Chapter 5
Methods

Abstract This chapter discusses the challenges of drawing causal inferences from observational data, outlines the rationale behind the choice of statistical models used to test theory's implications, and discusses the data.

The first part of the chapter discusses the problems of drawing causal inferences from observational data. Building on the counterfactual approach to causality, I discuss the fundamental problem non-experimental research faces when attempting to draw causal inferences, i.e. unobserved heterogeneity or, in other words, the lack of an appropriate control group. Techniques of longitudinal data analysis, however, offer a partial remedy for this problem. I propose to combine the advantages of fixed effects and random effects regression models by estimating hybrid models (also referred to as correlated random effects models in the econometric literature).

The second part of the chapter presents the study's empirical basis. I discuss the data source, the German Socio-Economic Panel (SOEP) and the operationalization of the theoretical constructs. The SOEP is a representative, on-going, longitudinal household survey in Germany. It comprises several subsamples that especially sample immigrants and is thus particularly suited to investigate immigrant integration. It encompasses not only a rich set of question-items relating to aspects of immigrant integration, but also several items relating to transnational involvement, that are visits to the country of origin, duration of such visits, sending remittances, as well as the amount remitted.

Keywords Longitudinal analyses · Longitudinal data · Panel data · Fixed effects · Random effects · Unobserved heterogeneity · Hybrid models · Correlated random effect models · SOEP · Germany

The previous chapter was devoted to developing a theoretical model which allows us, on the one hand, to understand and explain the relationship between immigrant integration and transnational involvement and, on the other hand, to infer concrete hypotheses on this relation. The next step now consists of testing the model by investigating if its hypotheses conform to reality or if they are falsified. As the hypotheses specify cause and effects, the interest therefore is in drawing causal inferences. However, drawing causal inference from non-experimental, i.e. observational, data is a difficult task; some even maintain that it is impossible (Kempthorne 1978). As Hausman and Wise (1981, p. 365) put it, "[u]nbiased parameter estimates, although

R. Schunck, *Transnational Activities and Immigrant Integration in Germany*, International Perspectives on Migration, DOI 10.1007/978-3-319-03928-2_5,

illusionary, are thought by many researchers to be the primary objective of empirical analysis in the social sciences." Whether or not we see this claim to be true, there are some fundamental problems we have to be aware of when we attempt to draw causal inferences from observational data. Ignoring these issues will for sure lead to biased and thus incorrect inferences.

The most fundamental threat to causal inferences in non-experimental research is endogeneity. In a broad sense, endogeneity is given if we cannot determine consistently what is the cause and what is the effect. This problem may originate from *unobserved heterogeneity*, *simultaneity*, and *measurement error*. Unobserved heterogeneity arises if cases differ with respect to unobserved characteristics that are correlated with the observed characteristics and the outcome. Neglecting to control for these characteristics means that any relationship that is found between an independent and a dependent variable is likely to be biased, because these omitted variables may determine that relationship. To give a practical example of unobserved heterogeneity, consider that we are interested in estimating the impact of transnational activities on a migrant's receiving country language skills. If we estimate this without controlling other relevant characteristics, such as education, years of residence, ability, etc. the estimate of the causal effect is likely biased, because these unobserved characteristics are prone to influence both transnational activities and language proficiency. The notion of unobserved heterogeneity thus amounts to observations being conditionally different, i.e. heterogeneous, in terms of unobserved characteristics in ways that are unaccounted for in the estimation. Unobserved heterogeneity is sometimes discussed in a different terminology that is *selection bias* or *omitted variable bias*. Simultaneity, also known as reverse or reciprocal causation, describes a situation in which the value of the dependent variable in fact causes the value of the independent variable. In our example, this corresponds to a situation in which the level of receiving country language proficiency causes transnational activities. Lastly, measurement error arises if the characteristics we are interested in are imperfectly measured. The consequences of measurement error are unfortunately not straightforward and depend on the specified statistical model. In a simple case—in the context of a linear regression—measurement error creates downward biased estimates. This downward bias is known as the attenuation bias. In nonlinear regression models, the direction of the bias is not that straightforward. We will pick up this point at the end of the chapter.

It is quite likely to encounter all of these problems at once in observational data analysis. While simultaneity cannot be easily solved with tools of data analysis—one has to rely on a strong theoretical argument to justify the supposed direction of a relationship—there are methods which to some degree remedy the problem of unobserved heterogeneity and measurement error. Unfortunately, there is a trade-off between correcting for measurement error and controlling for unobserved heterogeneity. Both problems are severe. One might argue that we always face the problem of unobserved heterogeneity; no study possibly includes *all* relevant measures. At the same time, measurement error is likely, but not necessarily given. Moreover, we can aim at reducing measurement error by developing better methodology for data collection, while some unobserved heterogeneity is likely to remain even in face of

perfect data collection. There are some concepts that we just cannot measure. Therefore, we can make a cautious claim that unobserved heterogeneity is the bigger of the two problems; although there are scholars who would disagree with this stance. Since the choice of methods is very much motivated by these problems, the following section will discuss causality and how unobserved heterogeneity threatens causal inferences. It might appear as a lengthy discussion, but the choice of methods should be well justified. Far too often, it appears that methods are used without sufficient reflection on their adequacy.

5.1 Causality

In recent years, the so-called "counterfactual approach to causality" has become very popular in the social sciences, foremost in economics, and is becoming increasingly popular also in sociology (Brand and Halaby 2006; Caliendo and Kopeinig 2008; Gangl and DiPrete 2004; Morgan and Harding 2006; Morgan and Winship 2007). This approach is mostly attributed to the work of economist Donald Rubin (1973, 1974, 1976, 2005), but others made important contributions to its development, too (e.g. Heckman et al. 1996, 1998, 1997; Rosenbaum 1999; Rosenbaum and Rubin 1985).

Recall the example from above: We want to assess an immigrant's receiving country language skills dependent on her or his transnational activities. In this case, our dependent variable, language skill level is denoted by Y_{it} for person i at time t. The individual causal effect is then defined as

$$\delta_i = Y_{it}^1 - Y_{it}^0 \tag{5.1}$$

where Y_{it}^1 is person i's language skill level being transnationally active and Y_{it}^0 the skill level without being transnationally active and where δ_i is the individual causal effect. Clearly, we can never observe both outcomes for a given time-point t. Either a person is transnationally active or not. This has been called the fundamental problem of causal inference (Holland 1986, p. 947). Instead, what we can observe is

$$Y_{it} = Y_{it}^1 \quad \text{If} \quad D_{it} = 1$$
$$Y_{it} = Y_{it}^0 \quad \text{If} \quad D_{it} = 0$$

where D_{it} is a variable indicating whether the attribute of interest (e.g. transnational activity) is present ($D_{it} = 1$) or not ($D_{it} = 0$). If we put it into experimental wording, D_{it} is the exposure (or treatment) indicator. It is thus equal to 1 if an individual is exposed to the cause and 0 if she or he is not exposed (Morgan and Harding 2006, p. 8). The fundamental problem of causal inference is displayed in Table 5.1

We can only observe one of the potential outcomes. By formally stating this relationship, we can set up an observation rule, written as (Morgan and Winship 2007, p. 35)

$$Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0 \tag{5.2}$$

Table 5.1 The fundamental problem of causal inference. (Source: Morgan and Winship 2007, p. 35)

		Outcome	
		Y_{it}^0	Y_{it}^1
Exposure	$D_{it} = 0$	Observable	Counterfactual
	$D_{it} = 1$	Counterfactual	Observable

If $D_{it} = 1$, i.e. a person is exposed, then we observe Y_{it}^1 and Y_{it}^0 is unobservable. If $D_{it} = 0$, then we observe Y_{it}^0 while Y_{it}^0 is unobservable. In light of this fundamental problem, the question arises if causal inferences are, nevertheless, possible. Fortunately, they are. To see how, it is best to start with situations in which causal inferences are almost certainly impossible. Unfortunately, these situations are rather common. Since one cannot estimate the true individual causal effect, causal effects are oftentimes estimated by investigating differences in the expected outcomes between persons. Thus, we shift the attention to aggregate causal effects in a defined population (Morgan and Harding 2006, p. 8).

$$\begin{aligned} E[\delta_i] &= E\left[Y_{it}^1 - Y_{it}^0\right] \\ E[\delta_i] &= E\left[Y_{it}^1\right] - E[Y_{it}^0] \end{aligned} \tag{5.3}$$

For the example given above, this is the expected effect of being transnationally active for all persons from a randomly drawn sample regardless of the treatment status. In the counterfactual tradition, (5.3) is referred to as the average treatment effect (ATE). The average treatment effect is a weighted average of two other treatment parameters, the average treatment effect on the treated (ATT) and the average treatment effect on the non-treated (ATC), with the last letters referring to 'treatment' and 'control,' respectively (Brand and Halaby 2006, p. 756; Winship and Morgan 1999, p. 665; Morgan and Winship 2007, p. 45). The average treatment effect for the treated (ATT) is given by

$$\begin{aligned} E(\delta_i^1 \big| D_{it} = 1) &= E\left[Y_{it}^1 - Y_{it}^0 \big| D_{it} = 1\right] \\ &= E\left[Y_{it}^1 \big| D_{it} = 1\right] - E\left[Y_{it}^0 \big| D_{it} = 1\right] \end{aligned} \tag{5.4}$$

The superscript 1 in δ_i^1 indicates that this is the average treatment effect for the treated. The average treatment effect for the non-treated (ATC) by

$$\begin{aligned} E\left(\delta_i^0 \big| D_{it} = 0\right) &= E\left[Y_{it}^1 - Y_{it}^0 \big| D_{it} = 0\right] \\ &= E\left[Y_{it}^1 \big| D_{it} = 0\right] - E\left[Y_{it}^0 \big| D_{it} = 0\right] \end{aligned} \tag{5.5}$$

Here, the superscript 0 in δ_i^0 indicates that this average treatment effect refers to the non-treated. The decomposition of the average treatment effect takes the following form, where π is the proportion of the population which receives the treatment and $1 - \pi$ the corresponding proportion which does not

$$\begin{aligned} E[\delta_i] &= \pi E\left[\delta_i^1\right] - (1 - \pi) E\left[\delta_i^0\right] \\ E[\delta_i] &= \pi E\left[Y_{it}^1 - Y_{it}^0 \big| D_{it} = 1\right] - (1 - \pi) E\left[Y_{it}^1 - Y_{it}^0 \big| D_{it} = 0\right] \end{aligned} \tag{5.6}$$

Plugging (5.4) and (5.5) into (5.6) and rearranging the terms brings about (Morgan and Winship 2007, p. 45)

$$E[\delta_i] = \Big(\pi E\big[Y_{it}^1 \big| D_{it} = 1\big] + (1-\pi)E\big[Y_{it}^1 \big| D_{it} = 0\big]\Big) - \Big(\pi E\big[Y_{it}^0 \big| D_{it} = 1\big] + (1-\pi)E\big[Y_{it}^0 \big| D_{it} = 0\big]\Big)$$

Many approaches to causality do not differentiate between these three possible effects, the average treatment effect, the average treatment effect on the treated, and the average treatment effect on the non-treated. And, at first sight, this differentiation might appear somewhat confusing. But at second sight, it does make a lot of sense and is of practical importance. Regarding the example from above, it is quite fruitful to carefully consider what effect we are actually interested in. Are we interested in the effect of transnational activities on language skills for those who typically are transnationally active (ATT)? Are we interested in the effect of transnational activities for those who are typically not transnationally active (ATC)? Or are we interested in the average effect with regard to both groups (ATE)? In many instances the ATT (5.4) is the effect of interest. The biggest challenge we face in estimating (5.4) is evidently the fundamental problem of causal inference. It implies that Eq. (5.4) in its present form can only be estimated for $i \neq i$. To see why this is a problem, let us consider the example of a cross-sectional observational study. Imagine we collected data at $t = 0$ from a random sample of immigrants living in Germany and we want to investigate if transnational activities have a causal effect on a migrant's language skills. What we want to estimate is (5.4). What we are instead able to estimate is

$$E[\delta_i] = \pi E\big[Y_{it}^1 \big| D_{it} = 1\big] - (1-\pi)E\big[Y_{it}^0 \big| D_{it} = 0\big] \tag{5.7}$$

This between person comparison is called the naive estimator (Morgan and Winship 2007, p. 44). Only under very specific conditions does this estimator provide us with an unbiased estimation of the causal effect: if the treatment assignment mechanism is ignorable. What does this practically mean? We can easily imagine situations in which treatment assignment is not ignorable. In our example, the question would be whether immigrants who are transnationally active differ systematically from immigrants who are not. This is likely to be the case. First, there could be unobserved characteristics, like a general orientation toward the receiving or the sending country which concurrently determines whether or not an immigrant is transnationally active and influence a person's language skills. Second, it could be that receiving country language skills determine whether or not a person is transnationally active, i.e. treatment selection is not independent from the outcome. In both cases we face an endogeneity problem. The first is caused by unobserved heterogeneity and the second by simultaneity.

With respect to (5.3), this means that we compare persons that are not comparable, since they differ in relevant unobserved characteristics. As a result, the estimate of δ_i also includes baseline differences in Y_{it}. What we observe as language skill differences attributed to transnational activities might in reality only reflect baseline

differences.[1] Treatment selection is ignorable if the selection process is independent from the value of Y_{it}. Which is not given in the case of endogeneity; either Y_{it} is causing the value of D_{it} (simultaneity) or there is unaccounted covariation between Y_{it} and D_{it} (unobserved heterogeneity). The independence or ignorability assumption can formally be stated as

$$\left(Y_{it}^1, Y_{it}^0\right) \perp D_{it} \tag{5.8}$$

It means that treatment assignment is independent of the potential outcome, as is the case if D_{it} is completely random (Morgan and Winship 2007, p. 40). The underlying logic here is that if treatment assignment is independent from the potential outcome, then differences between Y_{it}^1 and Y_{it}^0 can be attributed to the treatment, as all other factors that might influence the potential outcome are balanced over the two groups (Brady and Collier 2004, p. 32). If exposure is randomly assigned, we can be sure that all relevant factors are balanced across the treatment and the control group. Yet, in observational data it is extremely unlikely that all confounding factors are by default balanced across the treatment and the control group. Nevertheless, if we can identify and observe possible confounders, we can solve this problem by conditioning on them. Thus we can achieve a weaker version of (5.8)

$$\left(Y_{it}^1, Y_{it}^0\right) \perp D_{it} \middle| S_{it} \tag{5.9}$$

where S_{it} is a set of observed variables (Morgan and Winship 2007, p. 75). By conditioning on S_{it}, we can meet the so-called Conditional Independence Assumption (CIA). It states that after controlling for S_{it}, the treatment assignment is ignorable. Although practically this is almost impossible as it is unlikely to identify and measure all relevant factors—some might be impossible or difficult to measure at all. When it comes to simultaneity, we have no reasonable means for tackling this problem with cross-sectional data, simply because we cannot establish causal ordering with this data. Even with panel data this issue remains problematic as we will see later. Coming back to the question regarding the validity of the naive estimator, we see that if the CIA is met, the naive estimator produces an unbiased estimate.

