A further comprises 30 Turkish and 10 Polish minority members, whereas neighborhood B accommodates 10 Turkish and 30 Polish minority members. Finally, assume that the respective error rates at which actual majority members are misclassified via a name-based approach turned out to be 20%, for a Turkish minority member 10%, and for a Polish minority member 40%. The simulated, name-based majority share in neighborhood A would then be calculated as follows: $(60 \cdot 0.8 + 30 \cdot 0.1 + 10 \cdot 0.4)/100 = 55\%$. The name-based measure thus underestimates the actual majority share in neighborhood A by five percentage points. For neighborhood B, the name-based majority share would be $(60 \cdot 0.8 + 10 \cdot 0.1 + 30 \cdot 0.4)/100 = 61\%$, thus slightly overestimating the actual majority share by one percentage point. The example thus clearly demonstrates that the extent and the direction of bias in measures of context due to name-based classification is not simply a question of the error rates but also of the actual ethnic mix present in the contexts to be measured.

5 Implications for the case of Germany

In the previous section we repeatedly referred to the case of Germany to exemplify our theoretical arguments. Given that the minority group in Germany is rather ethnically diverse and growing (adolescents especially so) and name-based classification approaches are commonly applied here, it makes sense to base also the empirical tests of the arguments on the German case. We therefore summarize our expected empirical implications concerning the name-based classification of the population in Germany in the following.

Due to the discussed causes of misclassification in name-based approaches we expect false negative rates to vary systematically across specific demographic groups: first-generation immigrants should show lower error rates than those from the second generation. Further, error rates should be lower for minority members of Turkish and FYR background than those of Polish or FSU background should.

Concerning the formation of name-based measures of context composition, these generational and ethnic differences in error rates have important consequences, as they affect the resulting bias. Given that ethnic groups with a lowever socioeconomic status (i.e., Turkish and FYR) will be more likely than other groups (i.e., Polish and FSU) to reside in minority-dominated neighborhoods, the error rates in minority-dominated neighborhoods should be somewhat lower than in majority-dominated neighborhoods. The resulting form of bias, however, depends to a large degree on the exact neighborhood compositions and is thus mainly an empirical question.

6 Analytical approach

The analyses proceed in two consecutive steps. First, we test whether the error rates of a name-based classification vary across ethnic groups and immigrant generations, relying on the example of adolescents living in Germany. More specifically, we use a representative sample of 14-year-old adolescents living in Germany for whom we have detailed knowledge about their names as well as their (parents') birth-countries (for more information, see section 7.2). This information allows us to conduct a name-based classification according to the HS approach (see section 2.1) – thereby determining whether a respondent is a minority member or not – and to directly assess its validity. In other words, we compare the binary name-based classification to a classification according to respondents' (parents') birth-country (from here on actual minority background). In order to disentangle ethnic- from generational-specific classification error, we apply multivariate logistic models, regressing whether a respondent is misidentified or not (with two separate models: one containing all respondents and one including actual minority members only).

In a second step, we investigate the extent of bias that measures of context composition face when being constructed via name-based classification. To do so, we simulate the name-based classification process.[7] We do so for two German cities with sizable minority populations for which information on their actual ethnic neighborhood composition is available from local statistics.[8] To simulate the name-

[7] Alternatively, one could test the extent of bias by applying a name-based classification to all residents of an exemplary larger region or city and compare the resulting neighborhood compositions to the actual ones (i.e., those reported by local statistics). However, given that official register information of an entire city or region is not available to us and given that telephone directories may be subject to bias – including only a subset of households holding landlines – we opted for a cost-efficient, yet equally telling simulation approach.

[8] To extend the scope of our analyses, we repeated the simulations based on the "IRB" dataset which contains information on the ethnic neighborhood compositions from local statistics of not only two but more than 50 German cities (BBSR, 2016). This greater coverage, however, comes at a cost: the IRB data provide information on residents' nationality, but not on their actual minority status. When taking residents' nationality as a proxy, simulations yield results very similar to those based on the more exact data from the two cities. Nevertheless, for the sake of brevity we chose to present only the less encompassing, yet more exact simulation results based on the two cities.

based classification we proceed as described in the numerical example in section 4.2: for every empirical neighborhood of the two cities we calculate a name-based neighborhood composition conditional on the observed group sizes and the ethnic- and generation-specific classification error rates attained in our first analytical step. Finally, we compare the simulated, name-based neighborhood measure to the actual majority share in the neighborhood from local statistics. The extent of bias induced by name-based classification will thereby depend both on the estimated error rates as well as on the ethnic mix in the empirically observed neighborhoods. All analyses are carried out in R (v.3.2.3).

