Ethnic fractionalisation and social cohesion: the relation between immigration, ethnic fractionalisation and potentials for civic, collective action in Germany

Schaeffer, M.

_Citation for published version (APA):_
Chapter 3
Research Design

The EDCA-Survey

All empirical analyses of this thesis rely on the data of the Ethnic Diversity and Collective Action Survey (EDCAS), the conduction of which was my responsibility in the project “Ethnic Diversity, Social Trust and Civic Engagement” under the supervision of Ruud Koopmans at the Social Science Research Center’s (WZB) department “Migration, Integration, Transnationalization”. The project was financed by the German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ). A detailed description of the EDCA-Survey is given by Schaeffer et al. (2011). The EDCA-Survey is a large-scale computer-assisted telephone interview (CATI) survey and was conducted in Germany, France and the Netherlands. The current work focuses entirely on the German data, which was gathered from October 2009 to April 2010 and consists of 7,500 standardised telephone interviews. We explicitly designed the survey to test theoretical arguments on the effects of ethnic fractionalisation on social cohesion and the survey is thus suited to test hypotheses and less to make representative statements for the entire population of Germany. Instead, the survey population consists of persons over the age of 18 who have sufficient language skills to conduct an interview in German (or Turkish if they were sampled for the Turkish migrants oversample, see below) and reside in one of 55 selected sub-national regions. The aim of the survey is to enable the comparison of inhabitants of these 55 regions, which vary on contextual characteristics of interest.

The EDCA-Survey includes a 24 percent oversample of migrants in general (sample 2) and a 14 percent oversample of Turkish migrants in particular (sample 3). The oversampling is the same within each of the 55 regions, resulting in a final sample that consists of about 60 completed interviews with respondents from

\[ \text{Schaeffer et al. (2011)} \]

the general population, 26 additional completed interviews with respondents from the migrant population and 14 additional completed interviews with respondents from the Turkish migrant population in each of the 55 contextual unit. This adds up to 100 completed interviews per sampling point. Within five regions, 500 interviews with the same proportions were conducted. The overall number of interviews differs, because of minor deviations from the sampling plan that are described in Schaeffer et al. (2011). This survey design that encompasses an extensive oversampling of migrants, which is proportionally the same in all 55 sampled regions and consists of at least 38 migrants per contextual unit is an important characteristic of the EDCA-Survey and distinguishes it from other available data. In particular we designed the EDCA-Survey in this way to allow for the aggregation of contextual characteristics from the survey itself, which was for example done to build some of the ethnic fractionalisation indices investigated in Chapter 4. Table 3.1 summarises the sample of the EDCA-Survey.

<table>
<thead>
<tr>
<th></th>
<th>General Sample (60%)</th>
<th>Migrant Oversample (26%)</th>
<th>Turkish Migrant Oversample (14%)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>n per Large City</td>
<td>300</td>
<td>130</td>
<td>70</td>
<td>500</td>
</tr>
<tr>
<td>n per Region</td>
<td>60</td>
<td>26</td>
<td>14</td>
<td>100</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td><strong>4,552</strong></td>
<td><strong>1,898</strong></td>
<td><strong>1,050</strong></td>
<td><strong>7,500</strong></td>
</tr>
</tbody>
</table>

Due to its design, the respondents of the EDCA-Survey were sampled in three stages. First, 55 regions were sampled theoretically as well as randomly. In the second stage, telephone numbers were sampled, relying on random digit dialling for the general population and the oversample of the migrant population, in combination with an onomastic (surname-based) random sample from telephone books for the oversample of Turkish migrants. All telephone samples were extended with numbers of mobile phones from the telephone book (Häder and Häder 2009). Finally, the person who last had his or her birthday (and was at least 18 years of age) was sampled within the called household. This procedure was chosen to prevent an oversampling of populations who are more likely to be at home, such as housewives, elderly or unemployed (Diekmann 2003). In the following I will briefly discuss each of these stages.

The Regional Sample  Since the aim of conducting the EDCA-Survey was to investigate the contextual effects of ethnic fractionalisation, three goals had to be maximised in terms of the kind of contextual units to be compared, i.e. federal states, municipalities, census tracts or postal codes etc. One goal was to choose an operationalisation of region or context that is an empirically meaningful operationalisation of peoples’ everyday environments. The other goal was to choose an
operationalisation of region for which a rich source of publicly available context data exists, most importantly on ethnic backgrounds of the population. In addition, we had to be able to actually sample telephone numbers from these regions in order to prevent unfeasible screening costs.

In Germany, we chose “Land-” and “Stadtkreise” as operationalisations of regions, meaning that if I write about cities and regions in reference to the EDCA-Survey, I refer to “Kreise” of which there are 470 in Germany. Kreise have an average population of 187,749 inhabitants with a distribution that is right skewed for large cities like Berlin, Hamburg or Frankfurt, making the median population per Kreis 140,307 inhabitants. They correspond to level 3 administrative units of the Nomenclature of Territorial Units for Statisticians (NUTS 3). Kreise are an intermediate level of administration, situated between the federal states and local municipalities, but in contrast to these Kreise are not political units. Some Kreise are cities, however, and are as such congruent with municipalities, meaning that they are also political contexts that may have own integration policies as the examples of Stuttgart and Frankfurt am Main show. Kreise are thus the smallest contextual unit for which nation-wide comparable public data exists. Furthermore, Kreise are generally identifiable via pre-dialling codes, so that we were practically able to regionally stratify the survey. Figure 3.1 shows the sampled regions in Germany, with the colours denoting three different sampling strategies. Table B.1 on page 241 lists all regions included in the study along with the way they were sampled and whether they are a rural region or a city.

- Red is the colour for the four largest German cities (Berlin, Hamburg, Frankfurt and Munich) along with Duisburg. We chose to sample those five cities to ensure the societal significance of the empirical data. Immigration is an urban phenomenon and therefore central and important cities as well as Duisburg with its large migrant population and history of migration research were sampled. These are the five regions, in which 500 interviews were conducted.

- Yellow denotes cities and regions that we sampled theoretically on the independent variable. We sampled the two regions with the largest and smallest (at least 10% for reasons of feasibility) percentage of foreign nationals within each of the 16 federal states. Under this scheme 24 regions were selected, since some cities such as Berlin are federal states themselves and some had

2 These and most other contextual informations were derived from the Federal Statistical Office of Germany (www.destatis.de) and its data portal for regional statistics (www.regionalstatistik.de) respectively.
Figure 3.1: The Regional Sample in Germany
already been selected as one of the largest five cities. We would have preferred to rely on the percentage of migrants rather than foreign nationals. However, the regional data sources on migrants do not allow to distinguish between different countries of origin and thus give no information about the regional ethnic composition. This information was crucial, however to guarantee a reasonable number of Turkish migrants to live in each region as necessity for the Turkish migrants oversample. In each of these regions, 100 interviews were conducted.