Of course, all these problems could be solved by an experiment. By relying on random selection into treatment and control group, unobserved characteristics will be balanced across groups and obviously do not influence treatment assignment. Moreover, in an experiment we can manipulate treatment and thus rule out simultaneity. The application of experiments is, unfortunately, seldom feasible for research questions in the social sciences, either due to practical or ethical issues. We just cannot randomly assign immigrants to a treatment group and a control group, either forcing them to be transnationally active or to abstain from these activities.

[1] Sometimes this is called unit homogeneity assumption (Brady and Collier 2004, p. 29; Rubin 1974), which states that the units used for comparison should be identical to each other in all relevant characteristics except for the treatment.

5.2 From the Counterfactual Approach to Regression Analysis

Still, causal inferences from observational data are possible. Two modeling approaches are able to deal with problems of endogeneity: Matching and a variant of the familiar regression analysis. The logic behind matching is intuitively very appealing and closely connected to the counterfactual approach: Persons are matched, who are very similar and (ideally) differ only with respect to the treatment. If the matches are indeed similar, then the difference in the outcome can be attributed to the treatment. In our example, this would mean that we compare immigrants, who are similar with respect to cultural, structural, social, emotional integration, and other relevant aspects but differ in whether or not they are transnationally active. Thus, we try to ensure that the conditional independence assumption is met by structuring the data in a way that mimics the outcome of random assignment into treatment and control group. The difference in language skills can then be attributed to the presence or absence of transnational activities. These methods are quite powerful and they are currently en vogue. One can get the impression that these methods are the panacea to all problems of causality in non-experimental research. But this obviously is not the case. They, too, have limitations. Although these limitations are discussed in the theoretical literature, often enough they are plainly ignored in applications. Consider, for instance, the assumption of conditional independence, which states that given a set of controls S treatment assignment is ignorable. To meet this assumption, we would have to identify all variables that predict treatment assignment or, in other words, the value of the main independent variable D.[2] This appears as a very challenging endeavor. It is neither likely that we are able to identify all relevant variables nor is it likely that we can measure all. We are once again faced with the problem of unobserved heterogeneity, which will bias our estimates.

Thus, I will focus on regression models, which under specific conditions can outperform matching methods. It should be noted, however, that most of the criticism directed at matching methods also applies to regression models.[3] Any statistical model should always be treated with great care and we are well advised to carefully consider if the underlying assumptions are met. Although there seems to be a natural affinity between the counterfactual approach and matching methods, it is easy to show how the regression approach can adopt this perspective (see e.g. Heckman and Robb 1989). Consider the observation rule in Eq. (5.2): $Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$. By rearranging the right part of the equation, we get

$$Y_{it} = Y_{it}^0 + \left(Y_{it}^1 - Y_{it}^0\right) D_{it} \tag{5.10}$$

$$Y_{it} = Y_{it}^0 + \delta_i D_{it} \tag{5.11}$$

[2] However, one could argue that this is a mere prediction problem, i.e. we only have to be concerned with a very accurate prediction of the treatment regardless of whether the variables we use are correlated with unobserved predictors.

[3] And there are approaches which try to deal with "selection on unobservables" (e.g. Morgan and Winship 2007, p. 184).

For people used to regression analysis this should appear familiar (Morgan and Winship 2007, p. 78). In this formulation, Y_{it}^0 corresponds to the intercept and δ_i to the regression parameter of the main independent variable. When it comes to predicting the concrete value of Y_{it}, the model specified in Eq. (5.11) will be imperfect, as there are other factors that also influence its value. Thus, Eq. (5.11) becomes

$$Y_{it} = Y_{it}^0 + \delta_i D_{it} + \mu_{it} \tag{5.12}$$

where μ_{it} is the error term (Morgan and Winship 2007, p. 78). We see that the naive estimator of the average treatment effect (ATE) equals the regression parameter of D_{it} in a linear regression model. Causal interpretation of regression coefficients is based, however, on the same assumption, although it is notated differently. Only if the core assumptions of the regression model are met, i.e. $\mu_{it} \sim N(0, \sigma^2)$ and $Cov(D_{it}, \mu_{it}) = 0$, then the regression coefficient will be unbiased. If $Cov(D_{it}, \mu_{it}) \neq 0$ we face an endogeneity problem. The standard answer to this problem is to include additional control variables which hopefully will break the correlation between D_{it} and the error term (Morgan and Winship 2007, p. 79), which equals achieving the conditional independence assumption. Including a set of observed variables S_{it} into Eq. (5.12) gives us

$$Y_{it} = \alpha + \beta D_{it} + \eta S_{it} + \mu_{it} \tag{5.13}$$

with S_{it} being a matrix of control variables and η a vector of the associated regression parameters. By including these controls we can achieve $Cov(D_{it}, \mu_{it}|S_{it}) = 0$. The last statement can equivalently be expressed as $E(\mu_{it}|D_{it}, S_{it}) = 0$. It states that the expected value of the error is zero given a set of control variables S_{it}; or that μ_{it} and D_{it} are conditionally independent given S_{it} (Heckman and Robb 1989, p. 522). This corresponds to the CIA as expressed in Eq. (5.9) (see also Morgan and Winship 2007, p. 141).[4] But this—just as the CIA—is hard to achieve if some of the characteristics which create the correlation between D_{it} and the error term cannot be identified or measured (Brady and Collier 2004, p. 34).

The choice between regression and matching is not as fundamental as it may appear. On the contrary, the methods are related and the difference is not that large. Matching has distinct advantages over simple regression models, such as the possibility to include individual level heterogeneity of causal effects by distinguishing between the different types of causal effects. Regression models usually assume that δ is constant over all individuals. Moreover, it forces the researcher to more carefully consider issues of endogeneity, most importantly selection biases, by trying to mimic a situation in which the data is balanced as in a randomized experiment. If conditions are met, matching results can provide better estimates than regression models (Morgan and Winship 2007, p. 142 ff.). Overall, matching and the counterfactual approach in general provoke a more thorough consideration of causality and how

[4] There are numerous different ways of expressing this assumption, but they all mean the same, e.g. $E(\mu_{it}|D_{it} = 1) - E(\mu_{it}|D_{it} = 0) = E((\mu_{it}|D_{it} = 1) - (\mu_{it}|D_{it} = 0) = 0$.

causal inferences might be flawed. But all this depends on whether the assumptions are met. If this is not the case, matching results are as biased as the results from a poorly specified regression model.

Even more, applications of matching procedures often rely on regression models to stratify the data, e.g. to compute propensity scores predicting treatment assignment. Matching is a two stage process, which strongly depends on a good estimate of treatment assignment. If the model in the first stage does not perform well in predicting treatment assignment, then causal effect estimates will be biased, because incomparable individuals will be chosen for comparison. How likely is it to meet the assumptions of matching? Probably as likely as meeting the assumptions of regression models, i.e. not too likely. It is a strong assumption that a set of control variables can be identified that ensures conditional independence or zero correlation between the error and the main independent variable, respectively. In most cases this will be hard to achieve. But there is a class of regression models which is able to consistently estimate the ATT, even in the presence of time-constant unobserved heterogeneity. Nevertheless, we need longitudinal data to estimate these models.

5.3 Excurse on Potential Exposure

But before we advance to discussing how longitudinal data can help us in dealing with the problem of unobserved heterogeneity, it seems necessary to critically examine the notion of causality within the counterfactual approach. Although its notion of causality is logically appealing, it has been criticized for being too narrow and as hardly applicable to many sociological research questions (e.g. Goldthorpe 2001). The counterfactual approach shares basic assumptions with experimental research, and, consequently, in its strict formulation accepts as causes only those characteristics which—in principle—can be treatments in experiments. This does not mean that the counterfactual approach rules out causes that cannot be manipulated because of ethical or practical reasons. It is sufficient if a treatment can be manipulated hypothetically—the counterfactual model thus rests on an idea of 'manipulative causality'. The basic assumption of the counterfactual approach thus is that each unit is potentially exposable to any of the causes (Holland 1986, p. 946). Does this hold for being transnationally active? In principle, we can imagine an immigrant being 'exposed' to being transnationally active or not. However, Holland (1986, p. 855) might disagree with this assessment, arguing that the "voluntary nature of much human activity makes causal statements about these activities difficult in many situations." Instead, causal claims in a strict sense can only be made if the exposure resembles a treatment. This then would allow for causal claims. In any case, the counterfactual approach rules out as causes any stable attributes, such as ethnic origin, gender, social background, etc., because they are time-invariant and cannot be manipulated. From the counterfactual perspective, stable attributes cannot cause other attributes. As Kempthorne (1978, p. 15) bluntly puts it, "it is [. . .] nonsense to talk about one trait of an individual *causing* or determining another trait of the

individual." Some authors therefore suggest that the general focus of analysis in the social sciences should change by concentrating on analyzing the effects of (time-varying) characteristics that, in principle, can be manipulated (Allison 1994, p. 192). But if we follow the counterfactual approach completely, this would not be enough. It is not sufficient for causes to be (potentially) manipulable; they also have to resemble (exogenous) treatments. Where do we go from here? Ruling out intentional human actions as causes would make a major part of social sciences research impossible. But there are other accounts of causality in which intentional actions even make for an integral part. For instance, the interventionist conception of causality in philosophy. According to this view, actors can intervene in systems and bring about changes that would not have occurred without the intervention (Wright 1971, p. 60 ff.). Most social scientists would agree that purposeful human action can be a cause and many, as we have seen in the previous chapter, even assume that human agency is the only cause for social change (e.g. Boudon 1980; Elster 1989; Esser 1999; Hedström 2005). Interestingly but not surprisingly, the term counterfactual is also used when the interventionist conception is discussed (see e.g. Tuomela 1976, p. 185). The logic of covering law model (or practical syllogism) applied to human agency as discussed in Chap. 4, in which intentional actions are linked to specific outcomes via individual motives, entails a counterfactual logic: If an action *A* had occurred, when in fact it did not, it would have produced a certain outcome *O*.

Moreover, from a sociological perspective, many stable characteristics, such as gender or ethnic origin, are not as stable as they first appear (Goldthorpe 2001, p. 7). It is true that a person cannot change her or his ethnic background. But a person's ethnic background has an effect only because it is constructed as socially relevant. Thus, it can be argued that immigrants do worse in the labor market not because of their ethnic background (as an essentialist attribute), but because of the social construction of ethnicity as a relevant criterion for differentiation. For instance, an employer might be less inclined to hire immigrants, because she or he believes that immigrants are less productive. In this case, the employer might use the (observable) ethnic background of a potential employee to infer unobservable characteristics—as assumed in the theory of statistical discrimination (Arrow 1971). Therefore, it is not the potential employee's ethnic background per se, but the employer's perception and evaluation of it. From this perspective, the effect of ethnic background is generated through a (potentially) manipulable mechanism. And in this sense, ethnic background becomes—hypothetically—manipulable. This conception is, by the way, highly compatible with Alba and Nee's (Alba and Nee 2003, p. 59; Alba 2008, p. 39) notion of assimilation as a decline in social salience and consequences of group membership—as discussed in Chap. 2. If we apply this perspective, making causal claims regarding such stable attributes still is difficult, because it is not the attribute itself but associated processes that have an effect. In many cases, a stable characteristic tends to be correlated with unobserved characteristics and processes. But the problem of unobserved heterogeneity is general and it applies to well manipulable characteristics, too.

It appears, therefore, that the strict counterfactual logic of causality (as, for instance, proposed by Holland (1986)) is too narrow to be applicable without modifications. But there are numerous less strict formulations, which are more compatible with sociological inquiry (see for instance Gangl 2010; Morgan and Winship 2007). The social sciences can surely only profit if they give more thought to the way in which stable characteristics affect outcomes. As argued above, the apparent relation between a stable characteristic and an outcome is often produced by a process—which can, at least in principle, be manipulated. In many cases it is thus a question of theoretical specificity, which brings us back to the discussion of the covering law model and insufficient explanations in Chap. 4. In any case, the counterfactual approach is very useful because it makes us sensitive to biases that can obscure our analysis. It emphasizes that making causal claims through applying statistical models relies on important, and in part untestable, assumptions.

