Discussion Paper No. 17-031

An Investigation of Record Linkage Refusal and Its Implications for Empirical Research

Arne Jonas Warnke

ZEW

Zentrum für Europäische Wirtschaftsforschung GmbH

Centre for European Economic Research

Discussion Paper No. 17-031

An Investigation of Record Linkage Refusal and Its Implications for Empirical Research

Arne Jonas Warnke

An Investigation of Record Linkage Refusal and Its Implications for Empirical Research*

Arne Jonas Warnke†

August, 2017

Abstract

Linking survey data to administrative records provides access to large quantities of information such as full employment biographies. Although this practice is becoming increasingly common, only a small number of studies in the field of social sciences have thus far investigated the variables associated with linkage consent. These studies have produced diverging results with regard to the relevance of certain characteristics for the provision or non-provision of linkage consent. In this study, we analyze two comparable German datasets, thereby shedding new light on the possible reasons for previously inconsistent results. This is also the first study in which possible linkage consent bias is investigated in applied models, via the replication of an existing study for the sample in which respondents did not consent to data linkage. Whilst similar results are found between standard socio-demographic variables and linkage consent, there are considerable inconsistencies between the comparable datasets in terms of variables such as individual personality traits and work satisfaction. Overall, however, the results are promising – results do not differ much where respondents who did not provide linkage consent are considered.

JEL Classification: C18, C83
Keywords: Linkage Consent, Consent Bias, Administrative Records, Record Linkage, Linked-Employer-Employee Data, Survey Data, Selection Bias

---

*Special thanks go to Bernd Fitzenberger, Martin Kiefel, François Laisney, Antoni Manfred, André Nolte and seminar participants at the 7th Conference of the European Survey Research Association (ESRA), Humboldt University, ZEW Mannheim and in the workshop "Nonresponse Bias: Qualitätssicherung sozialwissenschaftlicher Umfragen" at DIW Berlin. I am grateful to Antonia Entorf, Sven Giegerich, Cecilia Großmann, Timo Haller and Elsbeth Wright for their outstanding research assistance.

†ZEW Centre for European Economic Research Mannheim. E-mail: warnke@zew.de.

1 Introduction

The linkage of survey data to administrative sources ("record linkage") facilitates access to reliable data relating to near complete employment biographies or extensive medical records. This in turn enables survey institutes to reduce the length of interviews and to gain access to large quantities of information which are generally relatively free of measurement errors. It is for these reasons that record linkage has become an increasingly popular tool in research and in medical research (e.g. Künn, 2015).

Data privacy laws in many countries require that polling institutes obtain prior consent to link survey data to administrative sources or to medical records. Such datasets span almost entire employment biographies, providing detailed information about periods of unemployment. Similar considerations apply to medical records, which often contain sensitive personal information. Whilst individuals who prefer to maintain their privacy may indeed be willing to participate in a survey interview, they may be reluctant to share such sensitive information with other parties. It is often the case that this reluctance is non-negligible; consent rates have been found to be as low as 40% (Bates, 2005). Final datasets in which survey data is linked to official records will not include information about respondents who do not consent to record linkage. Non-consent can therefore be viewed as a new form of non-response.

Several studies documented in the medical literature, as well as a number of studies carried out in the field of social sciences, have investigated linkage consent. These studies have so far indicated little consistency in terms of predictors of linkage consent. Whilst a number of standard socio-demographic characteristics such as respondents' age or sex have been shown to have a statistically significant positive correlation with linkage consent in one study, the same predictors have been shown to have a statistically significant negative association in other studies (see Section 2). Furthermore, there is growing evidence that the transferability of results from the medical literature to social sciences, and vice versa, is limited (Jenkins et al., 2006).

This study seeks to shed light on the possible reasons for inconsistent findings on predictors of linkage consent, as documented in the literature. To this end, we compare two very similarly structured datasets from the same country. In the two datasets, both of which were collected in surveys conducted by the same polling institute, workers in different establishments were asked questions about work-related aspects relevant for social science research. We first use the same set of controls for both datasets, thereby confirming that the two studies are broadly comparable. Secondly, we add further variables to the datasets to see whether varying the set of controls gives rise to inconsistent results. These additional variables are not necessarily available in both datasets and include psychological attributes as well as job and firm characteristics. In a subsequent step, we make use of the matched employer-employee structure of the available data to investigate whether the individual consent decision is driven by the work environment.

Finally, we hope to answer the question of whether similar results would have been obtained if the analysis had not been restricted to the sample of respondents who provided linkage consent. Suppose our target population consists of all survey participants regardless of their linkage consent decision. Our sample consists of respondents from whom linkage consent was obtained. Applied researchers are mostly concerned whether one can use the sample at hand to derive consistent estimators for statistics of the target population without strong assumptions (see, for example, Solon et al., 2015). This is the case if the association between linkage consent decision and the outcome of interest depends only on observable characteristics. In this case, it is possible to derive consistent population statistics from the sample by adding these observable characteristics to the regression or by weighting. In contrast, if unobserved heterogeneity is correlated with linkage consent (and the outcome) it is necessary to make stronger assumptions.

We investigate this question by testing whether results for two economic models would have been different if data on individuals who refused to provide linkage consent had in fact been available. The first model is a replication of our own earlier research on participation in job-related training. In the original study, we excluded information on survey participants who did not provide linkage consent. In this study we replicate our original results using the survey data only. We investigate whether we would have drawn different conclusions if we had included survey participants who did not provide linkage consent (the non-consent sample). In the second model, we estimate an augmented Mincer-regression, a "cornerstone of empirical economics" (Heckman et al., 2003), to see whether different samples give different findings on wage returns to human capital investments.

In general, our findings regarding the role of (denied) linkage consent for applied research are rather promising. It is clear that linkage consent is not independent of many individual characteristics. In particular, younger and white-collar workers (in Germany) are more reluctant to share their administrative data. The decision to provide linkage consent is closely related to respondents' willingness to participate in future interviews or to select non-response when they are asked to report their income. This does not, however, seem to translate into a large bias in economic models in our two applications. Looking at the non-consent sample provides us with more or less similar results. Furthermore, while there is some form of establishment heterogeneity in terms of employees' tendency to consent to data linkage, this can mostly be explained by observable differences between workforces. The location of a firm for example, seems to be a determining factor for linkage consent. The workforces of establishments located in East Germany for example, exhibit much higher consent rates than the workforces of similar establishments in West Germany. Workforces consisting of higher numbers of white-collar workers, however, tend to have lower consent rates.

The structure of the paper is the following. In the next section we present the literature concerning linkage consent. In Section 3 we introduce the research questions to be answered in this

study. Section 4 describes the two datasets used here. Results are presented in Section 5 to 6 before we finally conclude.

2 Literature

Antoni (2011), Sakshaug et al. (2012) and Sakshaug and Kreuter (2012) review the literature on linkage consent bias and show that relevant predictors vary considerably between studies. Their reviews include studies which found significant positive, as well as significant negative associations between linkage consent and characteristics such as age, level of education, sex and income. We give a non-exhaustive summary of recent studies in Table 1.

The exact reasons for these inconsistent results regarding predictors of linkage consent remain largely unknown. Inconsistencies may be due to differences in the record domain, the design of surveys, the way in which consent questions are worded, differences between the populations surveyed or variations in the analyses of linkage consent patterns. Jenkins et al. (2006), for example, have looked at multiple consent questions (administrative as well as employer's records) using a special follow-up wave of the British Household Panel Survey (*BHPS*).[1] They have thereby shown that predictors of linkage consent differ widely between different consent requests (or record domains). The authors find that only one variable, the household context, is associated with linkage consent in both record domains. Due to the (additional) issues arising from comparing linkage consent across record domains, this study focuses on the literature in the field of social sciences. For a review of the literature on consent decisions in the medical field, we refer to Kho et al. (2009).[2]

A number of studies have investigated the total effects of non-consent and non-response on the representativeness of final survey data. Sakshaug and Kreuter (2012) analyze the IAB *PASS*-data and show that imbalances between the survey and the population seem to be small both in absolute terms and relative to *classical* non-response and measurement error. Measurement error in survey data is an issue which can often be mitigated by imputing data such as income from official records. The authors therefore consider the process of merging survey data with administrative records to be more reliable than directly asking respondents to provide similar information themselves.

Warnke (2015) provides an overview of non-response and non-consent patterns in the IAB *WeLL*-data (see Section 4). Using social security information, he assesses the total bias due to non-response and non-consent by comparing final survey participants willing to share their data, with the general workforce at 150 establishments. The results show that over a period

[1] This is one of the studies reviewed by Sakshaug et al. (2012).

[2] Kho et al. (2009) review 17 unique studies from different countries and find inconsistent results for linkage consent patterns with regard to age, sex, race, education, income, or health status.

of ten years survey participants have moderately higher wage growth than the average of the workforce of those establishments. However, this does not bias the estimation of 'returns to education' (the wage premiums that more highly educated individuals receive). Sakshaug and Huber (2015) have independently compared non-response and linkage consent bias across different waves of the *WeLL* data. Bias is here defined as differences between survey participants and the population as a whole (also taking into account the consent decision) with respect to sex, whether participants are over 55, non-German citizenship, a low level of education, employment status and whether participants are working for a low wage. The authors find a modest non-response bias for two variables – a low wage and a low level of education. Of greater concern are increasing trends in a non-response bias over time. When comparing different forms of non-response bias, the authors find that bias due to linkage consent is small compared to classic non-response and measurement error bias. They also find that linkage consent bias decreases over time.

This is the first study to compare associations with linkage consent across two samples. We investigate whether systematic differences remain when two similar datasets, which are as broadly comparable as possible, are analyzed. We thereby try to shed new light on the reasons for the inconsistencies regarding predictors of linkage consent.

Finally, we go beyond comparing a sample of respondents who did provide linkage consent with a non-consent sample (or the population) with regard to observable characteristics such as age or level of education. Applied researchers are generally more concerned with non-ignorable non-consent. This is because regression models (or matching approaches) give representative results if a sample differs from the population only with respect to variables which are observable to the researcher.[3] To examine the role of non-ignorable non-consent, we replicate an existing study for the sample of individuals who did not provide linkage consent.

3 Research Questions

We begin by investigating predictors of linkage consent and compare our results to those in the literature. We will contrast predictors of linkage consent across two comparable surveys (see Section 4) to determine if and to what extent there are diverging results between two comparable German datasets. The interviews conducted to collect the data contained in these datasets were carried out by the same polling institute. Both datasets concern similar populations and have a comparable structure. Participants in the surveys were asked to provide consent for linkage of their survey data to the same administrative records in an almost identical manner. Furthermore,

[3]To give an example, let us assume that individual motivation is a driver for linkage consent decisions. If motivation is also related to both of the right-hand-side variables and to the outcome in an empirical model, there is a classic omitted variable bias. In this case therefore, coefficients in applied models are not consistently estimated.

to make the analyses of consent patterns as comparable as possible, we start by using the same set of controls. We therefore expect similar results between both datasets.

Research Question 1 *Which control variables predict linkage consent? Are there any differences between the two surveys?*

As shown in Section 2, the literature has found little consistency with respect to predictors of linkage consent (see also Sakshaug et al., 2012). Linkage consent patterns differ between studies for several important socio-demographic variables, such as age, level of education, income or sex. In Research Question 2, we want to investigate possible reasons for the inconsistent results found so far.