7 Data and variables

7.1 Data

As outlined, the analyses rely on a representative sample of 14-year-old adolescents living in Germany, more specifically on data from the "Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU)" (Kalter et al., 2016). CILS4EU is a representative, school-based panel survey carried out in England, Germany, the Netherlands, and Sweden in 2010/11, with subsequent yearly follow-up waves. The survey applied a stratified, three-stage sampling approach (CILS4EU, 2016, cf.). In the first stage, schools were chosen at random from nation-wide lists of all secondary schools in a country. Given that the main aim of CILS4EU is the investigation of the integration pathways of young minority members, schools with high minority shares were oversampled.[9] In the second stage, two ninth grade classrooms within the selected schools were chosen at random. Finally, in the third stage, all students within the chosen classrooms became part of the gross sample. The first wave of CILS4EU yields a net sample size of 18,716 adolescents attending 958 classrooms in 480 schools. The following analyses rely on data from the first wave of the German part of CILS4EU, encompassing 5,013 students in 144 schools and 271 classrooms. To account for bias due to the stratified sampling approach as well as for potential nonresponse bias we analyze the data using a combined design- and adjustment weight.

Besides the CILS4EU sample, we further make use of neighborhood compositional data from two German cities, Nuremberg and Berlin, when investigating the consequences of name-based approaches in the realm of measures of context composition. Nuremberg has one of the largest minority proportions among all German cities and Berlin accommodates the largest number of minority members in absolute terms, thus yielding sufficient variation in terms of majority shares in their neighborhoods. The data stem from local statistics and provide information on the respective ethnic composition of the two cities in the year 2015 (Amt für Stadtforschung und Statistik für Nürnberg und Fürth, 2015; Amt für Statistik Berlin-Brandenburg, 2015). The spatial scale is rather fine-grained, with an average neighborhood size in Berlin of $\sim 8,000$ residents, and in Nuremberg of $\sim 6,400$ residents (see Table B1 in Appendix B). In both cities, the information on residents' ethnic background is based on a combination of their (parents') nationality and country of birth (Böckler, Schmitz-Veltin, & Verband Deutscher Städtestatistiker, 2013). Several ethnic groups reported in the local statistics were combined into aggregate categories such that the final ethnic grouping closely matches that chosen in the CILS4EU data (see next subsection).

7.2 Variables

Respondents in the CILS4EU-survey are said to be "actual" minority members if they themselves or at least one of their parents was born outside of Germany. Otherwise we define them as being majority members.[10] Further, given that the names of all respondents were available to us, the group membership was additionally defined according to the name-based HS approach.[11] Taken together, the information on respondents' actual group membership status and their name-based group membership identifies those respondents misclassified by the name-based approach (i.e., false negatives and false positives). The dummy variable *error* contains this information and serves as the dependent variable in logistic regressions.

Two variables enter the logistic regressions as independent variables: respondents' *ethnic background* and their *immigrant generation*. The former variable, ethnic background, is derived based on respondents' and their parents' reported countries of origin (for more information, see Dollmann, Jacob, & Kalter, 2014). The German CILS4EU sample contains respondents from more than 100 different ethnic backgrounds. Facing this enormous ethnic diversity we chose to conflate the ethnic background variable to seven categories, among them the four largest ethnic groups in the data – Turkish, FSU, Polish, FYR – as well as two residual categories combining the remaining smaller groups (i.e., other Western and other Non-Western). The second independent variable, immigrant generation, distinguishes whether a respondent is a native, a second-generation immigrant (i.e., born in Germany, at least one parent born abroad), or a first-generation immigrant (i.e., born abroad him-/herself; see Dollmann et al., 2014).

[9] Concerning this article's aim, the oversampling of minority-dominated schools guarantees that the data entail a sufficient number of respondents in concentrated contexts – who, under purely random assignment, would be very sparse.

[10] In doing so, we classify minority members from the third generation as part of the majority.

[11] Whereas all other datasets used in this article are publicly available, information on respondents' names is restricted and thus cannot be accessed publicly.

8 Results

8.1 Different error rates across ethnic groups and immigrant generations

In the first step, we test whether the error rates of a name-based classification vary across ethnic groups and immigrant generations. We start by taking a descriptive look at how the HS approach classified the German CILS4EU sample into minority or majority members. Table 1 shows that ~ 42% of the students in the sample were classified as minority members. When applying weights to account for the stratified structure of the sample, we are still left with ~ 24% minority members.

Based on their reported countries of origin, we can ascertain whether the name-based approach classified the respondents correctly. Almost 88% of those classified as majority members are truly majority members (i.e., negative predictive value). Among those classified as minority members, the percentage of correctly classified respondents ranges lower at ~ 80% (i.e., positive predictive value). This yields a total accuracy of ~ 86%.