5.4 Fixed Effects and Random Effects Models

The above discussion centers around the problem that we need a within person comparison ($Y_{it}^1 - Y_{it}^0$) but can only achieve a between person comparison ($Y_{it}^1 - Y_{jt}^0$), as we cannot observe both outcomes for one person. The strategies discussed—e.g. achieving conditional independence—aim at making the between person and the within person similar by adjusting for relevant controls. If we have repeated observations, regression methods for panel data analysis can offer an invaluable advantage, since a certain class of methods can adjust for time-invariant unobserved characteristics. These models are most often referred to as *fixed effects models*, but sometimes also as *change score models*. To see how these models do so, consider the regression equation from above

$$Y_{it} = \alpha + \beta D_{it} + \eta S_{it} + \mu_{it} \tag{5.14}$$

The error term μ_{it} captures all unobserved characteristics that we have not included in our model, as discussed above. We have not yet taken into account the possibility to decompose the error into a time-variant and a time-invariant part, which we can sensibly do if we have repeated observations. The error term can accordingly be decomposed into

$$\mu_{it} = \gamma_t + u_i + \varepsilon_{it} \tag{5.15}$$

γ_t is a period effect that is constant over all units, u_i is a time-invariant unit-specific effect that captures all unobserved heterogeneity (conditional on D_{it}), ε_{it} is a transitory idiosyncratic disturbance unique to the *ith* unit at time t (conditional on D_{it} and u_i) (Halaby 2004, p. 510). The key idea is that that u_i represents causes that are unobserved but stable over time. For now, we will ignore γ_t, since there are straightforward ways in which one can counter this problem (basically by estimating $\hat{\gamma}_t$) as I

discuss below. This decomposition can also be done with respect to the independent variables in S_{it}

$$S_{it} = X_{it} + Z_i \tag{5.16}$$

where Z_i refers equivalently to all time-invariant observed characteristics and X_{it} to the time-variant ones. Equation (5.15) can now be rewritten as

$$Y_{it} = \alpha + \beta D_{it} + \phi X_{it} + \eta Z_i + u_i + \gamma_t + \varepsilon_{it} \tag{5.17}$$

This equation might look a lot more complicated compared to Eq. (5.14), but it actually is not. All that has been done is to decompose the observed and unobserved characteristics into time-variant and time-invariant parts. Let us again consider the example. The primary interest still is to estimate the effect of transnational activities on host country language skills—thus β is the primary parameter of interest. Z_i captures all measured characteristics that might influence an immigrant's language skills which do not change over time. Examples might be ethnic origin, gender, age at migration, and the like. Similarly, X_{it} refers to measured characteristics that can influence an immigrant's language skills and which do change over time. Examples of these characteristics can encompass things like labor force participation, income, years of residence, the ethnic composition of personal networks, and the like. Regarding the unobserved parts in Eq. (5.18), u_i covers all time-invariant characteristics that are not measured and not included in the model. This might be unmeasured or immeasurable (language) ability or a general and constant orientation toward the receiving and the sending country, respectively. In this sense, u_i can be interpreted as a unit-specific term that adds to the overall intercept. Similarly, ε_{it} includes unmeasured characteristics that change over time. The causal parameter β provides us with an unbiased estimate if the unmeasured characteristics do not differ systematically between treatment and control group. That is we assume that $E(\varepsilon_{it}|D_{it} = 1) - E(\varepsilon_{it}|D_{it} = 0) = E((\varepsilon_{it}|D_{it} = 1) - (\varepsilon_{it}|D_{it} = 0)) = E(\varepsilon_{it}) = 0$ and $E(u_i|D_{it} = 1) - E(u_i|D_{it} = 0) = E((u_i|D_{it} = 1) - (u_i|D_{it} = 0)) = E(u_i) = 0$—in other words that the errors are uncorrelated with the independent variables: $Cov(D_{it}, \varepsilon_{it}) = 0$ and $Cov(D_{it}, u_i) = 0$ (disregarding X_{it} and Z_i in the following). Now, if we consider a case in which we have two consecutive observations for each person

$$Y_{it\,0} = \alpha + \beta D_{it\,0} + \phi X_{it\,0} + \eta Z_i + u_i + \gamma_{t0} + \varepsilon_{it\,0} \tag{5.18}$$

$$Y_{it\,1} = \alpha + \beta D_{it\,1} + \phi X_{it\,1} + \eta Z_i + u_i + \gamma_{t1} + \varepsilon_{it\,1} \tag{5.19}$$

Subtracting (5.18) from (5.19) brings

$$(Y_{it\,1} - Y_{it\,0}) = \beta(D_{it\,1} - D_{it\,0}) + \phi(X_{it\,1} - X_{it\,0}) + (\gamma_{t\,1} - \gamma_{t\,0}) + (\varepsilon_{it\,1} - \varepsilon_{it\,0}) \tag{5.20}$$

the *first-difference model*. Owed to the subtraction, the time-invariant unobserved heterogeneity cancels out. This also intuitively makes sense. We started from trying

to estimate the true causal effect through a within comparison, i.e. through $Y_{it}^1 - Y_{it}^0$. If we consider the ATT given by (5.4) as

$$E\left(\delta_i^1 \middle| D_{it} = 1\right) = E\left[Y_{it}^1 \middle| D_{it} = 1\right] - E\left[Y_{it}^0 \middle| D_{it} = 1\right]$$

it is obvious, as we have seen above, that this ideal form of this intra-personal comparison is impossible to achieve for $t = t$. Still, we need a within comparison, because only the within comparison can assure that the effect estimate is not biased by confounding factors. With longitudinal data, we can approximate the estimation of this effect by looking at the intra-individual change in the outcome variable after exposure. We could thus state the ATT from a longitudinal perspective as

$$\begin{aligned} E\left(\delta_i^1 \middle| D_{it1} = 1\right) &= E\left[Y_{it1}^1 - Y_{it0}^0 \middle| D_{it1} = 1\right] \\ E\left(\delta_i^1 \middle| D_{it1} = 1\right) &= E\left[Y_{it1}^1 \middle| D_{it1} = 1\right] - E\left[Y_{it0}^0 \middle| D_{it1} = 1\right] \end{aligned} \tag{5.21}$$

assuming that the treatment exposure occurred in $t_1 < t < t_0$. By looking at the difference between Y_{it1}^1 and Y_{it0}^0 it is clear that unit-specific time-invariant factors will not bias any estimate, because their influence on Y_{it} is *constant* over time. The estimate of β in (5.20) thus corresponds to the ATT, because all individuals which do not experience a change in D_{it} are disregarded in the estimation (this will be explained in more detail in the next section). But this is not really a problem, because most of the time we are not interested in the naive estimator—a plain between person comparison—but in some kind of intra-personal comparison. In this sense, we are more interested in intra-individual (co)variation than inter-individual (co)variation. Is the estimate of β unbiased? Let us again consider the three potential sources of bias as in Eq. (5.15). The observed (read: measured) value of Y_{it} for a given individual i at time t in the treatment group can be stated as

$$Y_{it}^1 = \delta_i + \gamma_t + u_i^1 + \varepsilon_{it}^1 \tag{5.22}$$

and equivalently for an individual i in the control group

$$Y_{it}^0 = \gamma_t + u_i^0 + \varepsilon_{it}^0 \tag{5.23}$$

What we eventually observe is accordingly the sum of these three, respectively four parameters. If we state Eq. (5.1) with respect to (5.22) and (5.23) for $i \neq i$ we get

$$Y_{it}^1 - Y_{it}^0 = \delta_i + \left(u_i^1 - u_i^0\right) + \left(\varepsilon_{it}^1 - \varepsilon_{it}^0\right) + (\gamma_t - \gamma_t) \tag{5.24}$$

As we can see, Eq. (5.24) will, in its unconditional, simple form produce an unbiased estimate only if $u_i^1 = u_i^0$, $\varepsilon_{it}^1 = \varepsilon_{it}^0$, and $\gamma_t^1 = \gamma_t^0$. This is merely a different way of stating the ignorability assumption. But the discussion hitherto has shown that this condition is hard to meet. While a cross-sectional between person comparison eliminates the influence of γ_t, the estimate is still likely to be biased by $(u_i^1 - u_i^0)$ and $(\varepsilon_{it}^1 - \varepsilon_{it}^0)$. As we have seen above, longitudinal data allow, however, to partial out $(u_i^1 - u_i^0)$. Since $t \neq t$, this amounts to

$$Y_{it1}^1 - Y_{it0}^0 = \delta_i + (\gamma_{t1} - \gamma_{t0}) + \left(\varepsilon_{it1}^1 - \varepsilon_{it0}^0\right) \tag{5.25}$$

From this perspective, one might argue that there is a natural affinity between fixed effects regression models and the counterfactual approach to causality. Bias may still come from $(\gamma_{t1} - \gamma_{t0})$ and $(\varepsilon^1_{it1} - \varepsilon^0_{it0})$. But with more than two observations, period effects can easily be controlled by directly including them in the list of covariates. When there are more than two observations, the fixed effects model can be estimated by time-demeaning, i.e. by subtracting the between-model

$$\overline{Y}_i - \alpha + \beta\overline{D}_i + \varphi\overline{X}_i + \eta Z_i + u_i + \overline{\varepsilon}_{it} \tag{5.26}$$

from (5.17) leading to

$$(Y_{it1} - \overline{Y}_i) = \beta(D_{it} - \overline{D}_i) + \varphi(X_{it} - \overline{X}_i) + \gamma_t + (\varepsilon_{it} - \overline{\varepsilon}_i) \tag{5.27}$$

The subtraction equally rids the equation of u_i, eliminating this potential source of bias. Including period dummies in (5.27) will moreover effectively control for any for common aggregate effects. Of course, ε_{it} remains as a potential source of bias. But there is no cure for this problem. We just have to assume that $Cov(D_{it}, \varepsilon_{it}) = 0$.

Instead of the fixed effects model, we could also estimate a so-called *random effects model* as specified in (5.17). More precisely, this is a random intercept model, because it adds a subject specific u_i to the overall intercept α. It assumes that u_i is a random variable that follows a normal distribution with mean zero and a constant variance $u_i \sim N(0, \sigma^2)$. In a way we thus model unobserved characteristics with the parameter u_i. In the example from above this means that we allow for the fact that persons have different baseline levels of language skills. Compared to the fixed effects model, the random effect model can be more efficient, because it uses variation between and within subjects to estimate β. The fixed effects model, in contrast, only uses within subject variation to estimate β. One practical implication of using only within variation is that all observations that do not change in D_{it} are discarded from the analysis. If we choose a non-linear version of the fixed effects model, mode, for instance, a conditional logistic regression model for binary outcomes (Allison 2009; Chamberlain 1980), the sample on which the estimation is based becomes even smaller, because not only are subjects excluded from the analysis that do not experience a change in D_{it}, but also all subjects are excluded that do not change in Y_{it}. In these situations, random effect models can be substantially more efficient (Neuhaus and Lesperance 1996, p. 445).

However, the gain in efficiency is based on the assumption of common between and within subject effects (Neuhaus and Kalbfleisch 1998, p. 644). Moreover, the assumption of zero correlation between observed variables and unobserved effect—$Cov(D_{it}, u_i) = 0$, $Cov(X_{it}, u_i) = 0$, and $Cov(Z_i, u_i) = 0$—is hard to justify as the above discussion shows. Often enough the random effect will not be random and we are dealing with latent (unobserved) classes, where unobserved and observed characteristics are correlated. And if this crucial assumption is violated, the estimates will again be biased. We certainly could try to include other covariates hoping to break the correlation. But at this point it should be clear that it is almost impossible

to find and measure all relevant confounders.[5] Consequently, the advantage of fixed effects regression models is invaluable, since we don't have to bother searching for time-invariant confounders any longer. Any fixed effects just cancel out. This great advantage, however, comes at a price. As the fixed unobserved characteristics cancel out from the model, so do the fixed observed characteristics. Therefore, we cannot estimate the effects of time-invariant factors. In our example, this means that we cannot include variables like ethnic origin and gender, because they do not change over time. As mentioned above, some authors argue that the social sciences should change their research focus, by concentrating solely on questions of *change causes change* (Allison 1994, p. 192). In many instances, however, the association between time-invariant characteristics and some property are of crucial interest, especially if we are interested in examining the possibility of systematic differences between groups.

Regarding this work's research questions, it would certainly be a huge drawback if we could not estimate the association between the different ethnic origins and transnational activities. Of course, any association between ethnic origin and transnational activities cannot be interpreted causally in a strict sense. But significant differences in ethnic origins can be understood as a measure of unobserved characteristics (processes) that are associated with the different ethnic origins. This is certainly interesting. One possible remedy to this drawback is the inclusion of interactions between time varying and time-invariant covariates in fixed effects models (Allison 2009, p. 37). If the model includes time, i.e. $t - 1$ period dummies, we can model interactions of time and ethnic origin. But these interactions do not estimate the effect of the time-invariant covariate. Instead, they estimate a possible change in its effect compared to a reference period. Therefore, this only models how the association of one ethnic background with transnational involvement changes over time. If our intention is to compare the 'effect' of different ethnic backgrounds, modeling interactions with time is, however, not what we are interested in.

5.4.1 Correlated Random Effects and Hybrid Models

Another possibility is to estimate a *hybrid model* (Allison 2005, 2009; Schunck 2013) or a *correlated random effect model* (Schunck 2013; Wooldridge 2005, 2010, p. 286, 332, 615 ff.), first proposed by Mundlak (1978). These models are modified random effect models which differentiate within- and between-cluster effects (in the present case observations are clustered within persons) (Mundlak 1978; Neuhaus and

[5] A different reasoning for using fixed effects models is provided by Neuhaus and Kalbfleisch (1998, p. 644): "When between- and within-cluster covariate effects are different, models that assume that these effects are the same do not provide estimates of any substantive interest; the misspecified models measure neither between- nor within-cluster covariate effects." If we rely on fixed effects estimates, we can at least be sure that they correctly model intra-subject variation. See below for more details on within- and between-variation.

Kalbfleisch 1998; Neuhaus and McCulloch 2006). Thus, the models and simultaneously estimates the effects of within and between subject variation in explanatory variables. This is done by decomposing the variation of time-variant variables into between- ($\overline{X}_i$) and within-subject ($\overline{X}_i - X_{it}$) variation (Schunck 2013).[6] The hybrid models is given by

$$Y_{it} = \alpha + \beta_W(X_{it} - \overline{X}_i) + \beta_B\overline{X}_i + \eta_m Z_i + u_i + \varepsilon_{it} \tag{5.28}$$

β_W is the within estimator, which, in the linear case, equals the fixed effects estimator and β_B estimates the between effect (Maddala 1987; Mundlak 1978; Neuhaus and Kalbfleisch 1998). An alternative formulation to (5.28) is the correlated random effects model (also known as the Mundlak model), which takes the form

$$Y_{it} = \alpha + \beta_W X_{it} + \varphi\overline{X}_i + \eta_m Z_i + u_i + \varepsilon_{it} \tag{5.29}$$

Here, the mean ($\overline{X}_i$) picks up any correlation between the time-varying covariates and u_i, relaxing the assumption that $Cov(X_{it}, u_i) = 0$. β_W from (5.29) also estimates the within effect, i.e. it is the fixed effects estimator (Mundlak 1978). Including the mean of a time-varying covariate in a random-effects model is therefore an alternative to cluster mean centering (Halaby 2003, p. 519). The difference between (5.28) and (5.29) lies in the between-effect. Rewriting (5.28) as

$$Y_{it} = \alpha + \beta_W X_{it} + (\beta_B - \beta_W)\overline{X}_i + \eta_m Z_i + u_i + \varepsilon_{it} \tag{5.30}$$

makes this obvious (Schunck 2013, p. 67). The hybrid model estimates the between effect (β_B) whereas the correlated random effect model estimates the *difference* of between and within effect ($\varphi = \beta_B - \beta_W$) (Mundlak 1978). Both models have similar properties, but since the interpretation of β_B—the between subject effect—is more straighforeward than ϕ, I will use the hybrid model in the following.