- Blue characterises another 26 cities and regions that were sampled randomly, with drawing chances proportional to their population size. Cities and regions had to have at least 10% foreign nationals in order to be sampled, which is why there was a very low likelihood for any East German region to be sampled. Again 100 interviews were conducted in each.

The Three Samples  Within each of the 55 regions, three samples were drawn by relying on two telephone number sampling strategies. First, there is the sample of people from the general population above 18 years of age (sample 1). For this sample, telephone numbers were generated using random (last two) digit dialling (RDD) as suggested by Häder and Gabler (1998) and Gabler and Häder (1997) to conduct 60 interviews (300 in the seven large cities) per region with people above the age of 18. Respondents of this sample were only offered to conduct the interview in German.

The second sample is the oversample of migrants in general (sample 2). Migration background was defined as having at least one parent who was born outside the country of residence. Here numbers were also generated via random digit dialling in order to conduct 24 interviews per region (120 in the five large cities) with people above the age of 18 who have a migration background. At the beginning of each interview, six questions on the parents’ national origin were asked in order to screen out the native population (see “Questionnaire Design” below). Since these telephone numbers were also sampled using random (last two) digit dialling (RDD), this procedure resulted in very high screening costs, especially in those regions with low migrant populations. German was the only language for conducting the interview proposed to respondents of this sample.

Finally, there is the oversample of Turkish migrants (sample 3). This sample consists of randomly chosen numbers from the telephone book that are listed with a Turkish given or surname. The concrete list of Turkish names is the property (and business capital) of the Zentrum für Türkeistudien, which conducted the fieldwork, but generally consists of surnames such as Öztürk or Aslan. Relying on this sample, we conducted 14 interviews (70 in the five large cities) per region with Turkish migrants above the age of 18. Surname sampling from telephone
books is suboptimal. One cannot ensure to have an adequate list of surnames that are exclusively Turkish, that the Turkish names on the list are not correlated with social status and that the population of Turkish migrants is well represented in the telephone books. But as screening costs for sampling a certain migrant population are too high to be feasible, surname sampling seems to be the best alternative. Moreover, the first and second problems are relatively small, because particularly Turkish surnames are rather unique (Ersanilli, 2010; Granato, 1999; Salentin, 1999) and tend not to be correlated with social status (Berger et al., 2004). Respondents who were contacted for the surname-based sample had the choice to conduct the interview in Turkish or German. Roughly 81 percent of the respondents sampled via their last names chose to do the interview in Turkish. All interviewers who conducted interviews for the Turkish migrant oversample were perfectly bilingual so that they were able to conduct the interview in German or Turkish.

All samples included 20 percent mobile numbers, which were randomly sampled from the telephone book rather than by random digit dialling, because mobile numbers do not have a regional pre-dialling code that could ensure respondents to live in one of the 55 target regions. In German telephone books, mobile numbers are listed along with addresses of the potential respondents. The logic of these three samples is visualised in Figure 3.2. After the introduction, people went through a screening procedure (see “Questionnaire Design” below) on the basis of which they were classified as natives, migrants or Turkish migrants. Given this information, respondents were either screened out, or sorted to answer a native or migrant questionnaire.

Figure 3.2 also shows that the populations of the three samples overlap, with migrants and especially Turkish migrants also being part of sample 1. A person with a Turkish migration background could have been sampled for the sample of the general population above 18 years (sample 1), for the migrant population (sample 2), or for the population with a Turkish migrant background (sample 3). A person with a Polish migration background could have been sampled for sample 1 or sample 2, yet not for sample 3. For the analyses, these samples are pooled.

This sampling procedure resulted in an overall sample of which about 45 percent of the respondents have a migration background, with 29 percent being second generation migrants. Sixteen percent of the participants are of Turkish origin. These numbers are lower than one might expect, given that the oversamples alone should result in a share of 40 percent migrants. However, this is not due to a sample selection bias, since the general sample does include about 14 percent migrants. This number is lower than the overall share of migrants in Germany of about 19 percent, but this results from our oversampling of rather homogeneous for theoretical purposes. The unexpectedly low number of migrants is due to the impossibility
to reach the aims of reaching at least 24 respondents with a migration background and an additional 14 respondents with a Turkish migration background in many of the rural and particularly East German Kreise. The chances of sampling a migrant household via random digit dialling as well as the number of Turkish surnames in the telephone books were too small to fulfill the sampling frame in some of the regions. Instead more interviews with the general population, i.e. mostly natives were conducted in these regions. More information on the deviations from the sampling plan are given in Schaeffer et al. (2011).

Field Phase  The fieldwork in Germany was conducted by the Zentrum für Türkeistudien und Integrationsforschung (ZfTI)\(^3\) which is an institute of the University of Duisburg-Essen. The field phase of the German survey started on the

---

\(^3\)Zentrum für Türkeistudien und Integrationsforschung (ZfTI): www.zfti.de
6th of October 2009 and ended on the 22nd of April 2010. Overall, it took about seven months to complete the 7,500 interviews.

The ZfTI conducted computer assisted telephone interviews (CATI). The different regions were called in parallel so as to prevent a conflation of regional and time effects. 60 percent of the interviews had to be conducted during the evening or at weekends and at least 25 percent during the afternoon on working days. The time variable shows that actually 30 percent of the interviews were conducted before 5pm. The sampled telephone numbers were called 15 times before they were deleted from the sample. All interviewers received a special training, most of which I personally supervised, along with the first week of fieldwork. These interviewers who had foreign names took on native German names to introduce themselves in order to prevent social desirability biases.

**Response Rate and Questions of Representativity** Although our aim in conducting the EDCA-Survey was to test theoretical arguments and not to make representative descriptive statements, the response rate is still a telling indicator of possibly biased representativity within the 55 sampled cities and regions. Yet, response and cooperation rates, particularly of telephone surveys, have dwindled over the last years (e.g. Curtin et al., 2005; Schnell, 1997). Complex and costly face-to-face surveys that rely on samples drawn from public registries achieve response rates of 40 percent (Allbus) or 46 percent (ISJP). The response rates of telephone surveys suffer especially for reasons such as the high number of calls from marketing research agencies and other factors that drive down people’s willingness to respond, but also for technical reasons such as the fact that for those numbers where nobody answered, one cannot tell whether the contact was actually valid or invalid such as a company number. The EDCA-Survey is no exception to this trend with a response rate of about 13 percent, as can be seen from Table 3.2.