Next, we test whether false classifications are especially prevalent among specific demographic groups, as the laid-out causes for misclassifications suggest. Table 2 provides a first indication in this regard, showing the actual ethnic composition of the German CILS4EU sample. The sample's actual minority share is, at ~ 29%, five percentage points higher than the same share identified according to name-based classification. The name-based approach thus underestimates the minority share actually present. Moreover, ~ 34% of all actual minority members are identified incorrectly via name-based classification (i.e., false negative rate) whereas among actual majority members it is only ~ 7% (i.e., false positive rate).

A closer look at the ethnic subgroups reveals substantial variation in error rates within the minority group. In line with expectations, culturally more distant ethnic groups (Turkish, FYR, Other Non-Western) show very low error rates, partly even lower than those of majority members. In contrast, among respondents with a Polish minority background more than 70% were identified incorrectly as majority members. With error rates above 50%, Polish respondents are thus more likely to be classified incorrectly as majority members than correctly as minority members. Similarly high error rates are also present among respondents from FSU countries.

However, not all of these observed differences in misclassification may be ethnically specific. Second generation immigrants are probably harder to identify correctly than first generation immigrants. From this perspective, the observed ethnic differences may be due partly to the fact that some ethnic groups are dominated by immigrants recently arrived, whereas other groups are composed mainly

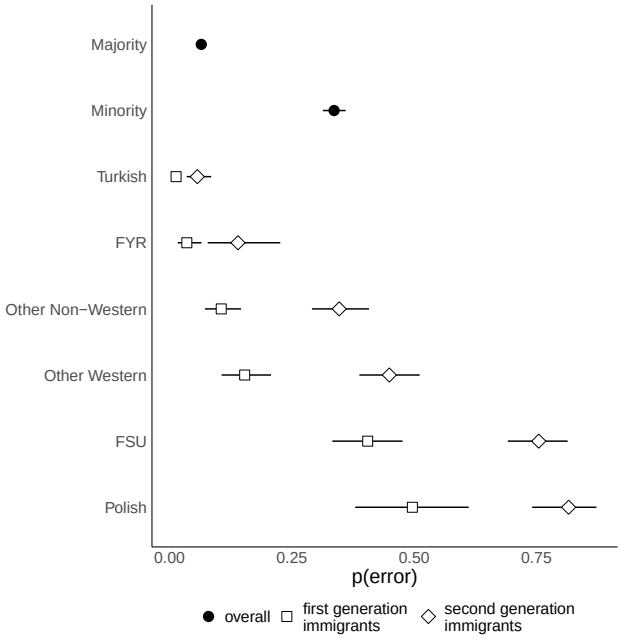

Figure 2. Predicted error rates in name-based classification (ethnic- and generation-specific)

of second-generation immigrants. To dissect ethnic from generational differences in misclassification we subsequently present results from multivariate analyses accounting for both attributes at the same time.

Results of the multivariate analyses are in line with expectations: the probability of incorrect identification varies substantially across ethnicities (see Figure 2). Majority members, Turkish and FYR minority members have the lowest probabilities of incorrect classification, while Polish and FSU minority members have clearly the highest. Moreover, first-generation immigrants are better identifiable than second-generation immigrants, which holds true for all ethnic groups, however differently pronounced.

A comparison of the patterns in Figure 2 to the gross ethnic differences reported in Table 2 reveals that some of the ethnic differences depicted in Table 2 are indeed due to compositional differences across ethnic groups in terms of immigrants' generational status. For example, the differences between FSU and Polish minority members seem to be largely attributable to the fact that second-generation immigrants are more prevalent in the Polish group. To summarize, the results corroborate our expectations concerning differential accuracies across ethnic groups and generations and are in line with findings of previous research (Schnell et al., 2013a, 2013b).

8.2 Resulting context compositional bias

In our second step, we investigate the extent of bias that measures of context composition face when being constructed via name-based classification. To do so, we first ex-

Table 1
Name-based classification of German CILS4EU sample (wave 1)

Classified as...	N (students)	Rel.freq. (unweighted, in %)	Rel.freq. (weighted, in %)	Correct (weighted, in %)
Majority	2,904	58.1	76.5	87.5
Minority	2,092	41.9	23.5	80.4
Total	4,996	100.0	100.0	85.8

Table 2
Actual composition of German CILS4EU sample (wave 1)

Actual group (true value)	N (students)	Rel.freq. (in %)	2nd-generation immigrants (in %)	Name-based error rates (in %)
Majority	2,609	71.5	–	6.5
Minority	2,387	28.5	78.4	33.6
Turkish	867	7.5	89.9	5.1
FSU	291	5.1	54.5	59.9
Polish	167	2.6	82.0	76.2
FYR	222	1.7	85.5	12.0
Other Western	363	5.6	84.6	40.4
Other Non-Western	477	5.9	75.0	28.5
Total	4,996	100.0	–	14.2

Notes: Data are weighted.

plore the actual ethnic mix present in neighborhoods in Germany. We then simulate the name-based classification for these neighborhoods – both under the assumption of homogeneous and heterogenous groups – and compare the resulting simulated majority shares to those actually observed in the neighborhoods.