We can further make use of a good feature of the hybrid model, namely testing for equality of within and between estimates, namely if $\beta_W = \beta_B$ (Allison 2009, p. 25). This test, which can be done easily though a Wald test, is also known as an augmented regression test (Jones et al. 2007, p. 217). It can be used as an alternative to the Hausman specification test (Baltagi 2008, p. 73; Hausman 1978). If between and within effects are the same, i.e. $\beta_W = \beta_B = \beta$, (5.29) and (5.30) will collapse into the standard random effects model.

In general, and if between and within estimates are not equal, between estimates can be seen as an approximation of the strength of the average association of dependent and independent variable between groups, that is i.e. including selection effects and unobserved heterogeneity. In other words, a comparison of between- and within-estimates can give us some idea of the strength of selection effects.

It is noteworthy that differentiating between- from within-estimates is not new and also applied in the multilevel literature, where this practice is referred to as

[6] To simplify notation I abstain from including the period effect γ_t in the notation. Moreover, I do not distinguish any longer between D_{it} and X_{it}, because both denote time varying covariates.

group-mean-centering (Raudenbush and Bryk 2002, p. 135 ff.; Kreft and Leeuw 1998, p. 105 ff.; Rabe-Hesketh and Skrondal 2005, p. 42; Snijders and Bosker 2004, pp. 52–56; Kaufman 1993; Kreft et al. 1995; Raudenbush 1989). For instance, Rabe-Hesketh and Skrondal (2005, p. 43) propose a similar method for comparing between and within effects of time-varying variables. Draper (2008, p. 108) suggests group-mean-centering to reduce positive autocorrelation of sampled draws.

5.4.2 Nonlinear Models

Random, fixed effects, and hybrid models can also be applied if the dependent variable is limited, i.e. it is categorical or binary. A practical example of a binary variable is whether or not an immigrant has visited her or his country of origin during the last year, a prominent example of a transnational activity. In the binary case, the dependent variable y is constrained to take either the value 0 or 1. We cannot apply a linear probability model, since it would violate the assumption of homogeneity in variance (homoscedasticity) (Giesselmann and Windzio 2012, p. 130). Moreover, while probabilities are limited to take values between 0 and 1, the linear probability model if applied to the binary case does not have lower or upper bounds for its prediction. Thus, if we apply the linear probability model, we can get probabilities below 0 and above 1—which do not make sense. A model which avoids these problems is the logistic regression model (Long 1997). In the following, I will thus discuss the logistic regression with random and fixed effects. The standard (pooled) *binary logistic regression model* is given as (not differentiating between time-variant and invariant covariates for simplicity)

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \beta X_{it} \tag{5.31}$$

with p_{it} being the probability that the dependent variable takes the value 1, given the values of independent variables ($P(Y_{it} = 1 \big| X_{it})$). The above equation is the log odds formulation of the logistic regression model. Alternatively, it can be displayed directly regarding probabilities as

$$P(Y_{it} = 1 \big| X_{it}) = \frac{\exp(\beta X_{it})}{1+\exp(\beta X_{it})} = \frac{1}{1+\exp(-\beta X_{it})} \tag{5.32}$$

The logistic regression with random effects or fixed effects, respectively, consequentially is

$$P(Y_{it} = 1 \big| X_{it}, u_i) = \frac{1}{1+\exp(-(u_i + \beta X_{it}))} \tag{5.33}$$

As in the linear case, the random effect logistic regression model assumes that u_i is a random variable following a normal distribution with a mean of zero and a constant variance ($u_i \sim N(0, \sigma^2)$). For the estimates to be consistent, the model likewise

requires the random effect to be uncorrelated with the (time-invariant) explanatory variables ($Cov(u_i, X_{it}) = 0$).[7] As we have seen above, the assumption that all regressors are strictly exogenous so that $Cov(u_i, X_{it}) = 0$ is rather unrealistic. It is quite likely that there are some characteristics influencing both the dependent and the independent variable on which we do not have any information. As in the linear case, the fixed effects model is, in this regard, more attractive, since it does neither require any assumption on the distribution of u_i nor its correlation with other regressors. However, the computation of the fixed effect is complicated by the so-called 'incidental parameter problem' (Lancaster 2000; Neyman and Scott 1948). Unfortunately, in contrast to the linear case, the individual fixed effects (u_i) cannot be eliminated by a simple linear transformation, such as the first-differences of time-demeaning (Hsiao 2003, p. 194). In general, maximum likelihood (ML) estimates are desirable because of their large sample properties. It has been shown that with an increasing sample size, ML estimates are asymptotically consistent, efficient, and normally distributed (Long 1997, p. 33). But the ML estimates' consistency depends on the assumption that the number of parameters remains constant as the sample size increases (Allison 2005, p. 57). This is not the case with Eq. (5.33)—the number of u_i increases with the sample size. Without going into detail ((for detailed discussions see Giesselmann and Windzio 2012, pp. 143–150; Wooldridge 2010, pp. 621–622)), the solution to the problem lies in applying conditional maximum likelihood as proposed by Chamberlain (1980). This approach *conditions* the likelihood function on the total number of events observed for each case. In the two period case, the conditional likelihood model takes the form

$$\log\left(\frac{P(Y_{i1}=0, Y_{i2}=1|Y_{i1}+Y_{i2}=1)}{1-P(Y_{i1}=0, Y_{i2}=1|Y_{i1}+Y_{i2}=1)}\right) = \beta(X_{i2}-X_{i1}) \tag{5.34}$$

The fixed effects approach in the binary case suffers from the same drawback as in the linear case: We cannot estimate the effect of time-invariant covariates. To remedy this shortcoming we can again rely on the hybrid approach, which combines the random and the fixed effects model and allows estimating the effect of time-invariant covariates (Allison 2009, p. 39; Neuhaus and Kalbfleisch 1998, p. 640; Wooldridge 2010)

$$P(y_{it}=1\big|x_{it}, u_i) = \frac{1}{1+\exp(-(u_i+\beta_W(X_{it}-\overline{X}_i)+\beta_B\overline{X}_i)} \tag{5.35}$$

But equality of within and fixed effects estimator is given in the linear case. In nonlinear cases, for instance for logit models, the differentiation into within and between effects in a random effect model approximates the fixed effects estimator (Allison 2009; Neuhaus and Kalbfleisch 1998; Neuhaus and McCulloch 2006; Wooldridge

[7] The parameters in standard logistic regressions are estimated via maximum likelihood (ML), an iterative method that produces estimates that make the observed data most likely. However, the parameters in the random effect logistic regression (Eq. 5.33) cannot be estimated by ML. Instead, quasi-likelihood approaches are applied (Hox 2002, p. 107 ff.) or, as some statistics programs do, numerical integration, such as the Gauss-Hermite quadrature (StataCorp LP 2007, p. 215 ff.).

2010, pp. 615–616, p. 620, 766). It has been shown that the estimates of coefficients and the standard errors of the deviation variables in the hybrid model come extremely close to those of the fixed effects model (Allison 2005, p. 67; 2009, p. 41; Neuhaus and Kalbfleisch 1998, p. 640 ff.). One reason for this may be trace back to fixed variance of the error in logit models, which is fixed at $(\pi^2/3) \approx 3.29$. This has peculiar implications for the estimation of β: if by inclusion of another predictor (e.g. z_i) the explanatory power of the model becomes higher (or the explained variance increases if we approach logistic regression with the idea of a latent, metric propensity Y_i^*, that predicts whether $Y_i = 0$ or $Y_i = 1$ (Long 1997, p. 41, 105)), the total variance of the dependent variable necessarily increases (see Mood 2010 for details). If the total variance of the dependent variable increases, its scale also increases and, as a consequence, the estimate of β has to change. It now expresses the relation between X_{it} and Y_{it} in another metric. Hence, a model in which $(X_{it} - \overline{X}_i)$ is the sole predictor will provide us with a different estimate of β than a model in which we include both $(X_{it} - \overline{X}_i)$ and $(\overline{X}_i)$ even though the $Cov(X_{it}, (X_{it} - \overline{X}_i)) = 0$ by construction. This is an important difference to linear regression models in which the error variance is unrestricted and estimates of β only change if the predictors are correlated. This also complicates causal interpretation of effect estimates in logistic regression models. As is well known due to the nonlinearity of logistic regression models, the effect of a covariate depends on the values of other covariates. This unfortunately also holds for excluded covariates. Thus, effect estimates from logistic regression models can be interpreted causally with respect to their effect size only in a fully, i.e. perfectly, specified model (Kühnel and Krebs 2010, p. 885). If we were able to specify such a model, it would, of course, be unnecessary to deal with complex panel data models, as a cross-sectional model that perfectly predicts treatment would be sufficient. Of course, the whole discussion so far underscores that it is impossible to specify such a model.

Besides linear and binary logistic regression models, the statistical analyses presented in Chaps. 7 and 8 also encompass *multinomial logistic regression models* and *count models*. Both classes of models (Long 1997; Long and Freese 2006) are also specified as hybrid models, thus decomposing the effect of time-varying covariates into within and between effects (Allison 2009, pp. 44–45; Wooldridge 2010, p. 766). A fixed effects version of the multinomial logistic regression model is

$$\log\left(\frac{F_{ijt}}{1 - F_{ijt}}\right) = \mu_{tj} + \beta_j X_{it} + u_{ij}, \; j = 1, \ldots, J-1 \tag{5.36}$$

where $F_{ijt} = \sum_{m=j}^{J} p_{imt}$ is the 'cumulative' probability of being in category j or higher (Allison 2005). u_{ij} denotes the time-constant but potentially response-specific individual fixed effect, that is the time-constant, response-specific unobserved heterogeneity. μ_{tj} denotes the time and category-specific intercept. In principle, this model can be estimated equivalently to the binary case, by conditional maximum likelihood with conditioning on the frequency counts of the different responses (Allison 2009, p. 44). However, this is not yet implemented in any commercial software package. To estimate a multinomial logistic regression model with individual fixed effects, we have several possibilities at hand. We could compute a series

$(J-1)$ of binary comparisons (Begg and Gray 1984). But this would strip us of the possibility of overall tests for predictors and a comparison of the relative risks of one outcome vs. a specific other in a joint model. Therefore, Allison (2009, p. 45) suggests to run a standard multinomial logistic regression model including the mean as well as time-demeaning formulations of the time varying covariates and adjusting for dependence of observations with cluster robust standard errors. This approach will be followed in this study.[8]

Besides categorical variables with several response categories, the analysis will also deal with so-called count variables that indicate how often something happened. In principle, count variables are metric. They have a natural zero-point and the response categories are ordered and equally spaced. However, count processes can produce strongly skewed distributions. For instance, count processes often produce distributions with very long right tales. This may cause overdispersion, meaning that the variance is greater than the mean. Technically, this is presented as $Var(y|x) > E(y|x)$. However, the "small-mean property" (Cameron and Trivedi 2009, p. 553) can also be caused by the presence of many zeros, i.e. many zero counts. In some cases there can also be an "excess" of zeros. One key indicator of transnational involvement, sending remittances, is such a count variable. And this variable is characterized by both overdispersion and excess zeros. In our data only about 10 % of the immigrants actually send remittances (see Chap. 6, Table 6.5). Therefore, 90 % of the observations have a zero count. Moreover, the variance is much greater than the mean. If we use linear regression to investigate the effect of predictors on such a count variable, this may lead to inefficient and biased estimates (Long 1997, p. 217). But there is a specific class of non-linear regression models—count models—which are able to handle such outcomes. In our case, the *zero-inflated negative binomial regression model* appears best suited for analyses. Zero-inflated count models assume that there are two latent (i.e. unobserved) groups: an always zero group and a not always zero group (Long and Freese 2006, p. 394). Therefore, zero-inflated models assume that the counts are generated by a two stage processes. First, a distinction is made between two latent groups, in our case non-remitting immigrants ($A = 1$) and remitting immigrants ($A = 0$). In the first group the outcome is always zero. For the second group a count process produces the actual count, i.e. the amount remitted (Greene 2000). In the first step, the probability of being member of latent group A is estimated by binary logistic regression:

$$P(A_{it} = 1 \big| X_{it}) = \pi_{it} = \frac{\exp(\beta S_{it})}{1 + \exp(\beta S_{it})} = \frac{1}{1 + \exp(-\beta S_{it})} \tag{5.37}$$

where π_{it} is the predicted probability of falling into the latent group of 'non-remitters' and S_{it} a vector of explanatory variables. The probability of remitting k Euros is then computed by mixing the probability of not-remitting and the probability of having a

[8] It should be noted that the estimates from such a model are population averaged and will be generally a bit smaller subject-specific random effects estimates (Allison 2009, p. 47).

count k, conditional upon a specified vector of explanatory variables X_{it}:

$$\Pr(y_{it} = k \mid X_{it}, S_{it}) = (\pi_{it} \times 0) + \Big((1 - \pi_{it}) \times \Pr(y_{ti} = k \mid X_{it}, A_{it} = 0)\Big) \quad (5.38)$$

$$= (1 - \pi_{it}) \times \Pr(y_{it} = k \mid X_{it}, A_{it} = 0) \quad (5.39)$$

This is the probability of remitting k Euros conditional of remitting, weighted by the probability of remitting $(1–\pi_i)$ with regard to the overall population. Unfortunately, just as for the multinomial logistic regression model, a panel model of the zero-inflated negative binomial regression model has not been implemented in commercial software yet. An alternative to the zero-inflated negative binomial regression would be a simple negative binomial regression—which can also be estimated as a panel model (with random or fixed effects). However, there are two reasons that speak against this model. First, the count process is not adequately described by a simply negative binomial distribution. Second, the fixed-effects estimator of the negative binomial regression model does not—contrary to other fixed-effects models—control for all time-constant unobserved heterogeneity (Allison and Waterman 2002, pp. 263–264). Because of this, I compute a zero-inflated negative binomial regression model and include the mean as well as time-demeaning formulations of the time varying covariates and adjust for dependence of observations with cluster robust standard errors.