<table>
<thead>
<tr>
<th>General RDD Sample</th>
<th>Turkish Migrant Oversample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Percent</td>
</tr>
<tr>
<td>Not Eligible</td>
<td>25,160</td>
</tr>
<tr>
<td>Not Reached</td>
<td>18,384</td>
</tr>
<tr>
<td>Refused</td>
<td>20,308</td>
</tr>
<tr>
<td>Language Problems</td>
<td>645</td>
</tr>
<tr>
<td>Partial Interview</td>
<td>1,565</td>
</tr>
<tr>
<td>Completed Interview</td>
<td>4,731</td>
</tr>
<tr>
<td>Overall</td>
<td>70,793</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Response Rate</td>
<td>13.80%</td>
</tr>
<tr>
<td>Cooperation Rate</td>
<td>23.67%</td>
</tr>
</tbody>
</table>

*Response and Cooperation Rates include partial interviews.*

These response and cooperation rates are estimated for the general sample and
the oversample of Turkish migrants according to the suggestions of the American Association for Opinion Research (AAPOR).

For the oversample of migrants no response rate can be computed, because only after the initial screening phase was it possible to tell whether a person was eligible for the oversample. I count call attempts as a valid interview in case we know the respondent’s gender, year of birth (both questions were asked at the beginning of the interview) and whether the respondent has children (this question was asked at the end of the interview). Note that I cannot differentiate between “no contact”, “unknown if household is occupied”, a general “other” and an “unknown other” as AAPOR demands. I rather summarise all these under the term “no contact”, which results in the following two formulas:

\[
\text{Response Rate} = \frac{\text{Completed} + \text{Partial Interviews}}{\text{Completed} + \text{Partial Interviews} + (\text{Not Reached} + \text{Refused})}
\]

\[
\text{Cooperation Rate} = \frac{\text{Completed} + \text{Partial Interviews}}{\text{Completed} + \text{Partial Interviews} + \text{Refused}}
\]

While both response and cooperation rates seem very low, recent research on survey methodology questions low response rates to affect survey results (Curtin, Presser, and Singer 2000; Keeter et al. 2000) and suggests trade-offs in favour of large sample sizes that yield more estimation power (Davern et al. 2010). This was also the consideration of the EDCA-Survey. Overall, this means that the ECA-Survey might deviate from representativity for two reasons. First, the low response and cooperation rates might indicate bias resulting from systematic non-response. Second, the EDCA-Survey was never meant to be representative and might deviate from representative surveys because of the partially non-random sample of regions. Given the findings on the overevaluation of response rates, how severe is the deviation of the EDCAS survey from representativity?

Figures 3.3 and 3.4 compare the population of the EDCA-Survey to the population of the Scientific Use-File (SUF) of the German Micro Census of 2008 in terms of six central socio-demographic characteristics. In the absence of a German census, the Micro Census serves as an equivalent and thus as the standard for representativity. This is not unproblematic, because the Micro Census is itself a one percent survey of the German population with a sampling frame that relies on the results of the last German census of 1987. The Micro Census might be the best, but certainly not the gold standard. The comparison is twofold, with the upper graphs of each of the three rows showing the results for all respondents of the EDCA-survey (excluding those from non-randomly sampled regions) and for all respondents of the Scientific Use-File (SUF) of the Micro Census. Unfortunately, it is not possible with the SUF to identify the respondents’ Kreis, but only the federal state they live in. However, Berlin, Bremen and Hamburg are Kreise.

---

4American Association for Opinion Research (AAPOR): www.aapor.org
Figure 3.3: Comparison Between the German Micro Census and EDCAS

![Average Age](image1)

![Percent Employed](image2)

![Average Age](image3)

![Percent Employed](image4)

![Percent Women](image5)

![Percent Married](image6)

![Percent Women](image7)

![Percent Married](image8)

---

Legend:
- **German Micro Census 2008**
- **EDCAS**
and federal city-states at the same time, meaning that they can be identified in the SUF. The lower row shows results of the respondents living in Berlin, Bremen and Hamburg. The lower row is thus not prone to the EDCA-Survey’s deliberate deviation from representativity, by having drawn a partially non-random sample.

Figures 3.3 and 3.4 show six socio-demographic characteristics, namely age, employment status, gender, family status as well as education. Disregarding education for a second, the EDCA-Survey does not do bad for a CATI survey. There is the common tendency for an oversampling of singles, which is probably due to the fact that it is difficult to reach the one person who last had birthday in a household of many. But in terms of age, gender and interestingly also employment status, the means are fairly similar and their confidence intervals mostly overlap. This is true for both the sampled inhabitants of Berlin, Bremen and Hamburg as well as the comparison that includes all regions, which suggests that the partial non-randomness of the regional sample does not introduce a strong deviation from representativity. Note also that for the percent married, the divergence from the Micro Census is similar in pattern for the EDCAS population at large and the EDCAS subpopulation of inhabitants from Berlin, Bremen and Hamburg. This also speaks for the conclusion that the partly non-random design of the EDCA-Survey
is hardly deviating from representativity.

In terms of education, the picture looks different. Here we find the general pattern that people with little education are systematically refusing to participate in surveys [Diekmann 2003], which results in an oversample of highly educated respondents. This is most drastic for native EDCAS participants from Berlin, Bremen and Hamburg of which only 0.8 percent are lowly educated, even though it should be 11. This bias is similar in pattern for the population at large and for the interviewed inhabitants of Berlin, Bremen and Hamburg.

In general, the comparison to the Micro Census shows that the EDCA-Survey does fairly well, but shows the common pattern of an undersample of lowly and oversamples of highly educated and single respondents. The partially non-random design does not introduce any obvious divergence from representativity. For this work, I conclude that one should condition on (i.e. control for) common socio-demographic characteristics, such as education, gender and age as adjustments when working with the EDCAS data, but that special weighting procedures to adjust for the over-representation of certain regions are not necessary if one wishes to draw representative conclusions.

**Questionnaire and Variables**

**Questionnaire Design** The standardised questionnaire of the EDCA-Survey consists of four main parts, the introduction and screening, the main questionnaire, standard demography and finally survey experiments. The questionnaire starts with a section on participants’ migration background, because of the necessary screening procedure for the oversample of persons with migration background. All information on migration background or origin results from six questions. The questions pertain to both parents’ country of birth, whether they ever migrated to Germany if born abroad and whether they did so after 1951. The latter question is important to identify WWII refugees and expellees, who are in general not defined as immigrants in Germany. If these questions indicate that at least one parent was born abroad, we define the person in question to have a migration background. People who were born abroad, but have parents who were born in Germany are either third generation migrants or children of native Germans who live or used to live abroad. They are all coded as natives. Given this information, respondents were subject to a native or migrant questionnaire. The questionnaires only differ in a few regards such as questions on identification or experiences of perceived discrimination.

After the initial screening, the questionnaire dealt with the following topics:

---

[5]The standard demographic questions were posed as suggested by the German Federal Office of Statistics and the Leipniz-Institut für Socialwissenschaften (GESIS): www.gesis.org
Ahead of the fieldwork, the questionnaire was pre-tested, especially to check the quality of the newly designed items. The German contractor ZfTI conducted 50 telephone interviews in five of the sampled regions using random digit dialling. The basic population of the pre-test thus did not differ from that of the final EDCA-Survey. I personally supervised the pre-test interviews and also conducted interviews myself to gather first-hand experiences of the instrument.