We are particularly interested in whether the choice of control variables matters. Standard, comparable socio-demographic variables are available for both datasets. However, further predictors for linkage consent, as discussed in the literature, are often not available for both datasets. Amongst other characteristics, these variables include personality traits or job attributes of the respondents. In this research question, we will therefore investigate whether including an extended set of further controls results in different conclusions being drawn with regard to standard variables. This may be one reason for the inconsistencies which have been found and discussed in the literature (see Section 2). We are particularly interested in what economists call non-cognitive skills such as risk aversion, or personality traits which differ between groups of a different age and education level. The inclusion or omission of these variables could therefore shed new light on predictors of linkage consent.

Research Question 2 *What reasons are there for the diverging results regarding predictors of linkage consent found in the existing literature? Does the choice of control variables matter?*

In a third question, we investigate whether response rates depend on the work environment (the firm a worker is employed by). Social sciences use increasingly linked employer-employee data in the context of inequality, globalization or innovation, among others (see Hamermesh, 1999). Little is known so far about whether co-workers within a firm who are interviewed at home, behave similarly when it comes to linkage consent. Analyses based on merged firm-worker surveys could be biased if there is a substantial intra-firm correlation in the response behavior. This could be the case, for example if individual survey-response is related to both firm effects and the outcome. Al Baghal et al. (2014) show that household members give similar consent decisions but, to our best knowledge, no information is available with respect to possible workplace heterogeneity.

Research Question 3 *Do workers of the same establishment respond similarly to the linkage consent question when asked in a telephone interview at home?*

The survey methodology literature *bias* primarily refers to imbalances in the distribution of observables such as age, education level or sex, between the available sample and the general population. Rubin (1976) terms this *missing-at-random* (MAR), and this can be formalized in the context of linkage consent as follows. We assume that X is the set of observed control variables or predictors which are available to all survey participants independent of the consent decision. Let $\mathbb{1}^{\text{refuse}}$ be an indicator variable which takes the value one if an individual did not consent to data linkage. Let Y be additional data from social security. Y includes the outcome of interest and it is observed only for the individuals from whom linkage consent was obtained. We therefore partition Y as $Y = \left(Y^{\text{consent}}, Y^{\text{refuse}}\right)$. Missing data is MAR if the following equation holds.[4]

$$P(\mathbb{1}^{\text{refuse}} \mid Y^{\text{consent}}, Y^{\text{refuse}}, X) = P(\mathbb{1}^{\text{refuse}} \mid Y^{\text{consent}}, X) \tag{1}$$

Equation (1) says that the probability of missingness depends only on X and Y^{consent}. Equation (1) is violated if missingness depends on information which is not available (Y^{refuse}).[5] MAR is closely linked to the concept of *ignorability* (see Gelman et al., 2014). MAR is often of little concern to researchers in the field of applied sciences. This is because X is usually controlled for in multivariate analyses. Researchers in this field are therefore more concerned with *missing-not-at-random* (MNAR) or with missingness which depends on unobservable heterogeneity which violates Equation (1). To give an example for which MAR does not hold, let us suppose that we are interested in measuring the wage effects of attending a job-related training course (see Section 6.2). Let us assume that individuals with higher earnings are less likely to provide linkage consent but more willing to participate in training. If training is associated with earnings, our data is not MAR anymore. In this case, we would estimate biased results for the wage returns to training if we restrict our analyses solely to the individuals who provided linkage consent ("non-consent bias"). Statistical techniques to deal with MNAR data are among others the Heckman selection approach (Heckman, 1979) or pattern-mixture models (Little, 1993).

We investigate whether the MAR assumption is justified in two economic applications in Section 6. To check the MAR assumption, we compare estimated regression results from two samples, individuals who provided or denied linkage consent. This is, however, not a full proof. Misspecification could give constant regression coefficients even in the presence of MNAR. Nonetheless, if we find similar results between both samples, it indicates that MNAR is at least

[4] This simple formula can of course be extended to more complicated patterns of missing data such as censoring (if X is always observed but y only up to a threshold) or truncation (if (X, y) is only observed for certain ranges of y).

[5] An even stronger assumption regarding missing data is *missing-completely-at-random* where $P(\mathbb{1}^{\text{refuse}} \mid Y^{\text{consent}}, Y^{\text{refuse}}, X)$ is assumed to be constant, e.g. independent of observable and unobservable variables.

not a severe issue for these applications.

Research Question 4 *Do we find non-consent bias in economic models? Is missing information due to linkage consent "missing-at-random"?*

Several studies carried out in different contexts have analyzed whether non-response and/or attrition in panel contexts is ignorable. Fitzgerald et al. (1998); MaCurdy et al. (1998) for example, have considered earnings regressions and van den Berg et al. (2006), unemployment durations. These studies generally find that non-response bias is rather small in magnitude. However, not all researchers agree with this view. In a recent working paper for example, Heffetz and Reeves (2016) found that official government statistics in the US, including the unemployment rate and labor force participation, depend on the ease or difficulty of contacting a respondent. To the best of our knowledge, no such evidence is available with respect to linkage consent.

4 Data

We compare two longitudinal, linked employer-employee datasets which have a similar structure. The first dataset is called *WeLL* (*Berufliche Weiterbildung als Bestandteil Lebenslangen Lernens*, which might be translated as "further training as a part of lifelong learning", see Huber and Schmucker, 2012). The second dataset is the *Linked Personnel Panel* (*LPP*, see Bellmann et al., 2015). The data contained in both of these datasets has been collected by the Research Data Centre of the Federal Employment Agency at the Institute for Employment Research, in cooperation with the polling institute infas. Table 2 shows basic information for both datasets.

WeLL is a four-wave panel which was conducted annually between 2007 and 2010. *LPP* is (currently) a two-wave panel, which was run in 2012 and 2014. Both surveys began with a firm survey (the relevant business units were establishments).[6] For this study, establishments were drawn from the IAB Establishment Panel (an annual employer survey, see Kölling, 2000, for more information) in 2005 (*WeLL*) and 2011 (*LPP*). Only establishments with at least 50 employees subject to social security contributions were eligible. For *WeLL*, the effective minimum number of employees is 100 and there is an upper limit of 2,000 employees. In a second step, outlined in detail in Sections A.3 and A.4, employees who work in these establishments were invited to participate in the employee survey.

In both datasets, survey data was merged with German social security records. These records contain employers' reports about all employees who are subject to social security contributions.

[6] We do not consider the establishment survey as data privacy laws mean that we are not entitled to link the survey information for the non-consent sample to the establishment sample. Moreover, the establishment survey has not been made public in the case of *WeLL*.

These reports are relevant for health insurance, the statutory pension scheme and unemployment benefits (Dorner et al., 2010). The reports include information about individuals' daily wages, periods of employment, citizenship, educational attainments and occupation. This information is combined with data from the Federal Employment Agency concerning social security benefits and periods of unemployment. At the end of the interview, individuals were asked whether they would agree to their data being linked to "other data [...] available at the Institute for Employment Research in Nuremberg", as outlined below. Only information on individuals who agreed to their survey data being linked with social security records was available to researchers using both survey data and administrative records (what we call the *consent sample*).[7]

In the following we will describe both employee surveys. These surveys are publicly available to the scientific community. Establishment information is available for those respondents who consented to their data being linked to the IAB Establishment Panel (and through the employer survey in the case of *LPP*). For *WeLL* we have further information relevant for the stratification of all respondents (see Section A.3). Since other establishment information is not available for the non-consent sample, we limit the following description to the employee survey.

The *WeLL* dataset focuses on job-related training and consists of data relating to 149 establishments and approximately 7,900 individual survey respondents. The *LPP* dataset focuses on human resources and management practices, job quality and corporate culture. In terms of its hierarchical structure, this dataset is very similar to *WeLL*, but it differs with regard to the ratio of workers to establishments. The number of establishments is more than six times that in *WeLL* (980 establishments), while the number of employees is approximately 50% higher than in *WeLL* (7,508 in the first wave plus 3,987 new employees in the second wave). The number of survey participants per establishment is therefore considerably lower in *LPP* than in *WeLL* (the median number of first-time survey participants per establishment in the final dataset is 25 in *WeLL* and 8 in *LPP*).

Both datasets sample full-time, part-time or marginally employed workers excluding apprentices and workers in partial retirement. We also exclude individuals for whom information on key variables such as education level, working hours or nationality is missing. We thereby restrict the sample to individuals who are still employed by the establishment (at least prior to the most recent interview). This leaves us with 11,385 first-time respondents in *LPP* and 5,753 first-time respondents in *WeLL*. We focus on first-time survey participants to investigate the initial linkage decision (with the exception of Section 6.1). We do this for two reasons. Firstly, considering multiple interviews per survey participant would mix results for linkage consent and panel non-response. Secondly, some individuals revised their consent decision – but only in one direction. Once individuals had agreed to record linkage, they were not consulted about linkage consent again in subsequent waves of the survey. It was only participants who did not

[7] In the case of *WeLL*, the merged data is called *WeLL-ADIAB* (Schmucker et al., 2014).

give initial consent who were again asked at a later interview. As a consequence, there are no individuals who withdrew their consent. There are, however, a number of survey participants who initially declined to give consent, but who later changed their minds (almost half of respondents who initially refused to provide linkage consent are included in our final WeLL sample). By restricting the (main) analyses to first-time participants we take this asymmetry resulting from the interview design into account. If we do not exclude panel participants, the results are generally similar. In Section 6.1 for example, we consider panel respondents in a replication of a study using *WeLL*.

Age is available in only four categories in *WeLL*. In contrast, *LPP* contains this information by year. In order to make the analyses between both datasets comparable we define similar age categories in *LPP*.

Table 3 shows descriptive statistics for all binary variables in *LPP* and *WeLL*.[8] Average consent rates amongst the first-time survey participants in our sample are 82% (*LPP*) and 94% (*WeLL*). The fact that the consent rates are above 80% means that they are relatively high compared to those of other datasets (linkage consent rates vary between 24% and 89% in the non-exhaustive review in Sakshaug and Kreuter, 2012). Besides higher average consent rates, we find for *WeLL* also a greater willingness to participate in future waves and lower item-non-response (for the question about net income). Compared to *WeLL*, the *LPP*-data is on average older, and there are fewer female respondents. While 80% of the participants in the *WeLL* survey state that they are in good or very good health, this is true for only 60% of the *LPP*-respondents (the age structure is probably an important reason for this difference). In *WeLL*, we find both more low-educated and highly educated workers (without a vocational qualification and with a tertiary degree respectively) but fewer workers with a vocational qualification.

Detailed information on *WeLL* and *LPP* is given in the Appendix A.3 and A.4.

5 Predictors of Linkage Consent and Establishment Heterogeneity

5.1 Predictors of Linkage Consent

In **Research Question 1** we look at possible predictors of linkage consent in *LPP* and *WeLL*. We will begin by comparing a selected set of variables which are available in both datasets. The variables used were chosen on the basis of information provided in the existing literature

[8] We standardize all non-binary variables by subtracting the mean and dividing by the standard deviation.

on non-response and linkage consent.[9] The selected variables include socio-demographic and job characteristics as well as personality traits, self-reported health and work satisfaction.[10] The results are provided in Columns 3 and 4 of Table 4. To ease interpretation, non-binary variables, such as a proxy for labor attachment and the item regarding work satisfaction, are standardized by subtracting the mean and dividing by the standard deviation.