5.5 Endogeneity Revisited

As we have seen, we can combine the virtues of the within estimator with the inclusion of time invariant explanatory variables in one model. But unobserved heterogeneity is not the only problem we encounter in data analysis. Measurement error is another major issue in statistical modeling, as mentioned at the beginning of this chapter. But how does measurement error influence estimation of fixed effects models? In simple linear regression models, measurement error in independent variables will lead to the so-called 'attenuation bias.' If the reliability of the measure is below 1, then the estimates will be downward biased by the factor $\sigma_*^2/(\sigma_*^2 + \sigma_\eta^2)$, where σ_*^2 is the variance of the true, unobserved variable and σ_η^2 the variance of the disturbance (Skrondal and Rabe-Hesketh 2004, p. 76). The greater the measurement error, the stronger will be the attenuation bias, since the fraction becomes smaller. The problem becomes more complicated in multivariate and non-linear models (see e.g. Bollen 1989, p. 159 ff.), although results from simulation studies suggest that we are also likely to face a downward bias (Schunck 2009). In the context of fixed effects models, it is generally assumed that reliability problems are magnified (Engel and Reinecke 1994, p. 19; Burr and Nesselroade 1990, pp. 9–10; Griliches and Hausman 1986) although some analyses suggest the opposite (see e.g. Bound and Krueger 1991). A related problem that fixed effects models face is dealing with the phenomenon of 'regression towards the mean.' In essence, this refers to the phenomenon that a variable that takes on an extreme value at its first measurement will tend to be

closer to the mean of the overall distribution at a later measurement. In face of this, fixed effects estimation may also produce biased results (Finkel 1995, p. 8). As a solution to both problems, the literature suggests dynamic models, sometimes also called level score models (Finkel 1995, p. 7 ff.; Burr and Nesselroade 1990, p. 9 ff.). These models include a lagged version of the dependent variable. A simple dynamic regression model takes the form

$$Y_{it} = \alpha + \beta X_{it} + \rho Y_{it-1} + u_i + \varepsilon_{it} \tag{5.40}$$

The inclusion of lagged versions of dependent variables allows us to control for the phenomenon of regression toward the mean, since we can estimate the impact of independent variable given the value of dependent variable at a previous time point. Moreover, the bias due to measurement error in β obtained from (5.40) will be lower as compared to a fixed effects estimate (Rodgers 1989, p. 443; Kohler 2002, p. 238). Furthermore, it is sometimes argued that there are substantial reasons to assume that the value of the dependent variable at time $t - 1$ has a causal effect on its value at time t (Finkel 1995, pp. 7–11). This is referred to as 'state dependency' as the state (i.e. value) of the dependent variable at previous time-points determines its present value. Taking up the example from above, this implies that an immigrant's language proficiency at time t is causally influenced by her or his proficiency at $t - 1$. Empirically, we are likely to observe state dependency rather often. But all that we really observe is that individuals who have high (low) levels of Y_{it-1} also have high (low) levels of Y_{it}. The problem is that there rarely is a sound theoretical justification for including the lagged dependent variable in the model (Liker et al. 1985, p. 86). We should keep in mind that there can be various reasons for a correlation between y_{it-1} and y_{it}. True state dependency is given when there is a causal mechanism so that y_{it-1} determines y_{it}. But a correlation between Y_{it-1} and Y_{it} can also be due to (unobserved) confounders. In this case, Y_{it} does not have a causal effect on Y_{it}. Instead, additional omitted explanatory variables determine both y_{it-1} and y_{it}, producing a spurious correlation. In many cases, an observed correlation between Y_{it-1} and Y_{it} is therefore not an indicator for a causal relationship, but for a misspecified model. If the latter can be ruled out and we have good (theoretical) reasons to assume that there is a causal effect of Y_{it-1} on Y_{it}, then we indeed face a problem. Morgan and Winship (2007, p. 254 ff.) provide an interesting example for an indirect (causal) relation between Y_{it-1} and Y_{it} In their example, Y_{it-1} does not directly determine the value Y_{it}, but it determines treatment selection (amongst other factors). This means, for instance, that language proficiency at $t - 1$ influences transnational activities at time t. This would be a case in point for using dynamic models. However, even in this situation fixed effects estimates appear to provide more accurate estimates of the true causal effect (Morgan and Winship 2007, p. 256).[9] Still, why not use dynamic models if

[9] However, the choice between a dynamic or a fixed effects model also depends on the (hypothetical) evolvement of Y_{it}^0 and Y_{it}^1 in the absence of treatment (for a thorough discussion, see Morgan and Winship (2007, p. 258 ff., 264)). The fixed effects model assumes that the values of Y_{it} evolve parallel between treatment and control group while the dynamic model assumes that the values of Y_{it} converge. Both assumptions might or might not hold. Neither model is, however, well suited when it comes to estimating causal effects in the presence of diverging, increasing differences between

they enable us to control for state dependency and even offer a partial remedy of measurement error and regression to the mean? The problem is that dynamic models will only provide unbiased estimates if (a) $Cov(X_{it}, \varepsilon_{it}) = 0$ and $Cov(X_{it}, u_i) = 0$ and (b) additionally $Cov(Y_{it-1}, \varepsilon_{it}) = 0$ and $Cov(Y_{it-1}, u_i) = 0$. Not only does the discussion so far suggest that the first three criteria are unlikely met, but the fourth criterion $Cov(Y_{it-1}, u_i) = 0$ is violated by definition (Wooldridge 2002, p. 256; Stewart 2007, pp. 515–516). If Y_{it} is a function of u_i, there is necessarily a correlation between Y_{it-1} and u_i and $Cov(Y_{it-1}, u_i) = 0$ will be violated. This intuitively makes sense, as time-constant unobserved factors by definition exert the same influence on the level of the dependent variable at any occasion. The lagged dependent variable is thus necessarily correlated with the time-constant error. Consequently, both estimates of β and ρ will be biased. Yet, we cannot rule out measurement error and this can bias our estimates, too. Indeed, if measurement error is present, dynamic models provide more accurate estimates than fixed effects models. Thus, we have to decide which problem we think is more pressing. Simulation studies indicate that the bias from unobserved heterogeneity seems to outweigh the bias from measurement error (Rodgers 1989, p. 444 ff.). Moreover, bias from measurement error usually does not change the direction of the estimates, whereas the direction of bias from unobserved heterogeneity is unpredictable (Palta and Seplaki 2002, p. 188) and the advantages of within estimators become more pronounced with an increasing number of cases (Rodgers 1989, p. 449, 452).

Nevertheless, the reader should take note that it is possible to estimate dynamic fixed effects models if we find an instrument for Y_{it-1}. In principle, with a sufficient number of time points (e.g. four), this can be achieved by substituting Y_{it-1} with, for instance, $(Y_{it-2} - Y_{it-3})$ (see e.g. Wooldridge 2002, p. 299 ff.; Arellano and Bond 1991).[10] In this study, with its unbalanced panel data (see Chap. 6 for details) and an average number of observations per person usually below three, such a specification would dramatically decrease the multivariate sample and therefore appears unsuitable.[11]

Y_{it}^0 and Y_{it}^1 in absence of treatment. A prime example of such a situation is the famous Mathew-effect (Merton 1988, 1968) or mechanisms of cumulative advantages and disadvantages (DiPrete and Eirich 2006). Arguably, such mechanisms are rather common. A potential remedy could be the combination of the fixed effects approach with growth curve modeling, thus allowing individual trajectories—or trajectories between treatment and control group—to develop differentially over time. The implementation of such an approach is unfortunately beyond the scope of this work (as it appears more complex than the mere inclusion of nonlinear time effects (Brüderl 2010, p. 984)). Thus, as is pointed out at the end of this chapter, we have to treat statistical effect estimates with care, as they always rely on assumptions. Some of which we cannot test.

[10] Depending on the assumptions a researcher makes, two time points might be enough. Crucial in this regard is the assumption that the initial condition is exogenous: $Cov(Y_{it1}, u_i) = 0$. For an application, see e.g. Kogan (2011). If the process in question begins with the observation period, this assumption may be plausible. However, in most circumstances this assumption cannot be upheld (Stewart 2007, pp. 515–516).

[11] An important approach to causal modeling—the instrumental variable approach (see Angrist and Pischke (2008) for an encompassing discussion)—has not been discussed in this work. The reason is simple. The approach is elegant and consistently estimates causal effects if we have good instruments at hand. However, this is rarely the case. And for the work at hand, the data does not offer instruments.

5.6 Conclusion

The concrete strategy for this work's multivariate analyses should be clear. Whenever possible, I will estimate the within- and between-effect of the explanatory variables. The within-estimate has a natural affinity to the counterfactual understanding of causality and will provide us with less biased estimate of the true effect, since we can rule out all time-constant unobserved heterogeneity. The between-estimate moreover provides an approximation of the gross association between the dependent and the explanatory variable. Overall, this analysis strategy appears as an improvement to the standard choice between fixed- or random-effects models.

Still, how much can we trust the estimated effects? This chapter started by pointing to the most pressing problems in data analysis: Measurement error, unobserved heterogeneity, and simultaneity. Regarding unobserved heterogeneity and measurement error, I have argued that the former is the more pressing problem and thus proposed using methods which are geared toward dealing with it. Simultaneity, however, remains a problem even if we have longitudinal data. Why is this so? Simultaneity can obscure our analysis, because the (assumed) cause and effect are most often measured contemporaneously. Therefore, we cannot establish a strict temporal order between the measured concepts. If we believe that D_{it} has a causal impact on the value of Y_{it}, we need to assure that D_{it} forgoes Y_{it}. With most longitudinal data, we cannot assure this, because both D_{it} and Y_{it} are measured at the same point in time. And if we do not know when D_{it} and Y_{it} occurred exactly, all we can do is to assume that Y_{it} follows D_{it}. In some cases, such an assumption can be plausibly justified. This is the case if D_{it} follows a stochastic process completely external to the subjects (Singer and Willett 2003, p. 178). If this applies, we can rule out any influence of Y_{it} on D_{it}. Many situations seem to satisfy this criterion at first glance, but at second glance it is often easy to come up with potential paths of reverse causation. Consider we are interested in investigating the effect of unemployment on health (behaviors) (Schunck and Rogge 2010, 2012). There are good reasons to believe that unemployment has a causal effect on health. But does job loss really follow an exogenous, stochastic process? And can we rule out reverse causation? If a person loses her or his job due to an economic crisis, then this reason is exogenous. But it is equally possible that people with worse health are selected into unemployment (e.g. Bockerman and Ilmakunnas 2009). Moreover, the worse a person's health, the harder it might be to find a new job. Therefore, we cannot rule out the possibility that there exists a simultaneous relationship between unemployment and health. What we can do, however, is using independent variables from $t-1$ to predict the outcome of Y_{it}. This does not guarantee the correct causal ordering, but it makes it more likely. The literature suggests that we have to rely on a strong theory justifying assumptions on the causal order of events. This is indeed good advice. It applies to data analysis in general, because statistics without theory does not allow us to distinguish between artifacts and facts (Freedman 1991; Hedström 2008, p. 40). However, as the relationship between unemployment and health shows, there are many situations in which reciprocal causation is theoretically possible and plausible. This also holds for the

research question at hand. As discussed in Chap. 4, we have to assume that an immigrant's degree of integration affects her or his propensity to engage in transnational activities and that transnational involvement again affects integration.

In any case, drawing causal inference from observational data is a difficult task and we have to apply ample caution in interpreting any observed relation as causal, since we can never rule out that measurement error, unobserved heterogeneity, or simultaneity biases what appears to be a causal relationship. In the analyses, I will thus abstain from interpreting the estimated effects as causal. Some may indeed be reflecting a causal relationship, others may not. The search for the "magic bullet estimator" (Smith and Todd 2005, p. 347) is futile. All estimators rely on assumptions that may or may not be correct.

5.7 Data and Operationalization

Now that the theoretical and methodological analysis strategy has been laid out, this section describes the data against which the hypotheses are evaluated. The primary data source for this work is the German Socio-Economic Panel (hereafter SOEP). The SOEP is a longitudinal household survey in Germany, carried out annually since 1984. What makes the SOEP especially suited for analyses on immigrant integration is the fact that it is made up of different subsamples, two of which have been explicitly designed to capture the immigrant population in Germany. This oversampling ensures a sufficiently large number of cases for multivariate analyses that a proportional sample cannot provide. For example, in 2000 the SOEP successfully interviewed around 3,181 first generation and 1,200 second generation immigrants, in 2009 around 1,519 and 1,143, respectively.[12] Moreover, as the SOEP has deliberately targeted immigrants, it also includes a wide range of items on immigrant integration and even a few on transnational activities. Table 5.2 gives an overview of the SOEP's different subsamples.

When analyzing SOEP data, there are some important aspects that we have to bear in mind. First, it is very important to take into account that the different subsamples have been gathered via different sampling schemes. Second, when using longitudinal data, we have to consider panel attrition. And third, we have to reflect on how sampling and panel attrition might leave us with a specific sample and how this might impair the possibility to generalize our findings to the greater population.

With respect to the first aspect, sampling design, we have to account for the fact that the SOEP subsamples are not simple random samples of the German population. As subsamples B, D, and F provide the most respondents with a migratory background, I will focus the discussion on these samples. Detailed information on the sampling design for the other subsamples can be found in Haisken-DeNew and Frick (2003). Of

[12] Depending on how one defines the samples, the figures can vary slightly.