We designed the main part of the EDCA-Survey questionnaire to focus especially on the neighbourhood level, because respondents can relate more easily to their neighbourhood than to any abstract region. Yet, it is not self-evident what neighbourhood means. It can be a spatially defined area (Sampson, 2006), a socially defined community (Tilly, 1973), or even vary among neighbours depending on how each of them engages with his environment on an everyday basis. We chose to rely on an individual spatial definition of the neighbourhood. Following the Detroit Area Study (DAS), all participants were told several times that neighbourhood refers to the area within ten minutes walking distance from their homes. This strategy has the advantage that respondents are asked about a spatial context that is meaningful to them in their everyday lives. The disadvantage of this approach is that individuals’ perceptions of their neighbourhood cannot be exactly compared to any objective characteristics of the regions as measured by public statistics. But the average responses of respondents within each region can be compared, because individuals are selected randomly within each region.

One aim of the EDCA-Survey was to use as few different scales as possible, so as not to confuse respondents on the telephone. Also we hoped familiarisation with the scales to enable faster answering by the respondents. Most questions had to be answered on a eleven point Likert scale that ran from 0 (no agreement, no trust, no

---

6Inter-University Consortium for Political and Social Research (ICPSR): www.icpsr.umich.edu/icpsrweb/ICPSR/studies/23820
identification etc.) to ten (total agreement, strong trust, strong identification etc.). The respondents were told several times, that zero means no agreement, no trust etc. and ten means strong agreement, strong trust etc. A frequency scale was also used often and ran from never, rarely, sometimes, often to very often. Very few questions, besides those asking about categories (family status, employment status etc.), diverge from this general pattern. More details on the concrete questions is given by Schaeffer et al. (2011).

**Dependent Variables** The different empirical chapters of this thesis draw on a wide range of dependent variables. In general, these are described in the single chapters, which is why I will not go into details here. Instead, I wish to discuss and justify my selection of social cohesion indicators. According to the definition of social cohesion given in the last chapter, I will rely on four indicators as dependent variables. For the cognitive dimension of social cohesion I will investigate trust in neighbours and collective efficacy as indicators. The EDCAS variable on trust in neighbours is identical to the measure Putnam (2007) uses.

> “Please indicate on a scale from 0 to 10, how much you trust the people in your neighbourhood”

Collective efficacy on the other hand was originally developed by Sampson et al. (1999) and is supposed to measure a community’s capacity to act collectively to solve neighbourhood problems. In contrast to trust, which might also be high in regions with a rather anonymous everyday life, collective efficacy particularly refers to solidarity and organisational capacities. Friedrichs and Oberwittler (2007) adopted this concept to fit the German context. Two items that are inspired by Friedrichs and Oberwittler’s work were posed to the respondents:

> “In neighbourhoods there are different problems. Let me give you some examples:

On a public green space lies bulky waste. On a scale from zero to ten, how likely is it that people from your neighbourhood would jointly try to find a solution?

In a dark alley several people have been mugged. On a scale from zero to ten, how likely is it that people from your neighbourhood would jointly try to find a solution?”

The two items correlate with 0.624 and mean values with listwise deletion of both items were used to construct the scale. I choose trust in neighbours and collective efficacy as dependent variables, because they directly refer to the neighbourhood, which was clearly defined for the respondents (see “Questionnaire Design”). Both measures thus relate to their everyday experiences in contrast to abstract concepts...
such as generalised trust. Moreover, all discussed theoretical explanations (see Chapter 2) are expected to work on the neighbourhood level, whereas the same cannot be said about larger contexts where inhabitants do not necessarily have face-to-face interactions. Table 3.3 shows the two variables descriptive statistics. Both measures are left skewed, as can be seen from the mean values and the density plots. But there is ample variance to be explained as the standard deviation shows.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>CV</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust in Neighbours</td>
<td>6.78</td>
<td>2.53</td>
<td>0.37</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Collective Efficacy</td>
<td>6.19</td>
<td>2.57</td>
<td>0.42</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

In both Chapters 4 and 5 additional analyses are performed to check the robustness of certain claims. These additional analyses rely on two further dependent variables, one being another indicator of cognitive social cohesion, the other being an indicator of failed collective action, which I however treat as a measure of cognitive social cohesion for its evaluative rather than behavioural nature. In particular, I will rely on neighbourhood satisfaction and respondents’ reported neighbourhood problems. *Neighbourhood satisfaction* was the final question on the general neighbourhood circumstances as a catch-all item:

“And considering all of the above, how would you rate your neighbourhood as a place to live on a scale from zero to ten?”

Reported *neighbourhood problems* on the other hand, goes back to Garofalo’s (1981) work on the broken windows theory, which states that residents observe their environment for socially deviant behaviour such as drug dealing or brawls.

\footnote{Note that all tables of regression outputs or descriptive statistics were generated with Jann’s “estout” Stata ado-file.}
Table 3.4: Descriptive Statistics of the Additional Indicators of Cognitive Social Cohesion

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>CV</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbourhood Satisfaction</td>
<td>7.77</td>
<td>2.00</td>
<td>0.26</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Neighbourhood Problems</td>
<td>0.76</td>
<td>0.79</td>
<td>1.05</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

(social disorder), as well as consequences of deviant be-heavier such as graffiti or broken windows (physical disorder). Observed disorder is then treated by persons as a sign of insecurity and insufficient social control, so that some will have increased fear of crime while deviant others will feel more secure to actually pursue their illegal activities (Xu et al., 2005). In short, disorder is a sign of failed collective action and of low capacities of a community to cooperate. In the EDCA-Survey respondents were asked two questions on their neighbourhood’s problems:

“How frequent do you face the following concrete problems in your neighbourhood? Is it never, rarely, seldom, sometimes or very often?

Waste lying about. Is that . . .

Harassment and verbal abuse. Is that . . .”

Table 3.4 shows the descriptives of these two additional social cohesion indicators. We see that they are structurally rather equivalent to trust in neighbours and collective efficacy. Note that the reported frequency of disorder is of course reverse in its meaning and is thus right instead of left skewed.

Next to these indicators of cognitive social cohesion, I analyse two indicators of the structural dimension of social cohesion. I here rely on the two most commonly used indicators of civic engagement, namely active membership in associations or initiatives on the one hand and voluntary engagement on the other. The instruments to measure civic engagement were replicated from the German
“Freiwilligensurvey” The battery asks about *membership in associations* by listing ten different social areas such as arts and culture or sports. Active membership was emphasised so as to exclude that people are members in associations, but do not participate in any activities on a regular basis. In particular, people were asked:

“There are many ways to get engaged as for example in an association or an initiative. I will now name different areas, in which people might engage. Please tell me if you are actively engaged in one or several of these areas. Are you actively engaged in the area of . . . ”

This battery was complemented by a question on whether the respondents perform *voluntary engagement* in any of their areas of engagement. This was explained to respondents as any kind of work for which they are not paid or receive only little financial compensation:

“Do you perform any voluntary work in any of the areas you are engaged in? With this we mean tasks and duties that you do for free or for only small compensation only.”