We estimate a random effects logit model including establishment random effects. This model allows us to capture possible intra-firm correlation in the linkage consent decision. Previous studies have considered the role of interviewers (Sakshaug et al., 2012) and/or households (Al Baghal et al., 2014). There is no information available, however, indicating whether employees who work for the same establishment tend to reach a similar decision with regard to linkage consent in a telephone interview in which they participate from home.

$$\mathrm{P}\big[\,\text{Linkage Consent}_i \,\big|\, \alpha_i,\, \boldsymbol{X}_i\,\big] = \text{logit}^{-1}\big(\alpha_i + \boldsymbol{X}_i\beta\big) \tag{2}$$

We model the linkage consent decision of individual i in the respondent's first interview. $\boldsymbol{X}$ includes varying sets of predictors such as wave-specific consent rates, socio-demographic or job characteristics, as listed in Table 4. α are random effects for the establishment in which a worker is employed (we assume $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$). We will start by discussing the role of certain observable predictors ($\boldsymbol{X}$). At the end of the section we will also briefly describe the role of establishment heterogeneity.

Our findings demonstrate that linkage consent is strongly related to age. Older employees are more likely to consent to data linkage than their younger counterparts. This tendency is particularly noticeable in the *LPP*-data. Amongst the variables included in Columns 3 and 4 in Table 4, it is by far the age variable, or more specifically, being over 55, which is the most important predictor of consent linkage (according to the size of the coefficient). The size of the coefficient is comparable across both surveys. The results gained for each of the surveys, however, diverge in terms of whether respondents aged between 35 and 45 were more or less likely to give consent than younger adults.[11] See below for a discussion of possible reasons for these findings regarding age and possible cohort effects.

Next, we come to respondents' level of formal education. Existing studies show inconsistent

[9] In the Appendix, we document a lasso variable selection approach (mainly for *WeLL*) in order to test the robustness and the relevance of the variable selection.

[10] Other studies such as Jenkins et al. (2006) or Sakshaug et al. (2012) have also included income as an additional predictor. We will refrain from using income as a predictor because item non-response is high, in particular among respondents who do not provide linkage consent. Furthermore, in *WeLL* only respondents who did not give their consent to data linkage have been asked to state their gross income. For the sample of respondents who did provide consent, gross income was taken from social security records. For *LPP* we find that gross wage is not significantly related to linkage consent and it does not alter the results for the other variables.

[11] The difference in consent rates for individuals aged 35-45 is significant according to a Z-test (p=0.043).

results regarding the role of (higher) education when it comes to linkage consent (see Table 1). Comparison of the two surveys considered within this study has revealed lower consent rates amongst individuals who have completed tertiary education in the *LPP*-data. This is not the case in *WeLL* (however, according to a Z-test, p=0.18 this difference is not statistically significant). In previous research which we carried out for *WeLL* (Warnke, 2015), more highly educated survey participants exhibited significantly lower linkage consent rates than less highly educated respondents. The discrepancy with our previous results might be explained by the fact that the samples used in the respective studies were selected according to different selection criteria and a different variable selection. In this study, we have focused exclusively on employed individuals in their first interview and have also included an indicator for white-collar workers.[12]

We now turn our attention to the jobs which are typically performed by more highly educated workers. Interestingly, we find much lower consent rates in both surveys for workers employed in white-collar jobs. Such positions generally presuppose that workers hold a tertiary degree: 95% of respondents holding a tertiary degree are employed as white-collar workers. In contrast, 'merely' 57% of those who do not hold a tertiary degree are employed in such a position. This may indicate a tendency towards increased reluctance to share data amongst the more highly educated workforce. This in turn may be explained by the greater concerns which more highly educated individuals generally have with regard to data privacy (Sheehan, 2002).

Next, we look at further socio-demographic characteristic considered in this study. We see virtually no differences between the consent rates of men and women throughout the entire sample. In addition, with regard to the household context, few predictors have been found to be significant. Neither marital status, nor whether the respondent has children has been found to be a significant predictor of linkage consent behavior. For this reason, we do not include them in the main specifications presented here.[13] Respondents living alone represent an exception to this rule; these individuals are considerably less likely to consent to data linkage.

With regard to job characteristics, we identify higher consent rates for individuals performing shift work in both surveys (this tendency is significant in *LPP* only). In addition, there are significant differences in the consent decision made by workers with and without personnel responsibilities in *WeLL*. Whilst approximately 29% of workers in both surveys have managerial responsibilities, such individuals included in *LPP* do not show higher consent rates (significant with p=0.06 according to a Z-test). Having said that, this finding might be a statistical artefact arising due to multicollinearity. The lasso approach, as described in the Appendix, suggests a much lower coefficient for *WeLL*.

We now consider the so-called Big Five personality traits and the more subjective items con-

[12] If we adapt the sample selection criteria by including unemployed or self-employed individuals and use a similar set of variables as in Warnke (2015), we are able to confirm our earlier results.

[13] Further results are available upon request.

cerning job insecurity and work satisfaction. Sala et al. (2012) identified no link between the Big Five personality traits and linkage consent for the United Kingdom. Looking at the *LPP*-data, we come to a similar conclusion with regard to Germany. According to an F-test (p=0.83), all Big Five personality traits are jointly insignificant in this sample. If we consider the *WeLL* data, we do find that coefficients are in general of a larger magnitude. There is a significantly positive correlation between conscientiousness and linkage consent. The opposite is true of extraversion. The coefficients for conscientiousness differ significantly between both surveys (p=0.02, Column 3 and 4). Both surveys use the same Big Five items and we accordingly find very similar relationships between personality traits and tertiary education, for example.[14] Nonetheless, different conclusions can clearly be drawn with respect to personality traits when considering *WeLL* as opposed to *LPP*.[15]

If we add further controls which are not available in both surveys, findings generally remain quite similar. Such variables include risk and trust for *LPP* (Column 5) and establishment characteristics, voluntary work and a proxy for labor attachment for *WeLL* (column 6). Few of the predictors discussed above are altered to any meaningful extent when further controls are added. We also added indicators of individuals' willingness to participate in future waves of the survey and an indicator of whether information about net income was available or not. We see that the provision of linkage consent is closely associated with item non-response with regard to net income and individuals' readiness to take part in a future survey with similar magnitude across both datasets.[16] 29.1% of the individuals who did not consent to data linkage also did not provide information about household net income in *LPP*, but we wage information is missing for only 9.3% of the individuals in the consent sample. In terms of rates of linkage consent provision, there are no large differences between individuals who refuse to state their net income and those who do not know it. We therefore considered such individuals in a single group.[17]

After controlling for risk aversion and trust, the confusion surrounding the role of personality traits and other psychological attributes for linkage consent increases. Both variables are significantly related to linkage consent with higher rates of consent seen amongst more risk-friendly and confident individuals. In *LPP*, we found that conscientiousness had a significantly negative correlation with linkage consent. After including the further controls mentioned above, it seems

[14] Workers who hold a tertiary degree tend to have a conscientiousness score which is 2.2 (1.8) standard deviations higher than those with no vocational education in *LPP* (*WeLL*).

[15] Again, results are similar if we use only the Big Five personality traits without including any further controls as predictors.

[16] Jenkins et al. (2006) have found that item non-response is related to rates of consent linkage where the item concerns permission to contact the respondent's current employer. No relation is found however, if the item concerns permission to request the respondent's national insurance number or permission to access administrative tax and benefits records.

[17] Information regarding the reasons for item non-response is only available for *LPP*.

that individuals who score more highly with regard to openness to new experiences, are more likely to consent to data linkage. These changes can be explained by the association between personality traits and risk aversion and trust.[18] We find the coefficients of both personality traits to be significantly different to our findings in *WeLL* (where the inclusion of further variables leads to very similar conclusions as before). In addition, the association between work satisfaction and the provision of linkage consent remains elusive. While more satisfied individuals are significantly more likely to provide consent for data linkage in *LPP*, we find a negative, albeit insignificant correlation in *WeLL*.[19]

5.2 Explanations for Inconsistencies Regarding Predictors of Linkage Consent

With regard to the reasons for the diverging results regarding predictors of linkage consent (**Research Question 2**), we have so far seen that results for socio-demographic variables are quite consistent and that they remain largely unaffected by the inclusion of a different set of controls. This is also true if we add further variables such as job tasks or household information. Psychological attributes and more subjective items such as work satisfaction are relevant for linkage consent. The interpretation of these predictors seems to depend on the inclusion of similar controls and we find here different results between both surveys.

In the *WeLL*-data (last column in Table 4), we find that individuals working in East Germany tend to agree much more often to their data being linked to social security information than individuals working in West Germany. The magnitude of this relationship is large and is comparable to the age gradient in linkage consent seen in individuals aged above 55 or below 35. A similar association has been found by Antoni (2011) and Korbmacher et al. (2013). If we assume that many of the workers concerned grew up in East Germany, this finding seems at first to be somewhat counterintuitive.[20] In the former German Democratic Republic (GDR), everyday life was subject to active surveillance, which we might justifiably presume would lead to heightened privacy concerns amongst the East German population. Furthermore, it is also known that the East German population generally exhibits lower levels of trust (Rainer and Siedler, 2009), a predictor which is positively related to linkage consent, as seen for the *LPP*-data. We do not have a trust measure in *WeLL* but we might expect to find on average lower consent rates in East than in West Germany as a result of the omission of this variable. However,

[18] In *LPP*, individuals who are more open to risk are on average less conscientious and, surprisingly, less curious. More confident individuals meanwhile, exhibit greater levels of conscientiousness and are more open to new experiences.

[19] According to a Z-test, p=0.07, the difference in coefficients for work satisfaction is significant.

[20] Korbmacher et al. (2013) show that higher consent rates in East Germany are mostly driven by individuals who have lived in the former German Democratic Republic.

opposing arguments can also be made. In the 1980s for example, there was an intense debate about privacy concerns in West Germany which lead to fierce opposition to the census carried out in 1987. Such experiences may have changed attitudes towards privacy in West Germany. Future research may look at the role of, on the one hand, trust or cohort experiences and, on the other hand, linkage consent.

If attitudes have changed over time due to experiences such as growing up in the GDR, this could mean that our age gradient in fact reflects cohort effects. The short time span of the two surveys means that we are unable to distinguish between age and cohort effects (and possible additional period trends). There are, however, no differences in the age-consent pattern between workers employed in establishments in East and West Germany in *WeLL* (according to a F-Test, p=0.24, see also Table 5). This could indicate that in Germany at least, reluctance to consent to data linkage to social security records declines as age increases. Such a finding has indeed been outlined at the beginning of this section. Considering the findings of earlier social science studies, this pattern is somewhat surprising.[21]

If we now turn our attention to differences in linkage consent rates between the first and subsequent waves of the survey, we find that individuals in the first wave more often consent to data linkage. The tendency toward lower rates of consent in later waves of the survey is particularly noticeable in WeLL where the survey wave is in fact the most important predictor of linkage consent (with the exception of item non-response and willingness to participate in future waves of the survey). One possible explanation for this finding is that more respondents participate in subsequent waves of the survey who had been initially reluctant to take part. These individuals are presumably also less willing to give their consent for data linkage (see also Heffetz and Reeves, 2016). This pattern should be more pronounced in WeLL than in LPP. In WeLL, there are on average fewer workers per establishment who had not been contacted, or who had not responded in the first wave of the survey. It might also be the case that the wave-specific consent rate in WeLL has been overestimated due to multicollinearity and small sample sizes. In the Appendix, we provide a regularized lasso approach which is better suited to multicollinearity (Tibshirani, 1996). This method suggests much weaker differences between the first and subsequent waves (but similar magnitudes for other predictors such as age and item non-response).