Figure 3 provides information on the actual ethnic mix in German neighborhoods, more specifically in the two exemplary German cities. The ethnic mix in minority-dominated neighborhoods (first quintile, all neighborhoods with at most 56% majority members) differs substantially from that in majority-dominated neighborhoods (fifth quintile, all neighborhoods with more than 86% majority members). Turkish minority members make up a much larger share of all minority members in the former type of neighborhood than in the latter.[12] The opposite holds for FSU and for Polish minority members. As discussed, these different ethnic mixes of the neighborhoods may have important consequences for the bias induced by name-based classifications.

Before deriving the exact form of these consequences, however, we first inspect the induced bias under the simplifying assumption of homogeneous groups. In other words, we derive a first approximation of the induced bias based on the overall error rates of majority and minority members (see upper two point estimates in Figure 2).

Following these error rates, an all-minority neighborhood would be misspecified as having 34% majority members (i.e., minorities' error rate at 34%). Vice versa, an all-majority neighborhood would be identified as having 93% majority members, given that majorities' error rate ranges around ~ 7%. Taken together, the error rates induce a bias that is described in Figure 4. The dashed, black line in Figure 4 lays out how the actual and the simulated name-based compositions of context relate under the assumption of homogeneous groups. In the absence of any bias, the dashed, black line should perfectly overlap the bisecting line (see dotted, grey line), as this would imply name-based majority shares to mirror those actually present in the neighborhoods. Clearly, this is not the case: majority shares in mixed neighborhoods tend to be overestimated (i.e., dashed, black line ranging above the bisecting line), while it is underestimated in all-majority neighborhoods (i.e., dashed, black line ranging below the bisecting line).

What does the bias look like if we finally relax the assumption of homogeneity in groups and account for ethnically specific error rates and for differences in the ethnic mix in the neighborhoods? To see this, we take the ac-

[12] Further analyses also based on the German first wave of CILS4EU reveal very similar patterns in school compositions, with culturally distant ethnic groups being overrepresented in minority-dominated schools (analyses not shown here, available upon request).

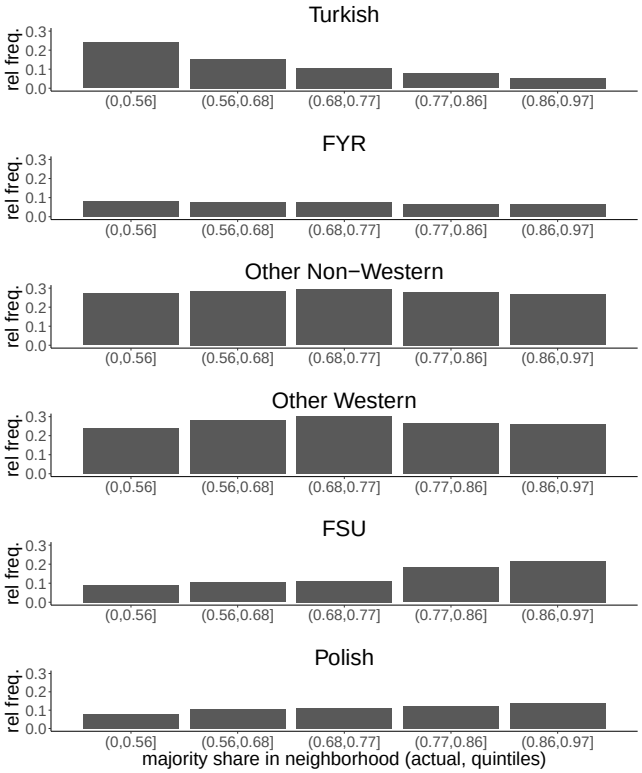

Figure 3. Ethnic mix in neighborhoods with different majority shares in two German cities.

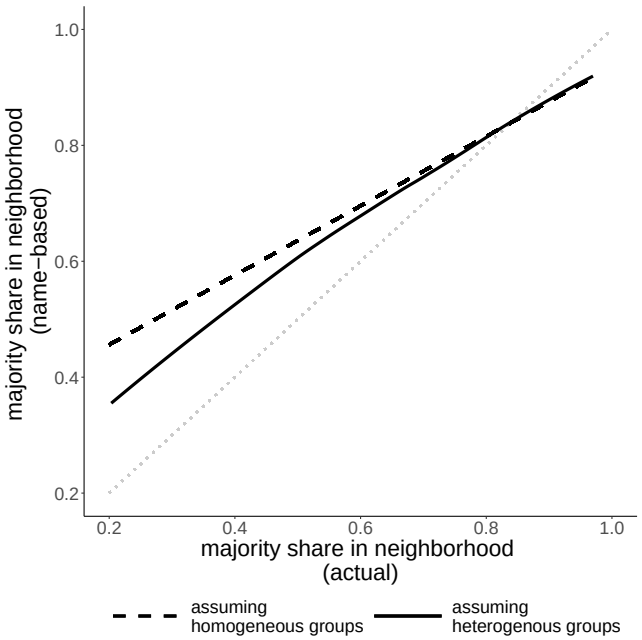

Figure 4. Actual and name-based majority shares in neighborhoods in two German cities. Trends across neighborhood compositions (black lines).