Table 5.2 SOEP Subsamples (Source: own computations, Kroh 2011, and Schupp and Wagner 1995)

Sample	Starting year	Description	Initial number of households
A/1	1984	German households in the FRG ("Germans-West")	4,528
B/2	1984	Foreign households in the FRG ("Foreigners-West")	1,393
C/3	1990	Households of the GDR ("SOEP-East")	2,179
D/4	1994/1995	Immigrant households in Germany ("Migrants")	236 (D1)/304 (D2)[a]
E/5	1998	Households in Germany ("Refreshment")	1,056
F/6	2000	Households in Germany ("Innovation")	6,052
G/7	2002	High income households in Germany ("High Income")	1,224
H/8	2006	Households in Germany ("Refreshment")	1,506
I/9	2009	Households in Germany ("Incentive")	1,531

[a]In 1995, 522 immigrant households were successfully interviewed. This number comprises 304 households that were first interviewed in 1995 (D2) and 218 households that were re-interviewed in 1995 coming from D1

course, other subsamples also include immigrants, but not through special sampling schemes.[13]

5.7.1 The SOEP Subsamples B, D, and F

Sample B (foreign households in FRG) covers persons in private households whose household head was either Greek, Italian, Spanish, Turkish, or Yugoslavian in 1984. It targeted the so-called 'Guestworkers' and their descendants (see Chap. 6 for a short overview of Germany's immigration experience since 1945). Subsample B itself consists of separate subsamples for each of the five nationalities. First, a random selection from counties and metropolitan areas ('Kreise und kreisfreie Städte') was drawn. These counties and metropolitan areas were the sampling units (PSU). 80 PSU were drawn for Turkish nationals and 40 for each of the other nationalities. Second, within the PSU, respondents were selected by probability sampling—i.e. systematic sampling—from official registration records (Haisken-DeNew and Frick

[13] All samples except for the high income sample (sample G/7) are used in the analyses. The latter is excluded because it makes for a very special population which is bound to differ substantially from the rest of the German population. Respondents from this sample are hardly comparable to other respondents. And since I will not use weights in the multivariate analyses, excluding this sample seems the only feasible strategy.

2003, p. 155). Overall, 1,393 and 3,169 individual respondents were interviewed successfully in the first wave.

The sampling scheme for subsample D is rather different. Subsample B was intended to cover Germany's foreign population. Due to the availability of official registration records on foreign population in Germany, addresses of eligible respondents are in principle straightforward to obtain. Sample D, however, intended to cover (naturalized) immigrants. As official registry data for naturalized immigrants is unavailable, the SOEP had to implement a different sampling strategy to cover this part of the population. Subsample D comprises yet again two subsamples, subsample D1 (1994) and subsample D2 (1995). Subsample D1 combines a random with a non-random (single level referral a.k.a. snowball) sample. The random sample includes households identified in 1992 through an address screening in the context of representative population surveys. Successfully contacted and interviewed households then provided additional addresses of other immigrants (Schupp and Wagner 1995). Of all 236 households in D1 that were successfully interviewed in 1994, 98 (41 %) were contacted via referrals from the originally identified households. Schupp and Wagner (1995, p. 18) argue that this enlargement of the sample is immune to the standard critique brought forth against snowball samples (see also Chap. 3), as the starting points for the snowball sample are not arbitrary. Instead, the snowball sample starts with randomly drawn addresses from population surveys. Moreover, it is only a single level snowball addition so that in principle the sampling process is not arbitrary and one could even estimate sampling probabilities for the referral sample. While this is true, the referral sample might still create biases in the overall subsample. Any survey on immigrants is likely to oversample assimilated immigrants (Schupp and Wagner 1995, p. 18). If we then ask those respondents to provide information on their network (i.e. collect addresses), we are likely to increase this (assimilation-) bias. Thus, with regard to integration into the receiving society, subsample D1 is likely to be a positive selection. What is more, a referral sampling creates clusters in which respondents are more similar than between clusters. Rendtel et al. (1997) thus maintain that D2 better represents the immigrant households in 1995. Subsample D2 relies on a random sampling scheme like the one in D1 but does not include a non-random part. Eligible respondents were identified through a screening procedure in a large population survey in 1994 and interviewed for the SOEP in 1995. Overall, 304 households were successfully interviewed for D2 in 1995 (see Table 5.2).

Starting in 2000, the SOEP has been enlarged considerably by including the new subsample F (see Table 5.2). Subsample F ("Innovation") covers 6,052 households (Haisken-DeNew and Frick 2003, p. 155). It was designed to include a share of foreign nationals that matches its share in the overall population. The design consists of a two stage scheme; in the first stage PSU were drawn and within these PSU, respondents were contacted via a random-route method. To correct for the usual underrepresentation of immigrants in population surveys, the inclusion probability for foreign nationals in the second stage was doubled (for details see Rosenbladt 2002). This made it possible to include 445 households with non-German members in subsample F, which makes for 7.5 % of the overall subsample. Thus, the goal of including a share of foreign nationals in the subsample that matches the share in

the population—around 8.8 % in 2000 (Statistisches Bundesamt 2010)—was almost achieved.

The differences in sampling designs are often disregarded in analyses with SOEP data, although respondents from the different subsamples may be hard to compare. This holds in particular if we are interested in descriptive analyses across different groups (and subsamples). Obviously, D1 may present a major problem in this regard. One solution to this problem might lie in an appropriate weighting scheme that accounts for these design features.

5.7.2 Cross-sectional Weighting

To meet this end, the SOEP provides cross-sectional weights for households and individuals to compensate for differences in sampling design. A detailed description on the construction of the cross-sectional weights for each wave can be found in Kroh (2009). The weights combine aspects of design-weighting, staying probabilities (see next section for more information on the staying probability), as well as post-stratification weighting. Design-weighting is necessary, as respondents in different subsamples have different inclusion probabilities. For instance, the inclusion probability for respondents in subsample A was 0.0002, whereas the inclusion probability for respondents in subsample B was four times as high (0.0008) (Haisken-DeNew and Frick 2003, p. 19). Post-stratification weights adjust the 'raw' cross-sectional weights to a set of characteristics of the underlying population in the specific years. On the household level, this encompasses information on the number of households per federal state ('Bundesland'), the district magnitude, the household size, as well as home ownership status (Kroh 2009, p. 4). On the individual level, the adjustment is achieved with respect to the marginal distributions of age, gender, and number of non-German nationals in the household. The adjustments are based on the German Microcensus, which is an annual 1 % survey of the German population.

Cross-sectional weighting may appear intuitively plausible and the decision to apply weights may appear equally straightforward. However, with longitudinal data and, in particular, with the SOEP things are more complicated. First and foremost, implementing an appropriate cross-sectional weighting scheme for the SOEP so that the waves properly represent Germany's population structure in the specific years is far from obvious and the assumptions on which the construction is based are debatable (for details see Diehl and Schnell 2006, p. 798). Weights may correct for (sampling- and attrition-) bias if they are adequately constructed. But they can also intensify bias if they are not correct. The more complex the weighting scheme, the more pitfalls there are. Second, with regard to subsample D1, it is hardly possible to compute design weights for the snowball sample (Rendtel et al. 1997, pp. 275–276; Spieß 2004, p. 10) and thus cross-sectional weights for D1 are unavailable.[14] At this

[14] Respondents from this subsample get the value zero as a weighting factor. It would be possible to compute sampling weights if respondents' personal network size and information about the recruitment process were known (Salganik and Heckathorn 2004).

point, there are two possible strategies to deal with this problem. On the one hand, we could use the available weights, assume that they are correct, and exclude sample D1. On the other hand, we could treat all waves of the SOEP as independent, unweighted samples as Diehl and Schnell (2006, p. 798) suggest. Both strategies are not optimal. If we do not weight, we ignore design aspects and (potentially systematic) non-response, which can bias our sample. But weighting may also aggravate problems if the weights are incorrect. What is more, weighting will reduce the number of respondents, as D1 will be excluded, which, in the multivariate analyses, poses a problem as the number of cases with non-missing information on all relevant variables is not very high.

Table 5.3 may provide some guidance. It presents the immigrants' and the autochthonous population's estimated age distribution based on the Microcensus and the SOEP for 2008. On average, the immigrant population is significantly younger than the autochthonous population, as the Microcensus data clearly shows. The autochthonous population's median age lies between 45 and 54 years, while that of the immigrants lies between 25 and 35 years. This may be due to the immigrants' higher fertility rate (but see Milewski 2010). Table 5.3 also shows that the SOEP's age distribution is close to that of the Microcensus, which serves as a benchmark. They do not match perfectly, as the percentage point differences indicate. Interestingly, applying the cross-sectional weights to the SOEP reduces the difference in the age distribution for the autochthonous population ($9.9 - 4.5$), measured as the sum of the absolute differences in percentage points, but it increases the difference for immigrants, albeit only slightly ($10.5 - 12.2$).

As mentioned before, Diehl and Schnell (2006, p. 798) point out that a plausible weighting scheme for the SOEP, especially for immigrants, has yet to be developed. This might explain that applying the cross-sectional weights results in better estimates for non-immigrants in the sample and slightly poorer estimates for immigrants.

In the following, this work will pursue a two-fold strategy. In the descriptive analyses, when immigrants are compared to the autochthonous population with regard to their integration into the German society, cross-sectional weights will be applied. Even if non-response adjustment by post-stratification is far from perfect, we have to control for the differences in inclusion probabilities. For the multivariate analyses, I will abstain from applying any weights. Some classes of models (logistic regressions) estimate unbiased coefficients, even if the sample is stratified on endogenous variables (Kalter 2006, p. 149; Hosmer and Lemeshow 1989, p. 177). Still, longitudinal data forces us not only to take into account the sampling process, but also the attrition process which refers to the loss of observations in the course of the survey.

5.7.3 Panel Attrition and Longitudinal Weighting

Not all households and respondents interviewed in wave t can be re-interviewed in wave $t + 1$. This process is called panel attrition. It can occur at two levels (Kohler 2002, p. 124). On the household level, the complete household may refuse to be

Table 5.3 Comparison of unweighted and weighted SOEP age distributions with the Microcensus for 2008. (Source: Microcensus 2008 (Statistisches Bundesamt 2009), SOEP 2008 ($N = 25{,}173$), own computations)

Age groups from . . . up to . . .	Microcensus 2008		SOEP 2008							
	Without migration background	With migration background	Without migration background		With migration background		Without migration background		With migration background	
			Unweighted				Weighted			
	%	%	%	Difference (pp)	%	Difference (pp)	%	Difference (pp)	%	Difference (pp)
00–05	3.3	7.4	2.7	0.6	8.6	−1.2	2.3	1.0	7.7	−0.3
05–10	3.8	7.4	4.6	−0.8	6.8	0.5	4.1	−0.3	6.6	0.7
10–15	4.1	7.1	5.3	−1.1	7.2	−0.1	4.2	0.0	6.1	1.0
15–20	5.2	7.5	6.0	−0.7	7.0	0.4	5.0	0.2	6.5	1.0
20–25	5.6	7.2	5.7	0.0	7.5	−0.3	5.3	0.4	7.4	−0.1
25–35	10.7	16.2	9.8	0.9	13.8	2.4	10.8	−0.1	15.4	0.8
35–45	15.9	15.9	15.1	0.8	15.7	0.2	15.9	−0.1	14.7	1.1
45–55	15.6	12.8	16.9	−1.2	11.1	1.7	16.5	−0.9	11.3	1.5
55–65	12.5	9.3	13.4	−0.9	9.8	−0.5	12.5	0.0	10.4	−1.1
65–75	13.5	6.0	13.8	−0.3	7.8	−1.8	14.3	−0.9	8.5	−2.5
75–85	7.3	2.6	5.6	1.7	3.6	−0.9	7.0	0.4	4.2	−1.5
85–95	2.2	0.5	1.3	0.9	0.9	−0.4	2.0	0.2	1.1	−0.6
95 and older	0.2	0.0	0.0	0.1	0.1	−0.1	0.1	0.1	0.0	0.0
Total[a]	100.0	100.0	100.0	9.9	100.0	10.5	100.0	4.5	100.0	12.2

Weighting implemented by the cross-sectional weights provided by the SOEP-group (Kroh 2009)

[a] The total is calculated as the sum of the absolute differences

re-interviewed, or may move to an unknown destination or abroad. Similarly, on the individual level, a respondent may refuse to be re-interviewed, may move abroad or to an unknown destination, or may die. Even though panel maintenance schemes can have strong impact on the amount of dropouts, panel attrition is inevitable. In the course of the survey, subsample B, for instance, decreases from 1,393 households in 1984 to 500 in 2009. Although it is undesirable because it decreases the sample size, it remains rather unproblematic for the analysis if the attrition is at random. This would require that the dropouts are a random subsample of the original sample. Panel attrition becomes problematic, however, if it is non-random. An example would be that unemployed respondents are more likely to drop out of the panel than employed respondents. This could have various reasons. For instance, a new job might require previously unemployed respondents to move and thus they drop out of the survey. Or unemployment by itself might decrease a respondent's willingness to participate in the survey and thus cause the dropout. In this case, the probability to drop out of the sample is related to relevant characteristics and we are left with a selective sample from which it becomes more difficult to generalize to the population. If panel attrition is following a systematic pattern, we can try to control for this process by weighting the observations with longitudinal weights, giving more weight to those respondents who are less likely to remain in the sample. This, of course, requires us to understand and model the attrition process, e.g. to identify the relevant factors which are driving attrition. The SOEP provides longitudinal weights, which are the product of the inverse of the staying probabilities. These staying probabilities are estimated via logit models, predicting successful follow ups and interviews at $t + 1$ with covariates from t (for details, see Kroh (2011) and Haisken-DeNew and Frick (2003)).

As for cross-sectional weights, longitudinal weights can improve our estimates if the weights are constructed correctly. If not, they, too, will increase bias. As Brüderl (2010, p. 993) points out, weighting might not even be necessary to correct for systematic panel attrition if attrition is driven by observables. Controlling these observed variables in the analyses (i.e. in the regression models) can correct for the attrition process. In addition, fixed effects estimates are also robust to attrition that is driven by time-constant unobserved variables. Moreover, in the context of random and fixed effects models, implementing the longitudinal weights is not straightforward. By construction, these weights differ for each wave. Standard fixed effects estimations, however, require that these weights are constant across individual observations (Bjerk 2009, p. 409). As a consequence, I will abstain from longitudinal weighting. We do not know whether more complicated modeling approaches really improve our estimates, but they certainly increase potential sources of error (Brüderl 2010, p. 993).