Our analyses indicate that both psychological attributes and job characteristics play an important role in an individual's decision whether or not to grant permission for data linkage. Psychological attributes include personality traits, the degree of risk aversion, trust and subjective measures such as work satisfaction or expectations regarding future labor market activity (or labor attachment). Job characteristics included in this study include white-collar occupations, managerial responsibilities, shift work or certain job tasks. Our study suggests that the role of

[21] In the medical literature, however, some studies have found higher consent rates amongst older individuals, whilst other studies have found no differences with respect to age, see Kho et al. (2009).

environmental indicators, such as our indicator for East or West Germany, should be the object of greater research. This is a potential source of the inconsistencies seen in the literature. In the *WeLL* for example, we find that female respondents employed in East Germany are significantly more likely than their male colleagues to grant consent for data linkage. No such difference relating to gender is identified in Western Germany.[22] Similar results are found for shift work which strongly correlates with the provision of linkage consent in establishments in East Germany only. These results reflect what we have found in the *LPP* data.

5.3 Information on Establishment Heterogeneity

We next consider possible establishment heterogeneity (**Research Question 3**). We thereby wish to investigate whether respondents employed by a particular firm reach a similar decision regarding data linkage consent when asked in a telephone interview. If the work environment plays a role for the individual linkage consent decision in a telephone interview at home, this could matter for the analyses of linked employer-employee data. Suppose, for example, that workers who fear they might lose their jobs do not provide linkage consent in private firms but are more inclined to give linkage consent when working in the public sector (or in other firms with stronger job protection). This could bias studies analyzing public sector motivation, for example. Furthermore, it could possibly also help to explain the inconsistencies regarding predictors of linkage consent found in the literature.

Establishment heterogeneity can be assessed by the variance of the establishment random effects terms (Table 4, Row σ^2_α). Aggregated consent rates between establishments vary much more than we would expect given pure random variation (taking into account the fact that sample sizes differ) as shown in columns 1 / 2. The variance of the establishment random effects is comparable between WeLL$\big(100{\cdot}\text{Var}(\sigma^2_{\text{Firm}}) = 2.74\big)$ and *LPP* $\big(100{\cdot}\text{Var}(\sigma^2_{\text{Firm}}) = 4.32\big)$. According to a (conservative) likelihood-ratio test this heterogeneity is significant for both datasets.[23].

Establishment heterogeneity can also be evaluated via a simple Monte-Carlo simulation (see the Appendix A.2). Here, we model the distribution of (aggregated) mean consent rates by establishment. This distribution is clearly non-normal for two reasons. First, the number of respondents per establishment varies widely (in particular in the *LPP* data). Second, average consent rates are by construction limited to $[0, 1]$. In the Monte-Carlo simulation, we assume that in the absence of intra-establishment correlations with respect to linkage consent, average consent rates represent series of Bernoulli trials. In Figure 1, we compare the empirically

[22]Results are based on a regression including the same set of controls as those listed in Columns 3 and 4 in Table 4 separately for establishments located in East and West Germany. Results are available upon request.

[23]95-th confidence intervals are $[0.002, 0.40]$ in WeLL and [0.12,0.16] in LPP.

observed distribution to 10,000 simulated draws.

If we include individual-level controls (Column 3/4 in Table 4), we see that the variance of the establishment random effects is reduced by half in *LPP* and is close to zero in *WeLL*. This indicates that there is some amount of establishment heterogeneity which can be primarily explained by differences in observable characteristics between the respondents employed by different firms.

For *WeLL* we find that an indicator for East Germany can alone explain most of the establishment heterogeneity.[24] We therefore run two separate logit regressions for East and West Germany (without random effects). This allows us to investigate whether predictors of linkage consent are consistent between East and West Germany. The results are shown in Table 5. Interestingly, we often find quite different results for workers employed in East German establishments compared to respondents who work in West Germany. We find significant discrepancies between female respondents working in establishments located in East and in West Germany (p-value 0.08).[25] It is only female workers in establishments in East Germany who exhibit higher consent rates than their male colleagues. We also find that the positive association between shift work and linkage consent is driven entirely by the consent behaviors of respondents employed in establishments located in East Germany.[26] Furthermore, there are other notable differences, for example, regarding part-time (which is negatively associated with linkage consent only in East Germany) or for less-educated workers (who tend to provide more consent in East Germany but somewhat less in West Germany). These differences are, however, not significant according to conventional levels (p-value 0.17 for part-time and p-value 0.18 for low educated).

6 Bias in Economic Models

In this final section we wish to shed light on possible unobservable heterogeneity between survey participants who consent to data linkage and those who do not (**Research Question 4**). In order to assess whether non-consent can be viewed as *missing-at-random* (MAR), we look at two economic models. The first application concerns participation in job-related training while the second looks at the wage effects of human capital investments. We test the MAR assumption by comparing regression results for the sample of respondents who gave linkage consent with those who did not provide linkage consent. If, on the one hand, both groups differ

[24]Due to data anonymization, we do not have information relating to region in *LPP* and therefore cannot say whether this is also the case in this sample.

[25]The p-value is calculated by running a joint regression in which we interacted all variables with an indicator for working in an East German establishment.

[26]The interaction term for shift work and East Germany is significant (p-value 0.06).

in unobserved variables which are correlated with both other predictors and the outcome of interest, this will generally give inconsistent regression coefficients. This would be a strong indication that missingness due to linkage consent is *not-missing-at-random* (MNAR). On the other hand, if regression coefficients are stable across groups, it is reassuring that MNAR plays only a limited role for these applications. It should still be noted that misspecification or other issues could give stable coefficients even if linkage consent is MNAR.[27]

6.1 Replication of Steffes and Warnke (2016)

The first analysis is a replication of our own previous work in Steffes and Warnke (2016). In that paper, we analyzed workers' participation in training using a matched employer-employee dataset. We explained to what extent training rates differ between workers within the same firm and between workers employed in different firms. For this purpose, we made use of the *WeLL-ADIAB*-data (Schmucker et al., 2014). *WeLL-ADIAB* links the *WeLL* survey data to social security data.[28] This dataset includes for example, full employment biographies (subject to social insurance contributions). It includes information about wages, periods of unemployment and levels of education. We used the social security data to measure, in particular, firms' rates of wage compression. This is an important variable discussed in the theoretical training literature.

In *WeLL-ADIAB*, survey data is available online only for those respondents who consented to data linkage. For the following replication, we use the *WeLL* survey data and re-run our original analyses based on *WeLL-ADIAB* for this sample. Further details are available in the Appendix A.6. As in our earlier study, we do not restrict the analysis in this study to first-time interview respondents, but instead also include panel participants. In order to analyze variation in training rates between workers and firms in Steffes and Warnke (2016), we ran a two-way random effects logit model estimated via maximum-likelihood in Equation (3). We thereby used the panel dimension of the *WeLL*-data to separate firm and worker heterogeneity. The original estimation equation reads as follows:

$$\Pr\big[\text{Training=1}_{it} \,\big|\, \alpha_{j(i)},\, \theta_i,\, \boldsymbol{T}_t,\, \boldsymbol{X}_{it}\,\big] = \text{logit}^{-1}\big(\boldsymbol{T}_t\tau + \alpha_{j(i)} + \theta_i + \boldsymbol{X}_{it}\beta\big) \tag{3}$$

Here, T_t capture time effects, θ_i are random-effects for worker i, $\alpha_{j(i)}$ are random effects for the establishment j where worker i is employed at time t.[29] This allowed us to analyze the (relative) importance of firms and workers in determining the individual's participation in training. By gradually adding worker, firm and job characteristics as predictors (X_{it}), we then explained

[27] We also carefully look at the variance explained for different groups, as shown to be important to check coefficient stability (Oster, forthcoming).

[28] Social security records are available for all employees of the establishments participating in *WeLL*.

[29] $\alpha_{j(i)} \sim \mathcal{N}(0, \sigma^2_\alpha)$ and $\theta_i \sim \mathcal{N}(0, \sigma^2_\theta)$ are assumed to be mutually independent, and independent of T_t and X_{it}.

these differences in participation in training.

For the purpose of replication, we pool survey information respondents for whom consent for data linkage was obtained with the sample of respondents who did not provide consent. We control for age (four categories), sex, level of education and citizenship of the given worker, as well as for the respondent's relationship status and whether he or she has recently experienced a period of unemployment. We control for the subjective health status of the respondent and for the probability that he or she will be active in the labor force in one year's time, as assessed by the respondents themselves (a proxy for labor attachment). We estimate here the following equation:

$$\begin{aligned} &\Pr\Big[\text{Training=1}_{it} \,\Big|\, \alpha^1_{j(i)},\, \alpha^2_{j(i)},\, \theta^1_i,\, \theta^2_i,\, \gamma,\, \mathbb{1}\big(\text{Linkage Refusal}_i\big),\, \boldsymbol{T}_t,\, \boldsymbol{X}_{it},\Big] = \qquad (4) \\ &\quad = \text{logit}^{-1}\Big(\boldsymbol{T}_t\tau + \alpha^1_{j(i)} + \theta^1_i + \boldsymbol{X}_{it}\beta + \mathbb{1}\big(\text{Linkage Refusal}_i\big)\,\big(\alpha^2_{j(i)} + \theta^2_i + \gamma + \boldsymbol{X}_{it}\beta^R\big)\Big) \end{aligned}$$

Equation (4) extents Equation (3) by adding the indicator "$\mathbb{1}\big(\text{Linkage Refusal}_i\big)$" which takes the value of one if an individual has not provided consent. γ_i is a (fixed) constant which represents the relative intercept of the group of respondents who do not give consent compared to the consent sample. θ^2_i and $\alpha^2_{j(i)}$ are random coefficients for the non-consent sample. We are interested in γ, θ^2_i, $\alpha^2_{j(i)}$ and the interaction terms of linkage refusal with X_{it}. These parameters show us differences in participation in training between the sample of individuals who provided linkage consent and those who did not.[30] For computational reasons, we assume that the covariances between the random effects and random coefficients are all zero.[31]

The results are presented in Table 6. We start with the findings for the respondents who provided linkage consent. As in Steffes and Warnke (2016), we find a strong association between training on the one hand, and age or education on the other hand.[32] Individuals with higher labor attachment and those with better health participate more in training and the opposite is true for workers with a migration background or who have experienced unemployment. We have shown that many of these associations disappear after controlling for job tasks performed at work.

In the following, we look at training patterns for the non-consent sample. Column 1 in Table 6 shows that individuals who did not provide linkage consent participate on average less in training. The (unconditional) average training rate is 46.2% among respondents who provided linkage consent and 45.3% for the sample of individuals who did not. The training gap of 0.9pp is rather small and becomes insignificant after including further variables (Column 2).

[30] Results are similar if we additionally interact T_t and $\mathbb{1}\big(\text{Linkage Refusal}_i\big)$.

[31] We further assume that random effects and random coefficients are independent of X_{it}.

[32] In Steffes and Warnke (2016), we used age and age squared. This showed a large negative but insignificant squared age term. Age is only available in four categories in the survey data.