tual neighborhood compositions in the two cities[13] as exogenously given and simulate – based on the ethnically-specific predicted error rates derived in section 7.1 – the neighborhood compositions that would be yielded if we constructed the contextual measure via name-based classification.

The solid, black line in Figure 4 (i.e., curve with locally weighted scatterplot smoothing of simulated neighborhoods) describes the resulting relation between actual and simulated neighborhood compositions under the assumption of heterogenous groups. Again, the line deviates from the bisecting line, most strongly so in neighborhoods with lower majority shares.

If we compare the solid, black line to the dashed, black line, however, we see that the former is closer to the bisecting line. In other words, if we account for heterogenenous groups and for the actual ethnic mix in the neighborhoods the name-based induced bias is smaller than under the assumption of homogeneous groups.

9 Summary and Conclusion

In this article, we investigated the accuracy of name-based classification approaches among adolescents living in Germany. The main contribution of the article is its test of how systematic misspecification may lead to bias in measures of context composition, more specifically in majority shares in neighborhoods.

Our analyses suggested the following: First, error rates varied substantially across ethnic groups and across immigrant generations. Majority members and minority members of Turkish or of FYR background were usually classified correctly (i.e., they showed very low error rates). Polish and FSU minority members, however, were severely at risk of misclassification as majority members. Second, given that the two latter groups are rather sizable in Germany, the observed error rates proved to have important consequences for measures of context composition that rely on name-based classification. Neighborhoods with very high or low majority shares were subject to bias, with both types tending toward values that were more moderate. In contrast, neighborhoods with moderate majority shares were captured correctly. In other words, name-based measures of composition underestimated the variation present in majority shares across neighborhoods. Third, simulations showed that we overestimate the bias if we do not account for heterogeneity in minority groups.

What practical implications do these findings hold? Most importantly, name-based approximations of measures of con-

[13] In both cities, information is available on the actual ethnic compositions of the neighborhoods, but not on immigrants' generational status. We therefore constructed two simulated neighborhoods for every empirical neighborhood, one assuming all minority residents to be of the first generation, the other assuming all minority residents to be of the second generation.

text composition seem to work rather well in areas with moderate majority shares. Only extreme values are biased. They may thus be a useful and powerful option for researchers interested in context effects in Germany, in the absence of more precise information.

When applying name-based measures of composition, however, researchers should account for the bias they carry. There are several options to do so:

1. If there is information available about the target groups' error rates, one would optimally derive a correction factor from these error rates and apply it to the name-based measures of composition. Without full knowledge about the actual ethnic mix in the neighborhoods of interest, however, we can conduct this correction only under the assumption of homogeneous groups (i.e., based on the overall error rates of majority and minority members).[14] Our simulations demonstrated that we overestimate the bias under this simplifying assumption. Consequentially, researchers should be aware that applying such a correction factor would overcontrol the bias – at least so in the case of neighborhoods in Germany.

2. If no information on the target groups' error rates is available, but instead information about additional characteristics (beside their names) that correlate with the target group's ethnicity (e.g., age structure of households), use the latter to correct the name-based compositional data ex-post. Of course, this second alternative calls for additional theorizing, sophisticated modeling, and rich data, thus being rather cumbersome.[15]

3. Finally, if no additional information about the target population is available, the only option left is to be aware of the name-based bias when interpreting one's results.

The analyses have several limitations. First, we focussed on one specific name-based approach only, the HS approach. As laid out, various other approaches exist and it is unclear whether they would perform similarly. Nevertheless, we chose to concentrate on the technique most frequently applied in the German context. Second, name-based classification can provide approximations to various ethnicity-related measures of context composition. Here again, we restricted our analyses to one of the most frequently applied context measures: majority shares in neighborhoods in terms of residents' (ancestors') country of birth. Had we evaluated name-based classification based on other ethnicity-related measures of context composition they may have performed differently. Third, our findings are based on the example of a representative sample of adolescents living in Germany in the year 2010. It may well be that the name-based approach performs differently in other targeted contexts. Finally, the simulations relied on data from two specific German cities only. Other German cities may show a different ethnic mix in their neighborhoods than what we observed in the two cities. Repeated simulations, however – based on more encompassing though less exact administrative data – yielded substantively identical results (see Appendix C). We are thus confident that our findings hold beyond the two described cities.