Table 3.5: Descriptive Statistics of the Indicators of Structural Social Cohesion

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>CV</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membership in Associations</td>
<td>0.51</td>
<td>0.50</td>
<td>0.98</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Voluntary Engagement</td>
<td>0.56</td>
<td>0.50</td>
<td>0.88</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Protest Participation</td>
<td>0.41</td>
<td>0.49</td>
<td>1.21</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.5 shows the descriptives of these two binary variables. Both variables are well suited for logistic regression, with mean values around 0.5. The table also shows the descriptives of another variable, namely participants’ protest participation. As above, some claims are tested via additional analyses, which in the case of behavioural social cohesion rely on protest participation as the dependent variable. In particular respondents were asked:

“Did you participate in any activities such as demonstrations, signature collections or fund-raising that had social or political goals within the last twelve months?”

In contrast to associational membership and volunteering, this dependent variable measures political mobilisation and a person’s aim to change social circumstances. It is thus well suited to test arguments on competitive collective action as discussed in the last chapter.

---

8Federal Ministry of Family Affairs, Senior Citizens, Women and Youth: www.bmfsfj.de/BMFSFJ/Service/Publikationen/publikationen,did=165004.html
Outlier Analysis  Given the complex survey design that consists of randomly and theoretically sampled contextual units with two regionally proportional oversamples of migrant populations, one might be concerned with the possibility of “social cohesion outliers” in relation to the cities’ and regions’ degree of ethnic fractionalisation. Such an analysis might suggest important control variables that should be included in later analyses. Figure 3.5 shows the aggregate relations of the two indicators of cognitive social cohesion to the degree of regional ethnic fractionalisation, as measured by the common Herfindahl-Hirschman index. The small table below the plots shows two highly significant aggregate level correlations of -.647 for trust in neighbours and -.576 for collective efficacy. The linear relations, which these correlation estimates express, are visualised as black lines (linear fit). Since five large cities have a sample size which is five times that of the other regions, the dashed lines show the linear fit weighted for these larger samples sizes. While the linear relation itself is the topic of later analyses, we assuringly see that these larger sample sizes hardly affect the relation. We also see that the randomly sampled regions, which are denoted as hollow circles, tend to fill the middle space between the extreme points of very homogeneous and rather fractionalized regions, even though a lot of randomly sampled regions are also rather homogeneous.

Besides the five large cities with their fivefold sample size, the plots also show outliers, which are defined in two terms. Either outlier regions are characterised by Cook’s $D > \frac{4}{n} = \frac{4}{55} = 0.073$ (Kohler and Kreuter [2001]). Cook’s $D$ considers both leverage, which is the abnormality on the dimension of the predictors, and discrepancy, which is the abnormality on the dimension of the dependent variable (Kohler and Kreuter [2001]). Alternatively there is also the measure of $DFBETA$, which denotes the degree to which the slope would change if a certain observation was not considered in the estimation. Values of $DFBETA > \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{55}} = 0.2697$ are seen as large (Kohler and Kreuter [2001]). All Kreise that lie beyond any of these two thresholds are shown in the plots.

There is at least one outlier for each dependent variable. For trust in neighbours, it is the city of Offenbach in Hesse with its roughly 120,000 inhabitants and a history as an industrial city that was known for its leather industry. It is the most diverse region and the city with the largest share of foreign nationals in all of Germany. Offenbach shows the lowest mean scores of trust in neighbours and collective efficacy. Yet, Offenbach is an outlier for its leverage, meaning its particularly high ethnic fractionalisation, because its discrepancy from the regression line is rather small. Still, Offenbach pulls down the regression line, as the $DFBETA$ suggests. This is balanced, however, by Munich, which similarly shows a high

---

There is at least one outlier for each dependent variable. For trust in neighbours, it is the city of Offenbach in Hesse with its roughly 120,000 inhabitants and a history as an industrial city that was known for its leather industry. It is the most diverse region and the city with the largest share of foreign nationals in all of Germany. Offenbach shows the lowest mean scores of trust in neighbours and collective efficacy. Yet, Offenbach is an outlier for its leverage, meaning its particularly high ethnic fractionalisation, because its discrepancy from the regression line is rather small. Still, Offenbach pulls down the regression line, as the $DFBETA$ suggests. This is balanced, however, by Munich, which similarly shows a high

---

The index is is described more thoroughly in Chapter 4. In this work, I rely on nationality to calculate the index.
DFBETA value, but pulls in the opposite direction. Munich with its 1.3 million inhabitants is Bavaria’s affluent and diverse capital. Its ethnic fractionalisation also stems from highly-skilled West-European migrants. In contrast to Offenbach, Munich has no particularly alarming Cook’s D value. Interestingly, not Offenbach is an outlier for collective efficacy, but the city of Rostock with 202,735 inhabitants. For its homogeneity, Rostock has a particularly low collective efficacy, which is true for some other regions as well, even though they do not surpass the outlier threshold. This means, Rostock has mediocre leverage but a strong discrepancy from its predicted value. This results in a problematic DFBETA value. A likely explanation is the regional unemployment rate, because Rostock and a couple of the other theoretically sampled regions from East Germany still suffer from the aftermath of reunification. Whereas the average unemployment rate for our 55 regions was about 8 percent in September 2009, it was 13.5 percent in Rostock.
West Germany is thus important.

Table 3.6: (Robust) Correlations Between Ethnic Fractionalisation and the Additional Measures of Cognitive Social Cohesion

<table>
<thead>
<tr>
<th></th>
<th>Nbh. Satisfaction</th>
<th>Nbh. Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic Fractionalisation</td>
<td>-0.537**, (0.449)</td>
<td>0.570**, (0.186)</td>
</tr>
<tr>
<td></td>
<td>-0.582**, (0.431)</td>
<td>0.558**, (0.197)</td>
</tr>
</tbody>
</table>

Standard errors in parentheses; + p < 0.1, * p < 0.05, ** p < 0.01.

In any case, Rostock and the other regions dampen the negative slope of the regression line, meaning they have a conservative effect, i.e. their influence on the regression line works against the hypothesis that ethnic fractionalisation is negatively associated with social cohesion. This conclusion is supported by the robust correlation coefficient. This coefficient is based on Stata’s “rreg” command, which implements a certain version of an M-Estimator that weights down influential observations (Jann, 2010). Standardised coefficients of binary regression models are nothing else than correlation coefficients, which is why the standardised coefficients of these robust regression estimates show correlation coefficients that are robust for the just discussed outliers. In line with my interpretation, the outliers hardly affect the estimates and mostly pull down the regression line for collective efficacy. Table 3.6 shows that these conclusions also hold for the additional indicators of cognitive social cohesion; robust regressions hardly alter the strength of the correlation to levels of ethnic fractionalisation. Furthermore, the heteroskedasticity one might be troubled with is later on taken care of by cluster-robust standard errors (see “Modelling Strategy” below).