For the other variables, we see that results are in general close to the previous results with two exceptions. Age above 55 and health status show a significant negative interaction term.[33] Older and less healthy workers participate less in training, in particular among the respondents who did not provide linkage consent. This indicates that, by focusing on the consent sample, we might have underestimated the already negative association between poor health and age on the one hand, and participation in training on the other hand. The association with predictors such as sex or level of education remains unchanged between the previous and current study.

Next, we look at the random effects and random coefficients. We test whether the inclusion of the random coefficients for the individuals who did not give linkage consent significantly improves our model using a likelihood-ratio test. The likelihood ratio statistic is marginally significant for the model including time effects only (Column 1, p-value 0.12) and significant at conventional levels for the model such as worker characteristics (Column 2, p-value 0.04). How large are these differences? To ease interpretation, we have presented variance components from separate regressions in Table 7. This table shows in (1) the original results from Steffes and Warnke (2016), in (2) results for the respondents who gave consent and (3) for the non-consent sample.[34] We see that the variance components are indeed slightly lower among respondents who did not give consent. Two points should be noted. First, linkage refusal is highly correlated with panel attrition (Table 4). Second, almost one-third (29.3%) of individuals who initially declined to provide linkage consent later reconsidered their decision and subsequently provided consent (and appear in the consent sample).[35] This implies that fewer respondents with linkage consent are observed at multiple periods, almost two-thirds of the individuals in the non-consent sample are observed for only one period compared to less than one-third of the consent sample. As a consquence, the estimation of random effects becomes less precise (similar to attenuation bias in the presence of classical measurement error).

One of the contributions of Steffes and Warnke (2016) to the training literature is a detailed variance decomposition based on the random-effects. Thereby, we show that firm heterogeneity plays only a minor role for workers' participation in training after taking into account differences in firm, worker and job characteristics. This result, among others, seems to be unaffected by the omission of respondents who never gave linkage consent. Even if we partition the variance components based on the non-consent sample only, this does not affect our interpretation of training differences between and within firms. This result is reassuring and indicates that unobserved heterogeneity associated with linkage consent is not (very) relevant for participation in job-related training. There is little evidence for missingness-not-at-random (see Research Question 4) in this context.

[33] The health status is standardized with lower values meaning better health.

[34] In Appendix A.6 we discuss why the analysis based on the survey data only does yield slightly different results compared to the original study.

[35] Results are similar if we restrict the analyses to individuals who never gave consent.

6.2 Earnings Regression

Various (augmented) forms of the Mincer earnings function have been estimated in microeconomics with the aim of estimating returns to schooling (Mincer, 1958). This model relates the logarithm of earnings on measures of the educational level, work experience and experience squared and other variables which often serve as controls for observable heterogeneity between educational groups. In the following we will estimate a standard Mincer earnings function using the *LPP* for which gross earnings are available for both the linkage consent sample and for hold-outs.[36]

As in Section 5, we will again focus on first-time participants only. We thereby avoid confusing inconsistencies due to linkage consent with possible panel attrition bias. We estimate a Mincer earnings function in which we interact all variables with an indicator for data linkage consent. This amounts to separate estimation on the two subsamples defined by the linkage consent indicator. These predictors include age and age squared (so-called potential experience) as well as indicators for individuals without a vocational qualification and with a tertiary degree, individuals without German citizenship, those with subjectively estimated good health or those withholding linkage consent. We add a further variable measuring participation in training in the last ca. 12 months prior to the interview, which is a common measure of returns on training (e.g. Bassanini et al., 2005). We consider this measure to be of particular interest because, as detailed in Section 6.1, respondents who never consent to data linkage tend to participate less in training. We are therefore particularly interested to find out whether this might have an effect on the association between training and wages.

$$y_i = \beta_0 + X_i\,\beta + \mathbb{1}\big(\text{Linkage Refusal}_i\big)\,\beta_0^R + X_i\,\mathbb{1}\big(\text{Linkage Refusal}_i\big)\,\beta^R + \epsilon_i \qquad (5)$$

y is the logarithm of hourly gross wages and X includes the list of variables described above. As in the previous section, $\mathbb{1}\big(\text{Linkage Refusal}_i\big)$ is an indicator function for linkage consent refusal. We cluster standard errors on the individual level.

Table 8 shows the results for the wage regression. The results are very much in line with those seen in the literature and indicate that education is rewarded in the labor market, that wages increase with age (with a negative squared term) and that women tend to earn less than men on an hourly basis. There is a slight wages penalty for individuals holding non-German citizenship, whilst individuals in good health earn more than their counterparts who report having health issues. Individuals who participate in training earn considerably more than individuals who do not (approximately two-thirds of the gender wage gap). This should not be interpreted as a

[36] In *WeLL*, only respondents who did not give their consent to data linkage were asked to state their gross income. For the sample of respondents who did provide consent, gross income was taken from social security records.

causal link. Individuals who participate in training could earn more even if they did not attend training due to favorable unobserved attributes.[37]

We next compare the results for respondents who gave linkage consent to the results for respondents who did not. In Table 8 this is expressed by the interaction effect. The results are generally reassuring for the MAR assumption (Equation (1)). Only one of the ten interaction terms is statistically significant. The majority of terms are small in magnitude. We find that individuals who refused consent earn somewhat lower wages, but the difference is not significant. Similarly, the general decrease in wages amongst less highly educated individuals is less marked amongst the non-consent sample than it is amongst respondents who did consent. The only significant differences concern participation in training, where we find that the wage difference between those who do participate in training and those who do not is higher among individuals who refuse to give linkage consent. The difference is 16.2% in the sample of respondents who did give consent, but 20.9% in the sample of respondents who did not.[38]

7 Conclusions

Survey data is increasingly being merged with administrative records. Due to survey data privacy laws, polling institutes must obtain explicit consent from individuals in order to link such data. However, not all individuals agree to their survey data being linked to such administrative records however, thereby giving rise to a new form of non-response. A growing body of literature has investigated predictors of linkage consent in order to ascertain whether surveys which are linked with administrative records can nonetheless be considered representative of the relevant general population. Previous studies in this field have thus far provided inconsistent results, in regard even to standard socio-demographic characteristics such as respondents' age or sex.

In this study, we have looked at two comparable German surveys, the data from which has been linked to social security data. Using these datasets, we have provided new insights about the characteristics of those individuals who tend to decide against allowing their data to be linked to social security data. Furthermore, we discuss the implications of these findings for survey practitioners and researchers.

We first shed new light on the relevance of possible reasons for the inconsistencies found in the existing literature. We have compared linkage consent patterns in a multivariate regression in both datasets using the same set of control variables. We have thereby shown that common predictors such as age, sex or non-German citizenship have comparable associations with link-

[37] In the empirical literature, researchers sometimes compare participants in a training course to a control group who planned to participate but cancelled due to more or less random events such as a cancellation by the provider.

[38] $\exp^{0.15} \approx 1.162$

age consent across the two datasets. The existing literature has suggested that linkage consent is closely related to panel attrition and item non-response, a pattern which we can confirm for both datasets. In addition, we have illustrated that an individual's decision to consent (or not consent) to data linkage is associated with other psychological items such as trust or risk aversion.

There are, however, some diverging results. This concerns formal education and, in particular, the Big-Five personality traits and levels of work satisfaction. Conscientiousness for example shows either no correlation or a negative correlation with linkage consent in *LPP*, while there is a statistically significant positive correlation between the provision of linkage consent and conscientiousness in *WeLL*. In contrast, work satisfaction positively correlates with the provision of linkage consent in *LPP*, whilst it shows a negative correlation with consent in *WeLL*. We have also found that including further predictors does not help to explain diverging findings regarding respondents' level of education, personality traits or an individual's level of work satisfaction. Taking further variables, which are not necessarily available in both datasets and which capture other psychological attributes or firm characteristics, into account does not alter our findings. The vast majority of correlations found remain unchanged when these further variables are included.

We have shown that the work environment plays only a minor role for the individual decision to provide linkage consent. Yet, there are large (average) differences in consent rates between respondents working in East and West Germany. Consent rates are higher amongst those employed in firms located in East Germany. This association cannot be explained by cohort differences and indicates that the role of shared experience may be important. We then have compared the linkage consent patterns identified for respondents in establishments in East and West Germany. Whilst female respondents employed in firms in East Germany are significantly more likely to consent to data linkage than their male colleagues, such a difference is not seen between male and female respondents employed in West German firms. We do find, however, that other results are generally comparable between East and West Germany. These findings indicate that differences between the populations surveyed may have contributed to inconsistencies with respect to predictors of linkage consent in the literature.

Our study is the first to analyze and compare the impact of linkage consent in two empirical models. Firstly, we have replicated one of our own previous studies in which we made use of the *WeLL* survey data linked with social security data. Accordingly, it was not possible to consider individuals who did not give linkage consent in this analysis. In respect to job-related training at least, there are few differences between the sample of individuals who consent to their survey data being linked to social security data, and those who do not give such consent. For this reason, we were able to confirm our previous findings on the sample of respondents who did not give consent to data linkage. Secondly, we have considered the results from a well-known empirical model which measures the wage returns to human capital investments

(schooling as well as job-related training). We find that wage differences between individuals who do participate in training and those who do not participate in training is larger among the sample of respondents who fail to provide linkage consent than for respondents from whom linkage consent was obtained. All other results differ very little by consent.

We therefore conclude with a promising view about linkage non-consent. The role of unobserved heterogeneity between respondents who gave linkage consent and respondents who did not seems to be rather small in the applications we have analyzed. Future research should address the role of psychological attributes in determining an individual's decision for or against linkage consent in more detail.

References

Al Baghal, T., Knies, G., Burton, J., et al., 2014. Linking Administrative Records to Surveys: Differences in the Correlates to Consent Decisions. Tech. rep., Understanding Society at the Institute for Social and Economic Research.

Antoni, M., 2011. Linking Survey Data With Administrative Employment Data: The Case of the German ALWA Survey. FDZ Methodenreport 12.

Bassanini, A., Booth, A. L., Brunello, G., De Paola, M., Leuven, E., 2005. Workplace Training in Europe. IZA Discussion Papers (1640).

Bates, N., 2005. Development and Testing of Informed Consent Questions to Link Survey Data With Administrative Records. In: Association, A. S. (Ed.), Proceedings of the Survey Research Methods Section. pp. 3786–3793.

Bellmann, L., Bender, S., Bossler, M., Broszeit, S., Dickmann, C., Gensicke, M., Gilberg, R., Grunau, P., Kampkötter, P., Laske, K., et al., 2015. LPP-Linked Personnel Panel* Quality of Work and Economic Success: Longitudinal Study in German Establishments (Data Collection on the First Wave), FDZ-Methodenreport, 05/2015 (en).

Bender, S., Fertig, M., Görlitz, K., Huber, M., Hummelsheim, S., Knerr, P., Schmucker, A., Schröder, H., et al., 2008. WeLL-Berufliche Weiterbildung als Bestandteil Lebenslangen Lernens. No. 5.

Dorner, M., Heining, J., Jacobebbinghaus, P., Seth, S., 2010. The Sample of Integrated Labour Market Biographies. Schmollers Jahrbuch 130 (4), 599–608.

Fitzgerald, J., Gottschalk, P., Moffitt, R., 1998. An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics. The Journal of Human Resources 33 (2), 251–299.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Vol. 1. Springer Series in Statistics Springer, Berlin.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2014. Bayesian Data Analysis. Vol. 2. Chapman & Hall/CRC Boca Raton.