From the above discussion it follows that I use an *unbalanced* panel design when estimating the statistical models. Unbalanced means that the number of observations may differ between respondents, as some are longer in the SOEP than others. A *balanced* panel design would require each respondent being observed at *every* time point. Obviously, panel attrition is one source of unbalanced observation across respondents. But in the SOEP this is also due to the fact that it combines different

subsamples that have been included in different years. Since a balanced panel design would drastically reduce the number of cases, its implementation does not appear sensible (see Wooldridge 2002, p. 577 ff. for a discussion). It may appear obvious to assign more weight to those respondents with fewer observations, but I refrain from doing so for the reasons enumerated above.

5.7.4 Operationalization

In the following, I will discuss the variables used in this study. In order not to prolong this chapter unnecessarily, I will abstain from a detailed description of all variables in the study and instead concentrate on the most important. A detailed list of all variables used can be found in table A.1 in the appendix.

Immigrants in the SOEP are identified via country of origin and nationality. This allows us to identify second generation immigrants who have acquired the German nationality. This work differentiates between first generation immigrants and second (and later) generation immigrants. The former are those who have migrated themselves, whereas the latter are direct descendants of immigrants. However, the second generation also includes those who have migrated themselves but arrived in Germany before the age of 6. This common procedure (e.g. Kalter 2006) is motivated by the fact that these immigrants are likely to pass through the German schooling system. Part of their secondary socialization will thus take place in the receiving country and its core educational institutions. Among the first generation, this work differentiates between immigrants from Greece, Italy, Poland, Spain and Portugal, Turkey, former Yugoslavia, other Eastern European Countries, other Western European Countries, other countries, and no answer.[15] Among the second generation, it is unfortunately not possible to unambiguously identify the parents' country of origin. In part, this is possible if the parents are interviewed in the SOEP, too (this is the case if the respondents live in one household). Unfortunately, this is not the case for many respondents. Consequently, the second generation has Germany as its country of origin. For the later analysis, this implies that one cannot compute joint models for first and second generation in which both generational status and country of origin are controlled for. Similarly to the country of origin, the respondents can also be distinguished by their nationalities, i.e. German, Greek, Italian, Polish, Spanish and Portuguese, Turkish, former Yugoslavian, nationalities from other Eastern European Countries, from other Western European Countries, and from other countries.

The analyses in general will put more emphasis on the first generation. This has a number of reasons. For one, the number of cases for the second generation is considerably lower and a detailed analysis of the second generation's transnational

[15] The latter is included as a separate category for most variables if the number of cases is high enough to justify a separate category. This is done also for other variables, e.g. CASMIN, to efficiently use the SOEP cases. The alternative would be listwise exclusion, which would considerably lower the number of cases in the multivariate analysis.

involvement is, in some instances, hard to realize. Moreover, the sample of second generation immigrants is rather heterogeneous regarding their inclusion into the SOEP. Some have been included directly through the sampling, while some grew up in (first generation immigrant) households and became part of the SOEP when reaching the 17th birthday (Wagner et al. 2007). As such, the sample of the second generation immigrants may not be fully representative of the second generation living in Germany.

The most important theoretical and empirical constructs of this work are certainly those capturing immigrants' transnational activities. Fortunately, the SOEP collects information on a set of important transnational activities: first, on remittances sent (abroad) to relatives and friends (Kivisto and Faist 2010, p. 140, 150 ff.; Haller and Landolt 2005; Waldinger 2008) and, second, on regular visits to the country of origin (O'Flaherty et al. 2007; Waldinger 2008; Haller and Landolt 2005). Remittances are typically considered to be an important aspect of transnational involvement. Remittance-arrangements are most likely within the context of families or extended kinship networks, as part of a household strategy to reduce risks (Massey 1990; Stark 1991; Stark and Bloom 1985; Landolt 2001). But they can also extend beyond close-kinship networks (Kivisto and Faist 2010, p. 141), being motivated by other than those quasi contractual arrangements of household risk diversification (Vanwey 2004), and may indicate emotional attachment to those left behind. Visiting the country of origin is a very tangible aspect of border-crossing involvement which, and as such, can be seen as a particularly important aspect of being transnationally active. As O'Flaherty et al. (2007, p. 820) put it, being in the country of origin in person is phenomenologically quite different from other forms of (electronically) mediated interaction. Hence, interacting with significant others in the country of origin is bound to be very important in evaluating investment decisions. Of course, these two dimensions of border-crossing involvement and the relating indicators cannot compete with the information on border-crossing involvement provided by a study like the CIEP (e.g. Portes 2001), for instance. But they very well match indicators used in other studies on transnational involvement, as discussed in Chap. 3 (Haller and Landolt 2005; O'Flaherty et al. 2007; Waldinger 2008).

An issue to consider with longitudinal data is that not every item is included annually. Some questions are asked biennially. The questions on a respondent's self-assessed language proficiency are, for example, included in the SOEP (after 1993) only in the years 1993, 1995, 1997, 1999, 2001, 2003, 2005, 2007, and 2009. One could limit the analysis to the years in which the relevant information was collected. This, however, severely limits the analysis, because it drastically reduces the sample and, more importantly, rules out analyses on characteristics that are included in the survey alternately. Alternatively, we can make a virtue out of necessity and can use leads and lags of the respective variables in the analyses, which offers some protection against simultaneity. For instance, a set of items on visits to the country of origin within the last two years prior to the interview has been included in the SOEP starting 1996. Since then, it has been part of the SOEP questionnaire every two years, i.e. 1996, 1998, . . . , 2006, 2008. Information on language proficiency is included biennially but in odd years. Thus, we have information on language proficiency for

the years 1995, 1997, . . . , 2007, 2009. If we are interested in estimating the effect of visits to the country of origin on German language proficiency, then we will estimate the effect of visits to the country of origin from $t-1$ (e.g. 2004) on German language proficiency at t (e.g. 2005). If, conversely, we are interested in estimating the effect of German language proficiency on the probability to visit the country of origin, then we estimate the effect of language proficiency at time t (e.g. 2005) on visits to the country of origin at $t+3$ (e.g. 2008)–recall that the items on visits refer to the last two years before the interview.[16] This strategy is chosen for all variables that are collected biennially.

The time lag between the independent and the dependent variables might appear long. In this example, the longest possible lag between the two variables is just under 3 years. However, this structure is necessary to ensure the correct temporal order between independent and dependent variables. How can such a lag influence the analyses? If there is a lag between exposure and effect, we have to rely on the assumption that effect will be observable after the lag. In some contexts such an assumption can be problematic. In the context of this work, however, it appears unproblematic. The theoretical perspective laid out above (Chap. 4) is compatible with the assumption of delayed responses. Investment decisions over the life course are surely shaped by more than the immediate experiences and opportunity structure prior to the decision. Quite the opposite: deciding to invest into a specific form of capital surely depends on previously accumulated capitals and previous experiences, as they also shape the expectations regarding gains and realization probabilities. Thus, it is a sound assumption to say that transnational activities in the last 2 years influence present immigrant integration (and vice versa). Besides, most variables are measured annually so that this problem only presents itself with a few indicators.

A related and potentially more severe problem presents itself for three other (sets of) variables (ethnic composition of one's network, the location of relatives, and neighborhood characteristics). Information on these items is only collected every few years (see Table A.1 in the appendix). Either we limit the analyses to the respective years or we have to fill in the missing values. The first option is out of question, as this would limit the analyses to at most three time points with considerable lags between them. With respect to the second option, if we perceive this as a missing data problem, we could try to impute the missing data from the available information (e.g. Allison 2000). The implementation of an elaborate imputation scheme, especially for longitudinal data, however, goes beyond the scope of this work. Thus, an alternative way of filling in the missing values is chosen: they are replaced by the values of the last available measurement. This procedure is admittedly not the best choice, although it appears to be commonly used (see e.g. Kalter 2006, p. 150). However, this is only done when these variables are used as predictors. And, in the case of neighborhood characteristics, additional information on moves is used to ensure that respondents are still living in the neighborhood their answers referred to. What are

[16] This example indicates that we are actually facing an endogenous causal structure, which is at odds with the assumption of certain panel models (such as fixed effects models). See Chap. 9 for a discussion.

the consequences of this procedure for the analysis? This is actually a measurement problem. By using last year's information as a proxy for the current year, we increase measurement error. Measurement error may bias our estimates. But measurement error is likely to create a 'conservative,' i.e. downward, bias, which is in any case preferable to an undefined bias due to unobserved heterogeneity. Thus, although this procedure is unlikely to obscure the estimates—and in the case of neighborhoods even seems well justified as most neighborhoods in Germany do not change completely within 4–5 years—we have to keep this in mind when conducting the analysis.

5.8 Theoretical and Empirical Implications of the Data

Besides the methodological issues discussed above, we should also reflect on how the data we have matches the theory we use and whether the data captures the population of interest. The group of immigrants which is most likely to be in a survey like the SOEP is at least partially permanent immigrants. Sojourners and pendular labor migrants are less likely to be part of such a survey. Thus, the immigrant sample in the data used is selective in the sense that the longer the residence in Germany, the higher the probability of being in the sample (Rendtel et al. 1997, p. 189). Some scholars might argue that precisely the group which has been excluded from the study is most relevant with respect to transnational involvement. This might well be the case. Still, it is very informative to investigate transnational activities among a population in which they are by default less likely, i.e. more permanent immigrants. Of course, the same line of criticism regarding the sample selection applies when investigating the relationship between transnational involvement and immigrant integration. One might say that it is uninformative to look at those immigrants who intend to settle permanently in the receiving country, because they are more likely to follow conventional paths of immigrant integration. Thus, inference regarding the overall validity and explanatory scope of the theoretical frame can be hard to justify. This is a valid critique, which has to be carefully addressed by studies on transnational involvement and immigrant integration. But a strength of the proposed theoretical model in Chap. 4 is its universality. Reconstructing the integration process as investment decisions that bounded rational actors make, facing a specific opportunity structure, can be easily extended to forms of temporary migration. Moreover, the theoretical model's core has been applied numerous times to explain the actual migration process (Esser 1980; Kley 2010; De Jong et al. 1983; De Jong and Gardner 1981; Huinink and Kley 2008) and even circular migration from Poland to Germany (Kalter 2011).

There are, nonetheless, important empirical and theoretical aspects to consider. First, the Comparative Immigrant Entrepreneurship Project (CIEP), which does not restrict its sample to (semi)permanent immigrants, has shown that transnational modes of making a living are uncommon among immigrants (see e.g. Portes et al. 2002). Second, although other forms of transnational involvement, be they political or socio-cultural, are more common among immigrants than assumed by critiques of

the concept, this is not limited to pendular migrants. Instead, permanent and semi-permanent immigrants do engage in these activities (see e.g. O'Flaherty et al. 2007; Portes 2003). But with respect to immigrant integration, one can as well argue that those immigrants who (intend) to stay for longer periods of time in a receiving country make up the population of interest. When it comes to issues of integration, this group is at the center of interest in the debates in science, politics, and media. Moreover, because a huge part of the non-autochthonous population living in immigration countries are permanent or semi-permanent migrants, it seems indeed reasonable to investigate transnational involvement among them.

References

Alba, R. D. (2008). Why we still need a theory of mainstream assimilation. *Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie, Sonderheft, 48,* 37–56.

Alba, R. D., & Nee, V. (2003). *Remaking the American mainstream: Assimilation and contemporary immigration*. Cambridge: Harvard University Press.

Allison, P. D. (1994). Using panel-data to estimate the effects of events. *Sociological Methods & Research, 23*(2), 174–199.

Allison, P. D. (2000). Multiple imputation for missing data. A cautionary tale. *Sociological Methods & Research, 28*(3), 301–309.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Cary: SAS Institute.

Allison, P. D. (2009). *Fixed effects regression models* (Quantitative applications in the social sciences, Vol. 160). Los Angeles: SAGE.

Allison, P. D., & Waterman, R. P. (2002). Fixed-effects negative binomial regression models. *Sociological Methodology, 32*(1), 247–265.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics. An empiricist's companion*. Princeton: Princeton University Press.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data—Monte-Carlo Evidence and an application to employment equations. *Review of Economic Studies, 58*(2), 277–297.

Arrow, K. J. (1971). *Some models of racial discrimination in the labor market*. Santa Monica: Rand.

Baltagi, B. H. (2008). *Econometric analysis of panel data* (4th ed.). Chichester: Wiley.

Begg, C. B., & Gray, R. (1984). Calculation of polychotomous logistic-regression parameters using individualized regressions. *Biometrika, 71*(1), 11–18.

Bjerk, D. (2009). How much can we trust causal interpretations of fixed-effects estimators in the context of criminality? *Journal of Quantitative Criminology, 25*(4), 391–417.

Bockerman, P., & Ilmakunnas, P. (2009). Unemployment and self-assessed health: Evidence from panel data. *Health Economics, 18*(2), 161–179.

Bollen, K. A. (1989). *Structural equations with latent variables* (Wiley series in probability and mathematical statistics. Applied probability and statistics). New York: Wiley.

Boudon, R. (1980). *Die Logik des gesellschaftlichen Handelns. Eine Einführung in die soziologische Denk- und Arbeitsweise*. Neuwied: Luchterhand.

Bound, J., & Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data. Do 2 wrongs make a right. *Journal of Labor Economics, 9*(1), 1–24.

Brady, H. E., & Collier, D. (2004). *Rethinking social inquiry: Diverse tools, shared standards*. Lanham: Rowman & Littlefield.

Brand, J. E., & Halaby, C. N. (2006). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. *Social Science Research, 35*(3), 749–770.

Brüderl, J. (2010). Kausalanalyse mit Paneldaten. In C. Wolf & H. Best (Eds.), *Handbuch der Sozialwissenschaftlichen Datenanalyse* (pp. 963–994). Wiesbaden: VS Verlag für Sozialwissenschaften.

Burr, J. A., & Nesselroade, J. R. (1990). Change measurement. In A. von Eye (Ed.), *Statistical methods in longitudinal research* (pp. 3–34). Boston: Academic Press.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using stata*. College Station: Stata Press.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies, 47*(1), 225–238.

De Jong, G. F., & Gardner, R. W. (1981). Motives for migration: A assesment and a value-expectancy research model. In G. F. De Jong & R. W. Gardner (Eds.), *Migration decision making. Multidisciplinary approaches to microlevel studies in developed and developing countries* (pp. 13–58) New York: Pergamon Press.