Figure 3.6 shows the percent of actively engaged and volunteering respondents in relation to the degree of regional ethnic fractionalisation. The correlations, which again are the topic of later chapters where they are modelled with a logistic link-function, are weak and insignificant as the table below the plot shows. Besides much stronger variation around the regression lines, we see a positive association for engagement and a slightly negative one for volunteering. Again, the larger samples sizes of the five cities do not affect the relation too strongly, even though the negative relation in the second plot is somewhat strengthened. The theoretically and randomly sampled cities and kreise also show no fundamentally different pattern.

For structural social cohesion we see other outliers than before. In terms of associational membership there are three outliers. Magdeburg and Potsdam are homogeneous and low in membership, meaning that they disproportionately contribute to the positive slope, whereas Stuttgart is rather diverse and low in membership rates and hence biases the regression line in favour of the diversity hypothesis.
Figure 3.6: Ethnic Fractionalisation and Indicators of Structural Social Cohesion

Just like Rostock, Magdeburg and Potsdam are two East German cities have little leverage, but huge discrepancy from the regression line. But only Potsdam has a large DFBETA value. Also in population size they are rather similar to Rostock with 231,525 inhabitants living in Magdeburg and 156,906 in Potsdam. And again just like Rostock it seems likely that unemployment rate and East German history to explain the outlier status at least of Magdeburg which had an unemployment rate of 13.5 percent in September 2009. Potsdam on the other hand is average in terms of unemployment with a rate of 8.1 percent. Here it might rather be the location close to Berlin and the high levels of mobility and gentrification that the city has seen in recent years. Finally, there is the city of Stuttgart, which has around 600,000 inhabitants and is the capital of the prosperous federal state of Baden-Württemberg. In contrast to the East German outliers, Stuttgart had a low unemployment rate of 6.6 percent in September 2009 and is rather diverse, which stems from the formerly industrial demands for guest workers. Stuttgart is an out-
lier both in terms of low membership rates and high voluntary engagement rates. In any event, Stuttgart problematic DFBETA values for both dependent variables. This is probably the direct result of its leverage, resulting from its degree of ethnic fractionalisation, and its positive discrepancy from the regression line. But since in terms of membership rates Stuttgart’s influence is dampened by Magdeburg and Potsdam and in terms of voluntary engagement Stuttgart exerts an influence against the hypothesis, Stuttgart hardly seems to be a problematic outlier. This conclusion is further validated by the robust correlation coefficients, which hardly differ and if at all suggest that the outliers work against the hypothesis. The same conclusion holds for the additional indicator of structural social cohesion. Protest participation is only weakly correlated with ethnic fractionalisation (.11) and also a down-weighting of potentially influential cases does not alter this result (.13).

Both the general plots and the discussion of the outliers show that there is no clear-cut relation between ethnic fractionalisation and structural social cohesion. Next to the suggested phenomenon of competitive collective action, meaning that exactly for the conflict and feelings of mistrust fractionalisation causes, people start to get involved, other factors such as the unemployment rate might also work as suppressor variables. In any case, the theoretical sample as well does not seem to have a systematic impact on the relation.

**Control Variables** A general set of control variables for the different analyses does not exist, because control variables serve the purpose to adjust for systematic self-selection into treatment conditions of interest [Morgan and Winship 2007]. For this reason and for the fact that the different chapters were originally written as articles for different journals and were hence subject to different review processes, the set of control variables is explicated in each individual chapter. However, the discussion of the response rate above has shown that one should control for common socio-demographic characteristics when working with the EDCA-Survey. In addition, the discussion of outlier cases suggests the local unemployment rate to be a critical control variable that possibly accounts for the few and generally unproblematic outlier cases. For these reasons, many of the control variables are actually the same, including most importantly educational level, gender, age, local unemployment and population density. Table [3.7](#) shows the descriptives of all individual-level control variables used in the thesis.

Similarly, Table [3.8](#) shows the descriptives for context-level control variables. Here it is not quite clear, whether one should focus on the level of regions or of individual cases and therefore the table shows both. These context variables come from two different sources and relate to the year of 2009, in which the survey was started. All information on foreign nationals derives from the “Ausländerzentralregister”, which is a public agency where all foreign nationals need to register.
Table 3.7: Descriptive Statistics of Individual-Level Control Variables

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>CV</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>48.28</td>
<td>16.95</td>
<td>0.35</td>
<td>13</td>
<td>97</td>
</tr>
<tr>
<td>Years in the Nbh.</td>
<td>19.05</td>
<td>16.09</td>
<td>0.84</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>I-E nbh. Acquaintanceships</td>
<td>4.72</td>
<td>4.32</td>
<td>0.92</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Home Owner</td>
<td>0.45</td>
<td>0.50</td>
<td>1.11</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Employed</td>
<td>0.59</td>
<td>0.49</td>
<td>0.84</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>0.54</td>
<td>0.50</td>
<td>0.93</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Married</td>
<td>0.56</td>
<td>0.50</td>
<td>0.89</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Education: Low</td>
<td>0.05</td>
<td>0.22</td>
<td>4.24</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Middle</td>
<td>0.65</td>
<td>0.48</td>
<td>0.73</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>High</td>
<td>0.30</td>
<td>0.46</td>
<td>1.54</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Religion: Atheist</td>
<td>0.40</td>
<td>0.49</td>
<td>1.21</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Protestant</td>
<td>0.18</td>
<td>0.38</td>
<td>2.13</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Catholic</td>
<td>0.19</td>
<td>0.39</td>
<td>2.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Muslim</td>
<td>0.16</td>
<td>0.37</td>
<td>2.27</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>0.05</td>
<td>0.23</td>
<td>4.19</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Surname Sample</td>
<td>0.12</td>
<td>0.33</td>
<td>2.65</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Turkish Interview</td>
<td>0.10</td>
<td>0.30</td>
<td>2.98</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.8: Descriptive Statistics of Context-Level Control Variables