Gensicke, M., Tschersich, N., 2015. Vertiefende Betriebsbefragung "Arbeitsqualität und wirtschaftlicher Erfolg" 2012. Tech. rep., Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Hamermesh, D. S., 1999. LEEping Into the Future of Labor Economics: The Research Potential of Linking Employer and Employee Data. Labour Economics 6 (1), 25–41.

Heckman, J. J., 1979. Sample Selection Bias as a Specification Error. Econometrica: Journal of the Econometric Society, 153–161.

Heckman, J. J., Lochner, L. J., Todd, P. E., 2003. Fifty Years of Mincer Earnings Regressions. Tech. rep., National Bureau of Economic Research.

Heffetz, O., Reeves, D. B., 2016. Difficulty to Reach Respondents and Nonresponse Bias: Evidence from Large Government Surveys. Working Paper.

Huber, M., Schmucker, A., 2012. Panel WeLL: Arbeitnehmerbefragung für das Projekt Berufliche Weiterbildung als Bestandteil Lebenslangen Lernens Dokumentation für die Originaldaten Wellen 1-4. FDZ Datenreport. Documentation on Labour Market Data (03/2012 DE).

Jenkins, S. P., Cappellari, L., Lynn, P., Jäckle, A., Sala, E., 2006. Patterns of Consent: Evidence from a General Household Survey. Journal of the Royal Statistical Society: Series A (Statistics in Society) 169 (4), 701–722.

Kho, M. E., Duffett, M., Willison, D. J., Cook, D. J., Brouwers, M. C., 2009. Written Informed Consent and Selection Bias in Observational Studies Using Medical Records: Systematic Review. BMJ 338, b866.

Knerr, P., Schröder, H., Aust, F., Gilberg, R., et al., 2009. Berufliche Weiterbildung als Bestandteil Lebenslangen Lernens (well) WeLL-Erhebung 2007-Methodenbericht. FDZ Methodenreport (6.

Kölling, A., 2000. The IAB-Establishment Panel. Schmollers Jahrbuch 120 (2), 291–300.

Korbmacher, J. M., Schroeder, M., et al., 2013. Consent When Linking Survey Data With Administrative Records: The Role of the Interviewer. In: Survey Research Methods. Vol. 7. Citeseer, pp. 115–131.

Künn, S., 2015. The Challenges of Linking Survey and Administrative Data. IZA World of Labor (214).

Little, R. J., 1993. Pattern-Mixture Models for Multivariate Incomplete Data. Journal of the American Statistical Association 88 (421), 125–134.

MaCurdy, T., Mroz, T., Gritz, R. M., 1998. An Evaluation of the National Longitudinal Survey on Youth. The Journal of Human Resources 33 (2), 345–436.

Mincer, J., 1958. Investment in Human Capital and Personal Income Distribution. The Journal of Political Economy, 281–302.

Oster, E., forthcoming. Unobservable Selection and Coefficient Stability: Theory and Evidence. Journal of Business Economics and Statistics.

Rainer, H., Siedler, T., 2009. Does Democracy Foster Trust? Journal of Comparative Economics 37 (2), 251–269.

Rubin, D. B., 1976. Inference and Missing Data. Biometrika 63 (3), 581–592.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., Weir, D. R., 2012. Linking Survey and Administrative Records Mechanisms of Consent. Sociological Methods & Research 41 (4), 535–569.

Sakshaug, J. W., Huber, M., 2015. An Evaluation of Panel Nonresponse and Linkage Consent Bias in a Survey of Employees in Germany. Journal of Survey Statistics and Methodology.

Sakshaug, J. W., Kreuter, F., 2012. Assessing the Magnitude of Non-consent Biases in Linked Survey and Administrative Data. Survey Research Methods 6 (2), 113–122.

Sala, E., Burton, J., Knies, G., 2012. Correlates of Obtaining Informed Consent to Data Linkage Respondent, Interview, and Interviewer Characteristics. Sociological Methods & Research 41 (3), 414–439.

Schmucker, A., Seth, S., Eberle, J., 2014. WeLL-Befragungsdaten verknüpft mit administrativen Daten des IAB:(WELL-ADIAB) 1975-2011. Tech. rep., Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Schütz, H., Gilberg, R., Dickmann, C., Schröder, H., et al., 2014. IAB-Beschäftigtenbefragung: Projekt "Arbeitsqualität und wirtschaftlicher Erfolg: Panelstudie zu Entwicklungsverläufen in deutschen Betrieben-Personenbefragung". Tech. rep., Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Sheehan, K. B., 2002. Toward a Typology of Internet Users and Online Privacy Concerns. The Information Society 18 (1), 21–32.

Solon, G., Haider, S. J., Wooldridge, J. M., 2015. What Are We Weighting for? Journal of Human resources 50 (2), 301–316.

Steffes, S., Warnke, A. J., 2016. New Evidence on the Determinants of Firm-based Training. Discussion Paper, Mannheim, Germany.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288.

van den Berg, G. J., Lindeboom, M., Dolton, P. J., 2006. Survey Non-response and the Duration of Unemployment. Journal of the Royal Statistical Society: Series A (Statistics in Society) 169 (3), 585–604.

Warnke, A. J., 2015. Verzerrung durch selektive Stichproben. In: Nonresponse Bias. Springer, pp. 305–323.

Yuan, M., Lin, Y., 2006. Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1), 49–67.

A Appendix

A.1 Tables

Table 1: Overview of Consent Patterns for Selected (Recent) Studies from Social Sciences

	Linkage Consent							
	(1)	**(2)**		**(3)**	**(4)**	**(5)**		**(6)**
	Benefits	Health	Benefits	Benefits		Health	Benefits	Benefits
Consent Rate	91.6%	41%	39%	67.8%	77.6%	66.9%	77.9%	93.9%
Age (higher)	+	0	0	0		-	-	+
Foreign-born	0	-	-	0		0	0	0
Female	0	0	-	0	0	-	0	0
Highly Educated	0	0	+	+	0	-	0	-
Partnership etc.	0	0	0	0	+	0	0	0
Children	0	0	0			0	-	0
Health Problems		0	+	0		0	0	+
Employed	+			0	0	0	0	+
Country	DE	UK		US	DE	UK		DE
Interview	1st-Interv.	F.-Up		F.-Up	F.-Up	F.-Up		F.-Up
Method		Bivar. Probit		R.-E. Logit		R.-E. Logit		Logit
Sample				aged 50+	aged 50+			Employed
	Controls							
Interviewer	Yes	Yes		Yes	Yes	Yes		No
Item Non-Response	Yes	No		No	Yes	No		No

Note: See also Antoni (2011) for an excellent related overview. Here, "+" refers to statistically significant positive associations, "0" insignificant and "-" to significant negative associations.
We do not illustrate age effects for Korbmacher et al. (2013) because the SHARE data covers only individuals aged 50 and older.
(**1**): Antoni (2011), Table 5 (Columns 1), ALWA dataset
(**2**): Sala et al. (2012), Table 2 (Columns 5 & 6), BHPS dataset
(**3**): Sakshaug et al. (2012), Table 4, HRS dataset
(**4**): Korbmacher et al. (2013), Table 2 (Column 5), SHARE dataset
(**5**): Al Baghal et al. (2014), Table 4 (Columns 1 & 2), Understanding Society dataset
(**6**): Warnke (2015), Table 4, WeLL dataset

Table 2: Two IAB Matched-Employer-Employee Datasets

	LPP Matched-Employer-Employee Panel	**WeLL** Further Training as a Part of Lifelong Learning	
Focus	HRM	Further Training	
Waves	2	4	
Establishments	869	149	(wave 1)
Individuals	7,508	6,404	
Method	Phone (CATI)	Phone (CATI)	contacted at home
Polling institute	infas	infas	
Eligible	employees subject to social security or minor employment (excl. apprentices)		
Response Rate	24.5%	31.7%	(of gross sample)
Response Rate	34.1%	38.7%	(contacted individuals)
Avg. Consent Rate	81.9%	94.2%	(first-time participants)

Note: The average linkage consent rate refers to the sample of first-time survey participations without missing information regarding key variables such as education, working hours or nationality (see Chapter 4).

Table 3: Descriptive Statistics (Binary Variables)

	LPP		WeLL	
	Absolute	**Percent**	**Absolute**	**Percent**
Total (First-Interview)	11,385	100%	5,753	100%
Linkage Consent	9,325	81.9%	5,321	92.5%
Wave 1	7,417	65.2%	4,483	77.9%
Wave 2	3,968	34.9%	536	9.3%
Wave 3	-	-	734	12.8%
Female	3,240	28.5%	2,236	38.9%
No Vocat. Qualif.	262	2.3%	583	10.1%
Vocational Qualif.	9,017	79.2%	3,786	65.8%
Tertiary Degree	2,106	18.5%	1,384	24.1%
Age below 35	2,257	19.8%	1,331	23.1%
Age ca. 35-45	2,361	20.7%	1,821	31.7%
Age ca. 45-55	4,325	37.8%	1,932	33.6%
Age over 55	2,442	21.5%	669	11.6%
Foreign-born	1,077	9.5%	343	6.0%
Part-time	1,424	12.5%	904	15.7%
White Collar	6,962	61.2%	3,833	66.6%
Child Under 14	2,983	26.2%	1,803	31.3%
Living Alone	758	6.7%	954	16.6%
Good Health	6,871	60.4%	4,593	79.8%
Managerial Resp.	3,396	29.8%	1,682	29.2%
Limited Contract	678	6.0%	766	13.3%
Shift Work	3,636	31.9%	2,452	42.6%
Panel	10,677	93.8%	5,666	98.5%
Net Wage Missing	1,463	12.9%	180	3.1%
Voluntary Work	-	-	1,591	27.7%
Firm Size 100-200	-	-	861	15.0%
Firm Size 200-500	-	-	1,397	24.3%
Firm Size 500-2000	-	-	3,495	60.8%
East Germany	-	-	2,267	39.4%
Service Sector	-	-	2,925	50.8%
Training Firm	-	-	4,990	86.7%
Investment Firm	-	-	2,604	45.3%

Note: First-interview sample. Excluded are respondents for whom information on key variables such as education level is missing. Non-binary variables are standardized and not presented here.