Despite these different limitations, the findings of the article carry a general message for research on integration and contextual effects: name-based approaches are clearly a useful tool for ethnicity-related classification. However, if there are no means to correct for the bias induced, researchers should always be aware of their limitations when applying such measures in the realm of context composition.

References

Amt für Stadtforschung und Statistik für Nürnberg und Fürth. (2015). Bezirksdatenblätter Migrationshintergrund Nürnberg. Retrieved from https : / / www . nuernberg.de/internet/statistik/

Amt für Statistik Berlin-Brandenburg. (2015). Einwohnerinnen und Einwohner in Berlin mit Migrationshintergrund nach LOR und ausgewählten Herkunftsgebieten am 31.12.2014. Retrieved from https://www.statistik-berlin-brandenburg.de

BBSR. (2016). Innerstädtische Raumbeobachtung des Bundesinstituts für Bau-, Stadt- und Raumforschung auf Basis der Kommunalstatistiken der IRB-Städte / Statistik der Bundesagentur für Arbeit.

Becker, B. (2009). Immigrants' Emotional Identification with the Host Society. The Example of Turkish Parents' Naming Practices in Germany. *Ethnicities*, *9*(2), 200–225.

Böckler, S., Schmitz-Veltin, A., & Verband Deutscher Städtestatistiker. (2013). Migrationshintergrund in der Statistik – Definition, Erfassung und Vergleichbarkeit. Materialien zur Bevölkerungsstatistik, 2. Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-344959

CILS4EU. (2016). Children of Immigrants Longitudinal Survey in Four European Countries. Technical Report. Wave 1 – 2010/2011, v1.2.0. Mannheim: Mannheim University.

[14] A simple correction factor taking account of the overall error rates of minority members (p_{fn}) and majority members (p_{fp}) looks as follows:

$$p(\text{maj})_{\text{ac}} = \frac{p(\text{maj})_{\text{nb}} - p_{\text{fn}}}{1 - p_{\text{fp}} - p_{\text{fn}}}$$

$p(\text{maj})_{\text{ac}}$ being the desired, actual majority share in a neighborhood and $p(\text{maj})_{\text{nb}}$ being the name-based majority share respectively. Deriving this correction factor is straightforward, as we demonstrate in Appendix D.

[15] Professional geomarketing companies make use of this second option. One example would be the ethnic neighborhood compositions provided by the German company "Microm". For sociological applications of these measures see, for example, Kruse (2017), Kruse et al. (2016), Sager (2012).

Dollmann, J., Jacob, K., & Kalter, F. (2014). Examining the Diversity of Youth in Europe. A Classification of Generations and Ethnic Origins Using CILS4EU Data (Technical Report). *MZES Working Papers*, *156*, 1–46.

Drever, A. I. (2004). Separate Spaces, Separate Outcomes? Neighbourhood Impacts on Minorities in Germany. *Urban Studies*, *41*(8), 1423–1439.

Ersanilli, E. & Koopmans, R. (2013). The Six Country Immigrant Integration Comparative Survey (SCIICS): Technical Report. WZB Discussion Paper SP VI 2013-102.

Gerhards, J. & Hans, S. (2008). Akkulturation und die Vergabe von Vornamen. Welche Namen wählen Migranten für ihre Kinder und warum? In F. Kalter (Ed.), *Migration und Integration. Sonderheft 48/2008 der Kölner Zeitschrift für Soziologie und Sozialpsychologie* (pp. 465–487). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Gramlich, T. (2015). Migranten als Untersuchungsgruppe in der Stadtforschung – Nutzung von Namen in der Städtestatistik. *Stadtforschung und Statistik*, *2015*(1), 65–75.

Humpert, A. & Schneiderheinze, K. (2000). Stichprobenziehung für telefonische Zuwandererbefragungen – Einsatzmöglichkeiten der Namensforschung. *ZUMA-Nachrichten*, *47*, 36–63.

Humpert, A. & Schneiderheinze, K. (2016). Sprachliche Analyse von Personennamen mit dem "Onomastik-Verfahren". Retrieved from http://www.stichproben.de/

Janssen, A. & Schroedter, J. H. (2007). Kleinräumliche Segregation der ausländischen Bevölkerung in Deutschland: Eine Analyse auf der Basis des Mikrozensus. *Zeitschrift für Soziologie*, *36*(6), 453–472.

Kalter, F., Heath, A. F., Hewstone, M., Jonsson, J. O., Kalmijn, M., Kogan, I., & van Tubergen, F. (2016). Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU). GESIS Data Archive, Cologne, ZA5353 Data file Version 1.2.0. doi:10.4232/cils4eu.5353.1.2.0

Kalter, F. & Schroedter, J. H. (2010). Transnational marriage among former labour migrants in Germany. *Zeitschrift für Familienforschung*, *22*(1), 11–36.