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>CV</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Foreign Nationals</td>
<td>10.94</td>
<td>7.16</td>
<td>0.65</td>
<td>1.10</td>
<td>29.71</td>
</tr>
<tr>
<td>Aggregate Level</td>
<td>8.42</td>
<td>6.18</td>
<td>0.73</td>
<td>1.10</td>
<td>29.71</td>
</tr>
<tr>
<td>∆%Foreign Nationals 07–09</td>
<td>0.07</td>
<td>0.92</td>
<td>13.57</td>
<td>-0.70</td>
<td>7.52</td>
</tr>
<tr>
<td>Aggregate Level</td>
<td>0.08</td>
<td>1.05</td>
<td>13.33</td>
<td>-0.70</td>
<td>7.52</td>
</tr>
<tr>
<td>Local Unemployment Rate</td>
<td>8.54</td>
<td>3.39</td>
<td>0.40</td>
<td>3.27</td>
<td>14.76</td>
</tr>
<tr>
<td>Aggregate Level</td>
<td>8.01</td>
<td>3.37</td>
<td>0.42</td>
<td>3.27</td>
<td>14.76</td>
</tr>
<tr>
<td>∆Unemployment Rate 07–09</td>
<td>-0.53</td>
<td>0.99</td>
<td>-1.88</td>
<td>-3.42</td>
<td>2.41</td>
</tr>
<tr>
<td>Aggregate Level</td>
<td>-0.54</td>
<td>1.11</td>
<td>-2.05</td>
<td>-3.42</td>
<td>2.41</td>
</tr>
<tr>
<td>Population Density</td>
<td>1.53</td>
<td>1.36</td>
<td>0.89</td>
<td>0.04</td>
<td>4.27</td>
</tr>
<tr>
<td>Aggregate Level</td>
<td>0.96</td>
<td>1.05</td>
<td>1.09</td>
<td>0.04</td>
<td>4.27</td>
</tr>
<tr>
<td>Crime Rate</td>
<td>0.09</td>
<td>0.04</td>
<td>0.43</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>Aggregate Level</td>
<td>0.08</td>
<td>0.04</td>
<td>0.44</td>
<td>0.03</td>
<td>0.16</td>
</tr>
</tbody>
</table>
While the EDCAS definition of persons with a migration background differs from that of persons of foreign nationality, I decided to rely on this data source for two reasons. First of all, this data source allows for the identification of different nationalities and thus for the calculation of fractionalisation indices. The German Micro Census, which is the only data source that allows to estimate representative shares of persons with migration background, has strong limitations on estimating shares of concrete migrant groups. Furthermore, the data of the “Ausländerzentralregister” are process-generated data and hence more reliable than any estimates from the Micro Census, which was not designed to estimate regionally representative shares of migrants. Both the local unemployment rate and population density are derived from the German Federal Statistical Office and its data base of regional statistics. Population density is measured as the number of inhabitants per square kilometre in thousands.

Modelling Strategy

In this work I generally estimate models that rely on a pooled dataset that includes all regions. The scatter plots have shown that the distributions of theoretically and randomly sampled cities and Kreise do not differ from each other in fundamental ways. I also generally estimate models based on the pooled sample of all respondents, including those of oversamples. By doing so, I control for migration background rather than sampling strategy, because I have shown the latter not to result in any significantly different responses (Schaeffer, 2011). As tests of robustness, however, Chapters 5 and 7 also report results of models that were separately estimated for different populations (mostly migrants and natives).

All estimated models take into account the clustered, hierarchical structure of the data. When observations are clustered and hence not independent it is inappropriate to use a standard OLS estimator, as Moulton (1990, 1986) has shown. Since the data I rely on are clustered in 55 regions and the analyses include context-level (Kreise) variables, I estimate linear regression models with cluster-robust standard errors. In comparison to OLS regression without clustering, these yield the advantage that the standard errors of parameters of both context-level and individual level regressors are not underestimated (Wooldridge, 2003; Williams, 2000; Liang and Zeger, 1986). Econometricians have developed estimators of cluster-robust standard errors for data “with a group structure—for example, the test scores of children observed within classes or schools. Children in the same school or class tend to have test scores that are correlated, since they are subject to some of the same environmental and family background influences” (Angrist and Pischke, 2009, p. 294). This motivation of econometricians to use cluster-robust standard errors parallels the use of multi-level models in other branches of the social
Cluster-Robust Standard Errors A regression model with cluster-robust standard errors is defined as:

\[ Y_{ig} = \alpha + X_{ig}\beta + \epsilon_i, \text{ with } \epsilon_{ig} \sim N(0, \Psi_g) \]

\( X_{ig} \) may include both variables measured at the individual and context levels. The decisive question is, however, what \( \Psi_g \) is. To explain this, I will derive the cluster-robust standard estimator from the well known OLS case (for this I rely heavily on Angrist and Pischke (2009), Fahrmeir et al. (2009) and Fox (1997)).

In matrix algebra, the OLS estimator for the \( \beta \) coefficients is defined as:

\[ \hat{\beta} = (X'X)^{-1}X'Y \]

Since we estimate \( \beta \) from a random sample, we assume \( \beta \) to vary if we repeated the same procedure many times. This variation of the \( \beta \) coefficients over an infinite amount of samples is the standard error. The variance covariance matrix \( \Omega \) encompasses this information. On the diagonal of \( \Omega \) we find the variance of the \( \beta \) coefficients, which is nothing less than the squared standard error. The estimator for the variance-covariance matrix \( \Omega \) of the \( \beta \) coefficients is defined as:

\[ \hat{\Omega} = (X'X_i)^{-1}(X' \hat{e}_2^2|X_i)(X'_iX_i)^{-1} \]

Under the assumption that the errors \( e_i^2 \) are independent and homoskedastic, specifically that \( E[e_i^2|X_i] \) is constant, \( E[e_i^2|X_i] = \sigma^2 \) is a constant. This means, we assume the same standard error for all observations. Mathematically, the expectation of a constant is the constant, which means the estimator of the variance-covariance matrix \( \Omega \) simplifies to

\[ \hat{\Omega} = \sigma^2(X'X_i)^{-1}(X'_iX_i)(X'_iX_i)^{-1} = \sigma^2(X'_iX_i)^{-1} \]

Of course, \( e_i^2 \) is unknown and unintelligible. Building on the Law of Large Numbers, we use the residuals to estimate \( \sigma^2 \), specifically, \( \hat{\sigma}^2 = \sum \frac{e_i^2}{n-2} \).