Table 4: Random-Effects Logit Regression Estimates (on Linkage Consent)

	Outcome: Linkage Consent					
	LPP	WeLL	LPP	WeLL	LPP	WeLL
Variable	**Coef (SE)**	**Coef (SE)**	**Coef (SE)**	**Coef (SE)**	**Coef (SE)**	**Coef (SE)**
Wave 2	-0.15*** (0.05)	-1.00*** (0.14)	-0.15*** (0.05)	-0.85*** (0.15)	-0.17*** (0.06)	-0.61*** (0.16)
Wave 3		-1.14*** (0.11)		-0.95*** (0.13)		-0.76*** (0.14)
Female			0.00 (0.06)	0.15 (0.14)	0.02 (0.07)	0.06 (0.14)
No Vocat. Qualif.			-0.03 (0.18)	0.00 (0.17)	0.01 (0.2)	-0.05 (0.18)
Tertiary Degree			-0.19*** (0.06)	-0.01 (0.13)	-0.25*** (0.07)	-0.02 (0.13)
Aged ca. 35-45			0.21*** (0.07)	-0.07 (0.12)	0.28*** (0.08)	-0.06 (0.13)
Aged ca. 45-55			0.28*** (0.07)	0.16 (0.18)	0.35*** (0.08)	0.15 (0.19)
Aged over 55			0.51*** (0.09)	0.64*** (0.23)	0.68*** (0.09)	0.57** (0.24)
Foreign-born			-0.20** (0.09)	-0.15 (0.21)	0.03 (0.10)	0.10 (0.25)
Part-time			-0.09 (0.08)	-0.19 (0.14)	-0.16* (0.09)	-0.12 (0.15)
White-Collar			-0.28*** (0.06)	-0.30** (0.14)	-0.27*** (0.07)	-0.30* (0.15)
Child Under 14			-0.09 (0.06)	0.01 (0.12)	-0.09 (0.07)	0.02 (0.13)
Living Alone			-0.24*** (0.09)	-0.13 (0.17)	-0.30*** (0.10)	-0.12 (0.17)
Good Health			0.10** (0.05)	0.08 (0.13)	0.09 (0.06)	0.01 (0.13)
Managerial Resp.			0.02 (0.06)	0.27** (0.12)	-0.02 (0.06)	0.24** (0.12)
Limited Contract			0.09 (0.11)	-0.02 (0.13)	0.17 (0.11)	0.00 (0.15)
Shift Work			0.20*** (0.06)	0.20 (0.12)	0.20*** (0.06)	0.08 (0.13)
Conscientiousness			-0.02 (0.04)	0.16** (0.07)	-0.09** (0.04)	0.11* (0.06)
Extraversion			-0.04 (0.04)	-0.1* (0.06)	-0.01 (0.04)	-0.10 (0.06)
Neuroticism			0.02 (0.04)	0.11 (0.07)	0.00 (0.04)	0.09 (0.07)
Agreeableness			-0.01 (0.04)	-0.11 (0.09)	-0.01 (0.04)	-0.13 (0.09)
Openness to new Exp.			0.02 (0.04)	-0.07 (0.08)	0.10** (0.05)	-0.08 (0.08)
Job Insecurity			0.03 (0.03)	-0.05 (0.06)	0.04 (0.03)	-0.03 (0.06)
Work Satisfaction			0.04 (0.02)	-0.05 (0.06)	0.05* (0.03)	-0.07 (0.06)
Panel					2.46*** (0.09)	2.02*** (0.21)
Net Income Missing					-1.26*** (0.07)	-1.48*** (0.23)
Openness to Risk					0.11*** (0.03)	
Trust					0.08*** (0.03)	
Justice					-0.01 (0.03)	
Voluntary Work						0.20 (0.12)
Labour Attachment						0.15** (0.06)
Firm Size 100-200						0.17 (0.19)
Firm Size 500-2000						0.15 (0.13)
East Germany						0.45*** (0.12)
Service Sector						0.10 (0.11)
Training Firm						-0.08 (0.15)
Investment Firm						-0.12 (0.1)
n_{Worker}	11,385	5,753	11,385	5,753	11,385	5,753
n_{Firms}	1,591	149	1,591	149	1,591	149
Intercept	1.59*** (0.04)	2.86*** (0.08)	1.49*** (0.09)	2.72*** (0.25)	-0.58*** (0.12)	0.67 (0.34)
σ^2_α	0.043 (0.028)	0.027 (0.037)	0.021 (0.023)	0.000 (0.000)	0.023 (0.025)	0.000 (0.000)
Log-Likelihood	-5377.0	-1482.6	-5298.2	-1461.5	-4635.0	-1385.1

Note: First-time respondents only. All observations with missing values for any of the predictors are excluded (results for Column 1 and Column 2 are very similar if the use the full sample of first-time respondents only). Non-binary variable have been standardized by subtracting the mean and dividing by the standard deviation (before applying sample restrictions).

Table 5: Separate Estimates for Linkage Consent for East and West Germany (WeLL)

	Outcome: Linkage Consent			
	East Germany		**West Germany**	
Wave 2	-0.82***	(0.25)	-0.87***	(0.18)
Wave 3	-1.12***	(0.24)	-0.9***	(0.15)
Female	0.42*	(0.23)	-0.08	(0.17)
No Vocat. Qualif.	0.38	(0.3)	-0.1	(0.2)
Tertiary Degree	0.1	(0.29)	-0.14	(0.13)
Age ca. 35-45	-0.22	(0.22)	0	(0.15)
Age ca. 45-55	0.47*	(0.27)	0.05	(0.23)
Age over 55	0.93**	(0.42)	0.46	(0.28)
Foreign-born	0.00	(0.00)		
Part-time	-0.37	(0.25)	0.06	(0.19)
White Collar	-0.35	(0.28)	-0.25	(0.16)
Child Under 14	0.25	(0.19)	-0.08	(0.16)
Living Alone	0.19	(0.24)	-0.21	(0.22)
Good Health	0	(0.29)	0.08	(0.16)
Managerial Resp.	0.36	(0.31)	0.28**	(0.12)
Limited Contract	-0.12	(0.24)	0.03	(0.16)
Shift Work	0.44**	(0.18)	-0.01	(0.16)
Conscientiousness	0.16	(0.11)	0.15*	(0.08)
Extraversion	-0.02	(0.11)	-0.15**	(0.07)
Neuroticism	0.1	(0.12)	0.12	(0.09)
Agreeableness	-0.11	(0.13)	-0.1	(0.12)
Openness to Exp.	-0.17*	(0.1)	-0.04	(0.11)
Job Insecurity	-0.16	(0.11)	-0.02	(0.07)
Work Satistfaction	-0.11	(0.13)	-0.03	(0.07)
Intercept	2.69***	(0.33)	2.76***	(0.33)
n_{Worker}	2231		3486	
n_{Firms}	61		88	
Log Pseudo-likelihood	-453.03		-988.86	

Note: First-time respondents only. All foreign-born respondents in East Germany provided consent and have therefore been disregarded in Column 2.

Table 6: Replication of Steffes and Warnke (2016) (WeLL)

	Outcome: Participation in Training			
Intercept	0.17**	(0.08)	-0.07	(0.13)
Intercept x Refusal	-0.24**	(0.11)	-0.44	(0.44)
Female			-0.04	(0.06)
Female x Refusal			0.01	(0.21)
Cohabitating			0.06	(0.06)
Cohabitating x Refusal			0.15	(0.24)
No Voc. Qualification			-0.07	(0.08)
No Voc. Qualif. x Refusal			0.00	(0.33)
Tertiary Education			0.88***	(0.06)
Tertiary Educ. x Refusal			0.26	(0.23)
Age ca. 35-45			-0.22***	(0.07)
Age ca. 35-45 x Refusal			-0.10	(0.25)
Age ca. 45-55			-0.39***	(0.07)
Age ca. 45-55 x Refusal			-0.04	(0.26)
Age above 55			-0.69***	(0.09)
Age above 55 x Refusal			-0.76*	(0.44)
Unempl. Exp.			-0.12	(0.18)
Unempl. Exp. x Refusal			-0.64	(0.63)
Labor Attachment			0.05***	(0.01)
Labor Attachm. x Refusal			0.01	(0.04)
Foreign Born			-0.58***	(0.11)
Foreign born x Refusal			0.49	(0.37)
Health Status			-0.12***	(0.02)
Health Status x Refusal			-0.16*	(0.09)
σ^2_{Firm}	0.6	(0.09)	0.48	(0.08)
$\sigma^2_{\text{Firm}} \times$ Refusal	0.23	(0.16)	0.32	(0.19)
σ^2_{Worker}	1.4	(0.10)	1.24	(0.10)
$\sigma^2_{\text{Worker}} \times$ Refusal	0.32	(0.53)	0.02	(0.28)
n_{Worker}	17269		17269	
n_{Firms}	149		149	
Wald χ^2	358.57		749	
Log Likelihood	-11082.43		-10840.85	

Table 7: Replication of Variance Components in Steffes and Warnke (2016) (WeLL)

		(1)		**(2)**		**(3)**	
Variable	**Model**	**Coef**	**(SE)**	**Coef**	**(SE)**	**Coef**	**(SE)**
σ^2_{Firm}	*Time Effects*	0.65	(0.10)	0.61	(0.09)	0.50	(0.22)
σ^2_{Worker}	*Time Effects*	1.5	(0.16)	1.42	(0.10)	1.26	(0.51)
σ^2_{Firm}	*+Worker Characteristics*	0.48	(0.08)	0.49	(.08)	0.43	(0.20)
σ^2_{Worker}	*+Worker Characteristics*	1.29	(0.14)	1.20	(0.09)	0.81	(0.41)
n_{Firm}		149		149		132	
n_{Worker}		12,560		16,263		666	

Note: (1) Original results in Steffes and Warnke (2016), (2) Re-analyses on survey data only,
(3) Replication on non-Consent Sample. n_{Worker} refers to the number of observations (interviews).

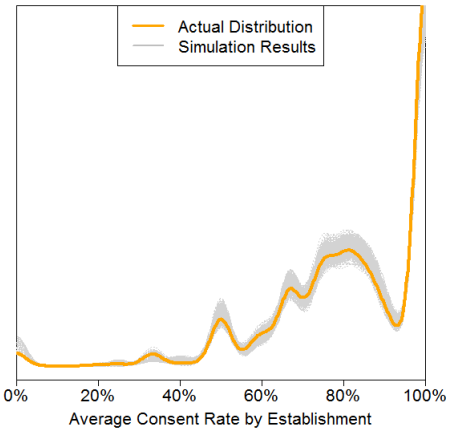
Table 8: Wage Regression (LPP-data)

		Outcome: Log Gross Hourly Wage	
		Data: LPP	
Variable		**Coef**	**(SE)**
Intercept		1.93***	(0.07)
Intercept	x Refusal	-0.16	(0.16)
2nd Wave		0.04**	(0.02)
2nd Wave	x Refusal	0.00	(0.02)
Female		-0.24***	(0.01)
Female	x Refusal	0.03	(0.02)
Poorly Educated		-0.16***	(0.03)
Poorly Educated	x Refusal	0.06	(0.07)
Highly Educated		0.34***	(0.01)
Highly Educated	x Refusal	0.01	(0.02)
Age		0.04***	(0.00)
Age	x Refusal	0.01	(0.01)
Age Squared		0.00***	(0.00)
Age Squared	x Refusal	0.00	(0.00)
Foreign-born		-0.03**	(0.02)
Foreign-born	x Refusal	-0.04	(0.03)
Good Health		0.08***	(0.01)
Good Health	x Refusal	-0.01	(0.02)
Training		0.15***	(0.01)
Training	x Refusal	0.04*	(0.02)
n		8964	
R^2		0.2553	

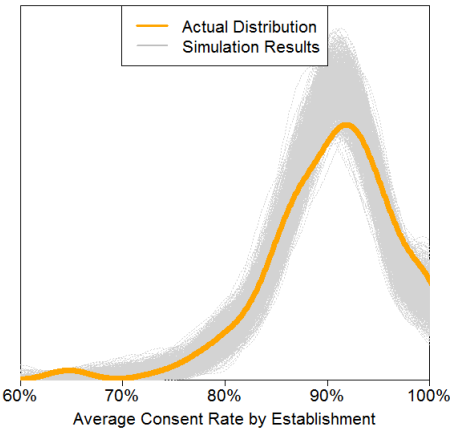
Note: First-time respondents only. Analyses restricted to individuals reporting working hours between 15 and 60 hours per week. The lowest and highest wage percentiles were trimmed.