Kogan, I. (2012). Potenziale nutzen! Determinanten und Konsequenzen der Anerkennung von Bildungsabschlüssen bei Zuwanderern aus der ehemaligen Sowjetunion in Deutschland. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *64*(1), 67–89.

Kristen, C. & Granato, N. (2007). The Educational Attainment of the Second Generation in Germany: Social Origins and Ethnic Inequality. *Ethnicities*, *7*(3), 343–366.

Kruse, H. (2017). The SES-Specific Neighbourhood Effect on Interethnic Friendship Formation. The Case of Adolescent Immigrants in Germany. *European Sociological Review*, *33*(2), 182–194.

Kruse, H., Smith, S., van Tubergen, F., & Maas, I. (2016). From Neighbors to School Friends? How Adolescents' Place of Residence Relates to Same-Ethnic School Friendships. *Social Networks*, *44*, 130–142.

Lieberson, S. (2000). *A Matter of Taste: How Names, Fashions, and Culture Change*. New Haven: Yale University Press.

Mammey, U. & Sattig, J. (2002). Zur Integration türkischer und italienischer junger Erwachsener in die Gesellschaft der Bundesrepublik Deutschland – Der Integrationssurvey des BiB. Materialien zur Bevölkerungswissenschaft 105a.

Mateos, P. (2007). A Review of Name-Based Ethnicity Classification Methods and Their Potential in Population Studies. *Population, Space and Place*, *13*(4), 243–263.

Mateos, P. (2014). *Names, Ethnicity and Populations. Tracking Identity in Space*. Heidelberg: Springer.

Rother, N. (2005). Wer zieht innerhalb der EU um und warum? Das PIONEUR-Projekt. *ZUMA Nachrichten*, *56*, 94–97.

Sager, L. (2012). Residential Segregation and Socioeconomic Neighbourhood Sorting: Evidence at the Micro-neighbourhood Level for Migrant Groups in Germany. *Urban Studies*, *49*(12), 2617–2632.

Schenk, L., Bau, A.-M., Borde, T., Butler, J., Lampert, T., Neuhauser, H., … Weilandt, C. (2006). Mindestindikatorensatz zur Erfassung des Migrationsstatus. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *49*(9), 853–860.

Schnell, R., Gramlich, T., Bachteler, T., Reiher, J., Trappmann, M., Smid, M., & Becher, I. (2013a). A New Method for Name-Based Sampling of Migrants Using N-Grams. German Record Linkage Center. Working paper series, 2013–04.

Schnell, R., Gramlich, T., Bachteler, T., Reiher, J., Trappmann, M., Smid, M., & Becher, I. (2013b). Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten. *Methoden, Daten, Analysen (mda)*, *7*(1), 5–33.

Schnell, R., Trappmann, M., & Gramlich, T. (2014). A Study of Assimilation Bias in Name-Based Sampling of Migrants. *Journal of Official Statistics*, *30*(2), 231–249.

Sharkey, P. & Faber, J. W. (2014). Where, When, Why, and for Whom do Residential Contexts Matter? Moving Away from the Dichotomous Understanding of Neighborhood Effects. *Annual Review of Sociology*, *40*, 559–579.

Waters, M. C. (1989). The Everyday Use of Surname to Determine Ethnic Ancestry. *Qualitative Sociology*, *12*(3), 303–324.

Appendix A

Table A1
Logistic regression results (dep. var.: error, complete sample)

	coef	s.e.
Intercept	−2.674	0.068**
Minority (ref.: *majority*)	1.994	0.088**
AIC	3,391.3	
N (students)	4,996	

Notes: Data are weighted.
* $p < 0.01$ ** $p < 0.01$

Table A2
Logistic regression results (dep. var.: error, minority members only)

	coef	s.e.
Intercept	−4.360	0.296**
Ethnic background (ref.: *Turkish*)		
Former Soviet Union (FSU)	3.967	0.288**
Polish	4.357	0.322**
Former Yugoslav Republic (FYR)	0.986	0.407*
Other Western	2.641	0.268**
Other Non-Western	2.203	0.272**
1st gen. immigrant (ref.: 2nd)	1.516	0.184**
AIC	1,297.5	
N (students)	2,387	

Notes: Data are weighted.
* $p < 0.01$ ** $p < 0.01$

Appendix B

Table B1
Neighborhood compositional data from local statistics (Nuremberg, Berlin)