De Jong, G. F., Abad, R. G., Arnold, F., Carino, B. V., Fawcett, J. T., & Gardner, R. W. (1983). International and internal migration decision making: A value-expectancy based analytical framework of intentions to move from rural Philippine Province. *International Migration Review, 17*(3), 470–484.

Diehl, C., & Schnell, R. (2006). "Reactive Ethnicity" or "Assimilation"? statements, arguments, and first empirical evidence for labor migrants in Germany. *International Migration Review, 40*(4), 786–816.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and emprical developments. *Annual Review of Sociology, 32*(1), 271–297.

Draper, D. (2008). Bayesian Multilevel Analysis and MCMC. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 77–139). New York: Springer.

Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.

Engel, U., & Reinecke, J. (1994). *Panelanalyse. Grundlagen, Techniken, Beispiele*. Berlin: Walter de Gruyter.

Esser, H. (1980). *Aspekte der Wanderungssoziologie*. Darmstadt: Luchterhand.

Esser, H. (1999). *Soziologie. Allgemeine Grundlagen* (3rd ed.). Frankfurt a. M.: Campus.

Finkel, S. E. (1995). *Causal analysis with panel data* (Sage university papers series. Quantitative applications in the social sciences no. 07-105). Thousand Oaks: Sage Publications.

Freedman, D. A. (1991). Statistical-models and shoe leather. *Sociological Methodology, 21,* 291–313.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology, 36*(1), 21–47.

Gangl, M., & DiPrete, T. A. (2004). Kausalanalyse durch Matchingverfahren. *Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie, Sonderheft, 44*, 396–420.

Giesselmann, M., & Windzio, M. (2012). *Regressionsmodelle zur Analyse von Paneldaten*. Wiesbaden: Springer VS.

Goldthorpe, J. H. (2001). Causation, statistics, and sociology. *European Sociological Review, 17*(1), 1–20.

Greene, W. H. (2000). *Econometric analysis* (3rd ed.). New Jersey: Prentice-Hall.

Griliches, Z., & Hausman, J. A. (1986). Errors in variables in panel data. *Journal of Econometrics, 31*(1), 93–118.

Haisken-DeNew, J. P., & Frick, J. R. (2003). Desktop companion to the German Socio-Economic Panel study (GSOEP). German Institute for Economic Research, Berlin.

Halaby, C. N. (2003). Panel models for the analysis of change and growth in life course studies. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the life course* (pp. 503–528). New York: Kluwer Academic/Plenum Publishers.

Halaby, C. N. (2004). Panel models in sociological research: Theory in practice. *Annual Review of Sociology, 30,* 507–544.

Haller, W., & Landolt, P. (2005). The transnational dimensions of identity formation: Adult children of immigrants in Miami. *Ethnic and Racial Studies, 28*(6), 1182–1214.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251–1271.

Hausman, J. A., & Wise, D. A. (1981). Stratification on endogenous variables and estimation: The gary income maintenance experiment. In C. F. Manski & D. McFadden (Eds.), *Structural analysis of discrete data with econometric applications* (pp. 365–391). Cambridge: MIT Press.

Heckman, J. J., & Robb, R. (1989). The value of longitudinal data for solving problems of selection bias in evaluating the impact of treatments on outcomes. In D. Kasprzyk, G. J. Duncan, G. Kalton, & M. P. Singh (Eds.), *Panel surveys* (pp. 512–538). New York: Wiley.

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1996). Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. *Proceedings of the National Academy of Sciences of the United States of America, 93*(23), 13416–13420.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605–654.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*(2), 261–294.

Hedström, P. (2005). *Dissecting the social. On the principles of analytical sociology*. Cambridge: Cambridge University Press.

Hedström, P. (2008). *Anatomie des Sozialen. Prinzipien der Analytischen Soziologie*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression* (Wiley series in probability and mathematical statistics. Applied probability and statistics). New York: Wiley.

Hox, J. J. (2002). *Multilevel analysis. Techniques and applications* (Quantitative methodology series). Mahwah: Lawrence Erlbaum Associates.

Hsiao, C. (2003). *Analysis of panel data* (2nd ed., Econometric Society monographs no. 34). Cambridge: Cambridge University Press.

Huinink, J., & Kley, S. (2008). Regionaler Kontext und Migrationsentscheidungen im Lebensverlauf. *Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie, Sonderheft, 48,* 162–184.

Jones, A. M., Rice, N., Bago d'Uva, T., & Balia, S. (2007). *Applied health economics* (Routledge advanced texts in economics and finance, Vol. 8). Milton Park: Routledge.

Kalter, F. (2006). Auf der Suche nach einer Erklärung für die spezifischen Arbeitsmarktnachteile Jugendlicher türkischer Herkunft. Zugleich eine Replik auf den Beitrag von Holger Seibert und Heike Solga "Gleiche Chancen dank einer abgeschlossenen Ausbildung?". *Zeitschrift Fur Soziologie, 35*(2), 144–160.

Kalter, F. (2011). Social capital and the dynamics of temporary labour migration from Poland to Germany. *European Sociological Review, 27*(5), 555–569.

Kaufman, R. L. (1993). Decomposing longitudinal from cross-unit effects in panel and pooled cross-sectional designs. *Sociological Methods & Research, 21*(4), 482–504.

Kempthorne, O. (1978). A biometrics invited paper: Logical, epistemological and statistical aspects of nature-nurture data interpretation. *Biometrics, 34*(1), 1–23.

Kivisto, P., & Faist, T. (2010). Beyond a border. The causes and consequences of contemporary immigration. Los Angeles: Pine Forge.

Kley, S. (2010). Explaining the stages of migration within a life-course framework. *European Sociological Review,* 1–18. doi:10.1093/esr/jcq020.

Kogan, I. (2011). New immigrants—old disadvantage patterns? Labour market integration of recent immigrants into Germany. *International Migration, 49*(1), 91–117.

Kohler, U. (2002). *Der Demokratische Klassenkampf. Zum Zusammenhang von Sozialstrukur und Parteipräferenz*. New York: Campus.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*(1), 1.
Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling* (ISM: introducing statistical methods). London: Sage.
Kroh, M. (2009). Short-Documentation of the Update of the SOEP-Weights, 1984–2008. pp. 1–5
Kroh, M. (2011). Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) (1984 until 2010). *DIW Data Documentation, 59*, 1–55.
Kühnel, S. M., & Krebs, D. (2010). Multinomiale und ordinale Regression In C. Wolf & H. Best (Eds.), *Handbuch der Sozialwissenschaftlichen Datenanalyse* (pp. 855–886). Wiesbaden: VS Verlag für Sozialwissenschaften.
Lancaster, T. (2000). The incidental parameter problem since 1948. [Proceedings Paper]. *Journal of Econometrics, 95*(2), 391–413.
Landolt, P. (2001). Salvadoran economic transnationalism: Embedded strategies for household maintenance, immigrant incorporation, and entrepreneurial expansion. *Global Networks, 1*(3), 217–241.
Liker, J. K., Augustyniak, S., & Duncan, G. J. (1985). Panel data and models of change—a comparison of 1st difference and conventional 2-wave models. *Social Science Research, 14*(1), 80–101.
Long, S. J. (1997). *Regression models for categorical and limited dependet variables*. Thousand Oaks: SAGE.
Long, S. J., & Freese, J. (2006). *Regression models for categorical dependent variables using stata* (2nd ed.). College Station: StataCorp LP.
Maddala, G. S. (1987). Limited dependent variable models using panel data. *Journal of Human Resources, 22*(3), 307–338.
Massey, D. S. (1990). Social structure, household strategies, and the cumulative causation of migration. [Proceedings Paper]. *Population Index, 56*(1), 3–26.
Merton, R. K. (1968). The Matthew effect in science. The reward of communication systems of science are considered. *Science, 159*(3810), 56–63.
Merton, R. K. (1988). The Matthew effect in science II. Cumulative advantage and the symbolism of intellectual property. *Isis, 79*(299), 606–623.
Milewski, N. (2010). Immigrant fertility in West Germany: Is there a socialization effect in transitions to second and third births? *European Journal of Population-Revue Europeenne De Demographie, 26*(3), 297–323.
Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*(1), 67–82.
Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects—prospects and pitfalls in theory and practice. *Sociological Methods & Research, 35*(1), 3–60.
Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: methods and principles for social research* (Analytical methods for social research). New York: Cambridge University Press.
Mundlak, Y. (1978). Pooling of time-series and cross-section data. *Econometrica, 46*(1), 69–85.
Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics, 54*(2), 638–645.
Neuhaus, J. M., & Lesperance, M. L. (1996). Estimation efficiency in a binary mixed-effects model setting. *Biometrika, 83*(2), 441–446.
Neuhaus, J. M., & McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 68,* 859–872.
Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*(1), 1–32.
O'Flaherty, M., Skrbis, Z., & Tranter, B. (2007). Home visits: Transnationalism among Australian immigrants. *Ethnic and Racial Studies, 30*(5), 817–844.

Palta, M., & Seplaki, C. (2002). Causes, problems and benefits of different between and within effects in the analysis of clustered data. *Health Services and Outcomes Research Methodology, 3,* 177–193.

Portes, A. (2001). Introduction: The debates and significances of immigrant transnationalism. *Global Networks, 1*(3), 181–193.

Portes, A. (2003). Conclusion: Theoretical convergencies and empirical evidence in the study of immigrant transnationalism. *International Migration Review, 37*(3), 874–892.

Portes, A., Haller, W. J., & Guarnizo, L. E. (2002). Transnational entrepreneurs: An alternative form of immigrant economic adaptation. *American Sociological Review, 67*(2), 278–298.

Rabe-Hesketh, S., & Skrondal, A. (2005). *Multilevel and longitudinal modeling using stata*. Texas.

Raudenbush, S. W. (1989). "Cenetering" predictors in multilelvel analysis: Choice and aonsequences. *Multilevel Modelling Newsletter, 1*(2), 10–12.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed., Advanced quantitative techniques in the social sciences 1). Thousand Oaks: Sage Publications.

Rendtel, U., Panneberg, M., & Daschke, S. (1997). Die Gewichtung der Zuwanderer-Stichprobe des Sozio-oekonomischen Panels (SOEP). *Vierteljahrshefte zur Wirtschaftsforschung/Quarterly Journal of Economic Research, 66*(2), 271–286.

Rodgers, W. L. (1989). Comparisons of alternative approaches to the estimation of simple causal models from panel data. In D. Kasprzyk, G. J. Duncan, G. Kalton, & M. P. Singh (Eds.), *Panel surveys* (pp. 432–456). New York: Wiley.

Rosenbaum, P. R. (1999). Using quantile averages in matched observational studies. *Journal of the Royal Statistical Society Series C-Applied Statistics, 48,* 63–78.

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics, 41*(1), 103–116.

Rosenbladt. (2002). Erprobung innovativer Erhebungskonzepte für Haushalts-Panel-Stichproben. Erstbefragung 2000 der SOEP-Stichprobe F. Methodenbericht. Infratest Sozialforschung (pp. 1–23).

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics, 29*(1), 159–183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing.1. Some examples. *Biometrics, 32*(1), 109–120.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331.

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology, 34,* 193–239.

Schunck, R. (2009). Selbstständigkeit von Migranten in Deutschland: Die Effekte sozialer Einbettung. Paper presented at the 6. Nutzerkonferenz Forschung mit dem Mikrozensus, GESIS Mannheim.

Schunck, R. (2013). Within- and between-estimates in random effects models. Advantages and drawbacks of correlated random effects and hybrid models. *Stata Journal, 13*(1), 65–76.

Schunck, R., & Rogge, B. G. (2010). Unemployment and its association with health-relevant actions. Investigating the role of time perspective with German census data. *International Journal of Public Health, 55*(4), 271–278.

Schunck, R., & Rogge, B. G. (2012). No causal effect of unemployment on smoking? A German panel study. *International Journal of Public Health,* 57(6), 867–874.

Schupp, J., & Wagner, G. (1995). Die Zuwanderer-Stichprobe des Sozio-oekonomischen Panels (SOEP). *Vierteljahrshefte zur Wirtschaftsforschung/Quarterly Journal of Economic Research, 64*(1), 16–25.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis. Modeling change and event occurrence*. Oxford: Oxford University Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*(1–2), 305–353.

Snijders, T. A., & Bosker, R. J. (2004). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage.

Spieß, M. (2004). Derivation of design weights: The case of the German socio-economic panel (SOEP). *DIW Data Documentation, 8,* 1–21.

Stark, O. (1991). *The migration of labor*. Cambridge: B. Blackwell.

Stark, O., & Bloom, D. E. (1985). The new economics of labor migration. *American Economic Review, 75*(2), 173–178.

StataCorp L. P. (2007). *Stata longitudinal/panel data. Reference Manual, Release 10*. College Station: StataCorp LP.

Statistisches Bundesamt. (2009). Bevölkerung mit Migrationshintergrund—Ergebnisse des Mikrozensus 2008—Fachserie 1 Reihe 2.2–2008. Wiesbaden.

Statistisches Bundesamt. (2010). Bevölkerung und Erwerbstätigkeit 2009—Fachserie 1 Reihe 2. Wiesbaden.

Stewart, M. B. (2007). The interrelated dynamics of unemployment and low-wage employment. *Journal of Applied Econometrics, 22*(3), 511–531.

Tuomela, R. (1976). Explanation and understanding of human behaviour. In J. Manninen & R. Tuomela (Eds.), *Essays on explanation and understanding: Studies in the foundations of humanities and social sciences* (pp. 183–208). Dordrecht: D. Reidel Pub. Co.

Vanwey, L. K. (2004). Altruistic and contractual remittances between male and female migrants and households in rural Thailand. *Demography, 41*(4), 739–756.

Wagner, G. G., Frick, J. R., & Schupp, J. (2007). The German socio-economic panel study (SOEP)—scope, evolution and enhancement. *Schmollers Jahrbuch, 127*(1), 139–169.

Waldinger, R. (2008). Between "Here" and "There": Immigrant cross-border activities and loyalties. *International Migration Review, 42*(1), 3–29.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25,* 659–706.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics, 87*(2), 385–390.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge: MIT Press.

Wright, G. H. von (1971). *Explanation and Understanding* (International library of philosophy and scientific method). London: Routledge and K. Paul.