The assumption of homoskedastic errors is often unrealistic. We have seen a larger variation in associational membership rates where ethnic fractionalisation is low for example. For this reason White (1980) developed a heteroskedasticity-consistent estimator of \( \Omega \), by proposing to take the individual observation’s squared residuals, \( diag(e_i^2, \cdots, e_n^2) = diag(\hat{\psi}_i) = \hat{\Psi} \) as alternative to \( \hat{\sigma}^2 \). Since heteroskedasticity means that \( E[e_i^2|X_i] \) is not constant, we cannot simplify \( \Omega \), which results in the so called “Huber-White Sandwich Estimator”, where \( \hat{\Psi} \) is “sandwiched” between the regressor matrices, and is an asymptotically consistent estimator even if the errors are heteroskedastically distributed:
\[ \hat{\Omega} = (X'X)^{-1} (X'\hat{\Psi}X)(X'X)^{-1}. \]

Starting from here, it is possible to generalise the Huber-White Sandwich Estimator so that it also accounts for clustered data (e.g. Williams, 2000; Liang and Zeger, 1986). To do so, one builds cluster-level scores, by summing the individual-level scores within each cluster and building a matrix by taking their outer product. This results in the following form of \( \hat{\Psi}_g \):

\[
\hat{\Psi}_g = a \hat{e}_g \hat{e}'_g = a \begin{bmatrix}
\hat{e}^2_{1g} & \hat{e}_{1g}\hat{e}_{2g} & \cdots & \hat{e}_{1g}\hat{e}_{ng} \\
\hat{e}_{1g}\hat{e}_{2g} & \hat{e}^2_{2g} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \hat{e}_{ng-1,g}\hat{e}_{ng} \\
\hat{e}_{1g}\hat{e}_{ng} & \cdots & \hat{e}_{ng-1,g}\hat{e}_{ng} & \hat{e}^2_{ng}
\end{bmatrix}
\]

This matrix is then applied to the variance-covariance estimator:

\[
\hat{\Omega} = (X'X)^{-1} \left( \sum_g X'_g \hat{\Psi}_g X_g \right) (X'X)^{-1},
\]

where \( X_g \) are group-level regressors and \( a \) is an adjustment factor for the degrees of freedom. This means, the cluster-robust estimator of \( \Omega \) does not rely on the average squared residual as the common OLS estimator does, but on each specific group’s squared sum of residuals. Thereby the cluster-robust estimator corrects for the dependence between observations from the same group and accounts for heteroskedasticity between groups.

**Multi-Level Models** In comparison to cluster-robust standard errors, a simple multi-level model, is defined in matrix notation as:

\[
Y_{ig} = \alpha_0 + X_{ig}\beta + \nu_i + \eta_g, \quad \text{with } \eta_g \sim N(0, \sigma^2_g) \text{ and } \nu_{ig} \sim N(0, \sigma^2_i)
\]

where \( \eta_g \) is a random error component that captures variation between clusters and \( \nu_{ig} \) is a within-cluster random error component that captures left-overs (Gelman and Hill, 2007). \( X_{ig} \) includes a vector of ones for the estimation of a constant and may include both variables measured at the individual and context levels.

Another frequent interpretation of these models is that the model’s intercept \( \alpha_g \) varies randomly over the \( g \) groups, so that multi-level models are also called *random intercept models*. From this angle, two regressions, one on the individual level and one that explains the variance in the intercept,

\[
Y_{ig} = \alpha_g + X_{ig}\beta + \nu_i, \quad \text{with } \nu_{ig} \sim N(0, \sigma^2_i)
\]

\[
\alpha_g = \alpha_0 = X_g\beta + \eta_g, \quad \text{with } \eta_g \sim N(0, \sigma^2_g)
\]
are estimated simultaneously (e.g. DiPrete and Forristal 1994). These models are commonly estimated under a maximum likelihood framework or alternatively with a generalised least squares estimator, where the importance of the within and between part of the model is weighted by the inverse of their variance, whereas an OLS estimator would give equal weight to both parts (Baltagi 2008). As always in the universe of statistics, the more clusters the more reliable the estimates. While there is a lively discussion on how many clusters are needed for a multi-level analysis, textbooks suggest about 30 clusters for a normal random intercept model without random slopes and little interest the exact estimates of the random parameters (e.g. Hox 2010).

Comparison of Both Approaches  

Econometricians discuss multi-level models under the term random effects models (Angrist and Pischke 2009; Baltagi 2008), but tend to favour the use of cluster-robust standard errors rather than multi-level models: “In any case, here as elsewhere we prefer a “fix-the-standard-errors” approach to GLS” (Angrist and Pischke 2009, p. 309, footnote 10). This has two reasons. First, in contrast to multi-level models, “The clustered estimator is consistent as the number of groups gets large given any within-group correlation structure and not just the parametric model in (8.2.3) [the multi-level model, author’s note].” (Angrist and Pischke 2009, p. 313) or in other words, cluster-robust standard errors assume “no particular kind of within-cluster correlation nor a particular form of heteroscedasticity” (Wooldridge 2003, p. 134). This means, multi-level models decompose the error into a within and a between error component both of which are estimated by one variance parameter. Thereby they assume a similar homoscedastic within-error variance for all clusters, whereas cluster-robust standard errors do not. This latter point makes multi-level models more efficient, but at the same time it is a strong and nonconservative assumption. In my case, this would assume that the standard error in a city like Berlin is the same as that of a rural region like Oberallgäu.

Second, because the estimation of multi-level models weights within and between parts by the inverse of their variance, all clusters are weighted toward the grand mean if the cluster-level variance is small. More than that, clusters with only few observations are generally weighted toward the grand mean. This is called shrinkage or partial pooling (Rabe-Hesketh and Skrondal 2008; Gelman and Hill 2007). Shrinkage affects both the estimation of standard errors and $\beta$ coefficients of context and individual-level variables, because clusters that are weighted toward the grand mean have a different leverage. The fewer observations per cluster, the stronger do multi-level models weight toward the grand mean. Since the shrink-

\footnote{Econometricians tend to estimate multi-level models with generalised least squares rather than maximum likelihood estimators}
The age of multi-level models affects the estimation of the $\beta$ coefficients, multi-level models are frequently seen as biased, when clusters have only few observations (e.g. Moineddin et al., 2007). Others however, like Gelman and Hill (2007) see the shrinkage as an advantage rather than a bias, since we should not have much confidence in the information of clusters with only few observations. However, as my investigation of outliers has shown, there are hardly influential context units that might better be handled by an estimator that shrinks them toward the grand mean. Cluster-robust standard errors do not result in any such shrinkage. The regression coefficients of models with cluster-robust standard errors are similar to those of OLS models, only the standard errors are adjusted. This makes sense, given that even if observations are clustered the OLS estimator is still the best linear unbiased estimator (BLUE) for the $\beta$ coefficients (Angrist and Pischke, 2009; Baltagi, 2008). This can also be seen from the discussion above, where only the variance-covariance matrix is affected by assuming independent and homoskedastic standard errors or allowing for heteroskedasticity and within-cluster variation. In general, multi-level models demand both many observations per cluster in order to reduce shrinkage, and many clusters in order to estimate context effects precisely, whereas cluster-robust standard errors only rely on the latter. As for the number of clusters, the EDCA-Survey has 55 regions surveyed in Germany and with “51 clusters, you are on reasonably safe ground” (Angrist and Pischke, 2009, p. 323). In any case, this second point is less important than the first, given the large number of observations within each region that the EDCA-Survey encompasses.

For both these reasons, but mainly the first, I chose to follow the econometricians’ lead and rely on cluster-robust standard errors in this work. How cluster-robust standard errors are computed in Stata, the statistical software package I rely on, is discussed by Rogers (1993).