	Nuremberg	Berlin	Combined
N (neighborhoods)	81	447	528
Neighborhood population			
mean	6,365.1	7,969.1	7,723.0
s.d.	3,838.5	5,319.5	5,149.7
Minority share	.422	.286	.303
Turkish	.061	.049	.050
Former Soviet Union (FSU)	.061	.031	.035
Polish	.038	.029	.030
Former Yugoslav Republic (FYR)	.036	.021	.023
Other Western	.165	.067	.080
Other Non-Western	.060	.088	.085

Appendix C

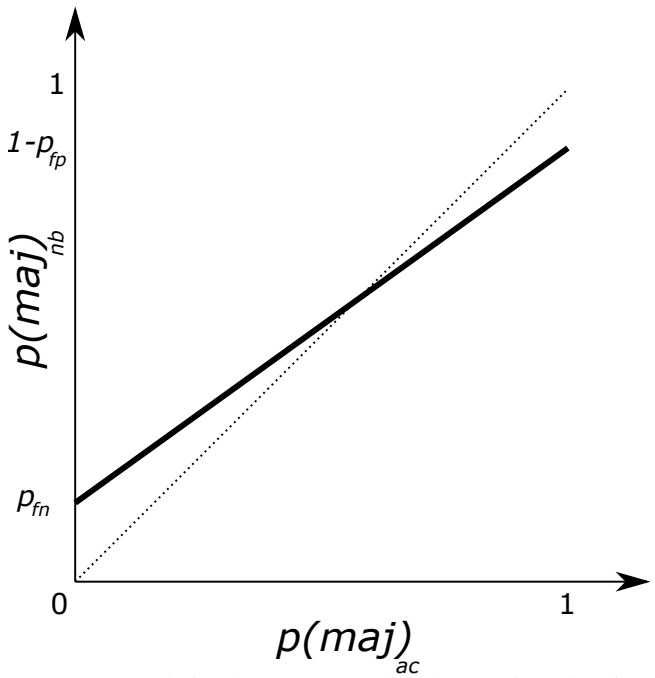

Figure C1. Relation between actual and name-based majority shares in neighborhoods given overall error rates of minority and majority members

We derive a correction factor based on the overall error rates of minority members p_{fn} and of majority members p_{fp} is a rather intuitive process. Assuming both error rates to be positive, we know that an all-minority neighborhood (i.e., $p(\text{maj})_{\text{ac}} = 0$) would be falsely identified as having a majority share of $p(\text{maj})_{\text{nb}} = p_{\text{fn}}$. Vice versa, an all-majority neighborhood (i.e., $p(\text{maj})_{\text{ac}} = 1$) would not be identified as such, but as having a majority share of $p(\text{maj})_{\text{nb}} = 1 - p_{\text{fp}}$. The resulting relation between name-based and actual majority share in the neighborhoods would thus look as depicted by the solid black line in Figure C1. It is easy to see that the function's intercept is p_{fn} and its slope is $1 - p_{\text{fp}} - p_{\text{fn}}$, yielding

$$p(\text{maj})_{\text{nb}} = p_{\text{fn}} + [1 - p_{\text{fp}} - p_{\text{fn}}] \cdot p(\text{maj})_{\text{ac}}. \tag{1}$$

Simple rearrangement leads to the correction factor:

$$p(\text{maj})_{\text{ac}} = \frac{p(\text{maj})_{\text{nb}} - p_{\text{fn}}}{1 - p_{\text{fp}} - p_{\text{fn}}}. \tag{2}$$

Appendix D

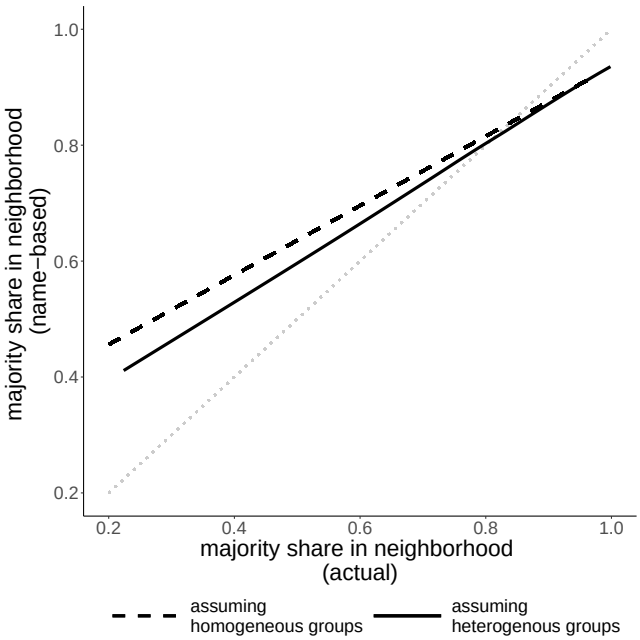

Figure D1. Actual and name-based majority shares in neighborhoods in the IRB data. Trends across neighborhood compositions (black lines).