Figure 1: Consent Rates on the Establishment Level

(a) LPP



(b) WeLL



Note: Simulation based on $10,000$ repetitions. First-time respondents only. Gaussian kernel estimation with bandwidth fixed at 0.025 (LPP) / 0.02 (WeLL).

A.2 Figures

In Chapter 5 we have shown that consent rates are partly driven by establishment heterogeneity (if no further controls are added). Another way to illustrate how consent rates differ across establishments is a simple Monte-Carlo experiment. Here we assume that establishment-wide average consent rates are a result of a series of Bernoulli trials. We model the distribution of the expect average consent rate $\bar{p}_j$ for each of the 149 establishments in WeLL $(j = 1, ..., 149)$: $p_j = \sum_{k=1}^{n_j} p$. Here $p = 89.7\%$ corresponds to the grand mean of consent rates for first-time respondents, and n_j equals the number of first-time respondents per establishment j. The result of 10,000 draws of the simulation are depicted in Figure 1. Figure 1 shows a kernel density plot of the average (aggregated) consent rate by establishment. The graph demonstrates that actual consent rates indeed seem to differ from what we would expect, as is apparent from the lower number of establishments with a consent rate around the grand average. The very different worker to establishment ratio in the *LPP*-data leads to a distinct shape of the simulation profile, but notable deviations from what we would expect due to random noise for probabilities close to zero and one.

A.3 Detailed Description of Sampling in *WeLL*

The sampling procedure in *WeLL* has been described in Bender et al. (2008) and Knerr et al. (2009). The population consists of 149 establishments located in five different German states (three states in West Germany and two states in East Germany) which were sampled in a stratified way from the IAB Establishment Panel.[39] These establishments were active in either the service sector or in the manufacturing sector and had between 100 and 2000 employees. The strata were defined on the basis of size (three categories with 200 - 500 employees being the middle category), sector (manufacturing or service sector) and location (East or West Germany). In addition, establishments were sampled according to their willingness to make investments and whether they indicated training provision. A survey of the establishments has been conducted but this has not been made available to researchers.

We will next describe the sampling of survey participants from the 149 *WeLL* establishments. We will start with the procedure used in the first wave. All employees who were subject to social security contributions and who were employed at the *WeLL*-establishment on December 31, 2006, were eligible. This excludes apprentices or workers in partial retirement and includes approximately 56,000 employees. Within this group, 20,190 individuals were sent a letter inviting them to participate in a telephone interview, along with information explaining the purpose of the survey. 16,552 individuals were finally contacted at home and 6,404 interviews were conducted. This gave a response rate of 38.7%.[40] Computer assisted telephone interviews (CATI) were conducted by infas, Bonn, between October 2007 and January 2008.

The telephone interviews lasted for an average of 32 minutes (Knerr et al., 2009). At the end of the interview, individuals were asked whether they were willing to participate in future waves of the survey, and whether they would provide consent for the interview data to be linked to their social security records. The question regarding data linkage read as follows, "We have now talked a lot about topics such as your job or your education. To shorten the interview, we would like to include for the analysis excerpts of data available at the Institute for Employment Research in Nuremberg. This includes information about previous periods of employment and unemployment. The Data Protection Act requires your consent for the purpose of linking such information to the interview data, to what I would like to ask you cordially. It is absolutely certain that all data protection regulations are strictly adhered to. Your consent is of course voluntary. You can also withdraw your consent at any time. Do you agree to this additional information potentially being merged with your details in the interview?" 91% of respondents

[39] Originally, *WeLL* had sampled 167 establishments but 18 of these had to be excluded for reasons of anonymity - less than 50 employees were eligible to participate in *WeLL*. See further details about the sampling of respondents from the establishments in the next paragraph.

[40] The difference between the 20,190 individuals invited and the final sample is due to missing addresses or telephone numbers, insufficient language skills or other similar reasons.

consented to data linkage. This consent rate is extremely high compared to those reported in the literature (see Section 2). Furthermore, 92% of respondents agreed to participate in future waves of the survey.[41]

While the second and the third waves of the survey made use of both panel participants and new respondents, the fourth wave was limited to panel respondents only. Individuals who joined a *WeLL* establishment in 2008 for the second wave (2009 for the third wave) were eligible as new respondents. This also includes apprentices who had become regular employees in the respective year. The second (third or fourth) wave was conducted in autumn 2008 (2009 or 2010, respectively).

In *WeLL*, basic establishment information which is relevant for the stratification has been made public for all respondents. This includes firm-size (100-199 employees, 200-499 and 500-1,999), sector (manufacturing or service), location (East or West Germany), whether the establishment provides further training (yes/no) and the establishment's willingness to make investments (yes/no). Further establishment variables are available for the consent sample through the link to the IAB Establishment Panel. The employer survey has not been made public.

A.4 Detailed Description of Sampling in *LPP*

LPP consists of an establishment and an individual survey. Establishments in the *LPP* were drawn from the IAB Establishment Panel 2011 (in the first wave). Letters of invitation to participate in the *LPP* establishment survey were sent to all 2,222 non-agricultural establishments with more than 50 employees. Establishments primarily owned by the state and those which operate as non-profit establishments were excluded. In total, 1,219 interviews were conducted between July and October 2012 by TNS Infratest Sozialforschung (which also organized the interviews for the IAB Establishment Panel in 2011). Further information is available in Gensicke and Tschersich (2015).

The individual survey was conducted by infas, which also carried out surveys for the collection of data for the *WeLL* dataset (Schütz et al., 2014). Individuals were drawn from 869 *LPP*-establishments which had expressed their willingness to participate in future waves of the survey and which employed a sufficient number of eligible individuals. The survey was conducted between December 2012 and April 2013 in the form of a CATI. The average duration of the interview was 30 minutes.

At the end of the interview, individuals were asked whether they would consent to data linkage. The question was similar to that asked in the surveys conducted for *WeLL*. It read as follows:

[41] The average consent rate is probably very high because the question is unspecific regarding the nature of the data with which survey information is to be merged. Furthermore, consent rates are much higher in East Germany which is over-represented in *WeLL*. This could explain the difference to *LPP*.

"To shorten the next interview by not asking about your full employment biographies, we would like to include excerpts of other data for the analysis. This data is available at the Institute for Employment Research in Nuremberg. This includes information about previous periods of employment. However, the inclusion of this data requires your consent. The Data Protection Act requires your consent for the purpose of linking such information to the interview data to what I would like to ask you cordially. It is absolutely certain that all data protection regulations are strictly adhered to. Your consent is of course voluntary. You can also withdraw your consent at any time. Do you agree to this additional information potentially being merged with your details from the interview".[42]

LPP does not include establishment information on individuals who did not consent to their data being linked to administrative records. For the sample who did provide consent (consent sample), establishment information is available both through the link to the IAB Establishment Panel and through the employer survey.

A.5 Lasso Estimation of Linkage Consent

We conduct a robustness test in order to illustrate that the identified predictors of linkage consent are indeed important and that they should therefore be taken into account in future research. In the main specification we left out a number of variables which may in fact be important. These variables include predictors which have been used in previous studies, such as marital status (available in *WeLL* only) as well as variables such as job tasks used in the field of social science and available in this study for *WeLL* only.[43] Furthermore, we have already included a large number of variables in Table 4. Multicollinearity can be an issue where unregularized methods are used for variable selection. Here, we want to assure that our findings are robust to using another approach which works well in the case of moderate multicollinearity.

Lasso regularization is a commonly used method in the machine learning literature and is suitable for variable selection (least absolute shrinkage and selection operator, Tibshirani, 1996). Lasso regularisation constrains the sum of the absolute values of the estimates. It thereby sets many coefficients to exactly zero and is therefore well suited to variable selection (Friedman et al., 2001). Many researchers prefer lasso to standard approaches such as stepwise selection models, in particular in the presence of highly correlated variables (e.g. Yuan and Lin, 2006).

[42]Individuals who agreed to participate in future waves of the survey were asked this question. The question addressed to respondents who declined to participate in future interviews was very similar.

[43]Many variables regarding the family status and household context are highly correlated. Individuals who are married for example, are likely to live with another person whilst widowed individuals are often fairly old. We offer a regularized approach in order to account for multicollinearity. It is for this reason that we have so far excluded these variables.

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N}\text{Linkage Consent}_i\cdot(\beta_0+x_i^T\beta)-\log(1+e^{(\beta_0+x_i^T\beta)})\right]+\lambda||\beta||_1. \quad (6)$$

Here x includes the 32 variables listed in Table 4 for *WeLL* and 20 other variables including marital status and household composition, 12 job tasks and future expectations regarding wage growth. We do not include any further variables for *LPP*, but we do check whether lasso confirms the results from the random effects logistic regression in Chapter 5. λ controls the amount of shrinkage and is estimated via 50-fold cross-validation.[44]

The results are overwhelmingly in line with previous findings with only a few exceptions. In *WeLL* we find that the main indicators, such as willingness to participate in future interviews, item non-response, age and employment in a firm in East Germany, are of an almost identical size. Contrasting results are found for individuals with managerial responsibilities, a predictor which exhibits only a small positive correlation with linkage consent according to lasso. The negative coefficient for white-collar workers is also only half as large as that shown in Table 4. The time effects are also much smaller and the coefficient for the third wave is even set to zero in the lasso approach. Results for *LPP* are fully in line with our previous findings.

For *WeLL*, lasso regularization suggests that alternative household situations should be included; e.g. a single household not living with a partner or being divorced rather than living alone. The negative coefficient for being divorced is, however, comparable to that found for those living alone (indeed, we find similar associations for other household situations). In addition, lasso regularization suggests that different job tasks should also be included. These include "measuring, testing, quality control" (positively associated with linkage consent), "teaching, training, educating" (positive association), "taking care, healing" (positive association), "operating, controlling machines" (positive association), "manufacturing of goods, planting" (positive association) and "repairing, renovating, restoring" (negative association). The absolute size of the coefficients in the lasso approach is between 0.16 for "measuring" and 0.05 for "repairing". If we add these variables to the specification in Table 4 column 6, we obtain similar results. Only one additional predictor, "measuring, testing, quality control", however, is significant at the 10 percentage level.

A.6 Further Details of the Replication of Steffes and Warnke (2016)

Here, we replicate our original results presented in Steffes and Warnke (2016). We restrict the replication to specifications which use only worker characteristics available in the survey data.

[44] k-fold cross-validation partitions the original data into k subsamples of equal size. $k-1$ samples are then used as a training set and one remaining sample is then used for validation. This exercise is repeated k times. As in Table 4, we use a logistic loss function.

As in our earlier paper, we exclude mandatory job-training courses and restrict the analyses to workers who are employed in a full- or part-time position and working at least 15 hours a week. This provides us with estimates comparable to those in Table 7, Columns 1 and 3, in Steffes and Warnke (2016). We excluded all individuals with missing information on key variables such as level of education or training attendance. In the earlier study we used age and age squared. In the survey data used here four age-categories are included instead.

These sample selection criteria give us a sample of 16,263 interviews and 6,731 unique respondents over four waves from whom linkage consent was obtained. Compared to the original study and for these reasons, there are approximately 29% more interviews and 16% more workers in this study. There are a number of different (related) reasons why this will provide slightly varying results. Firstly, not all variables which we used for the original data preparation are available in the survey. We cannot tell from the survey data alone, for example, whether a worker still works for a given *WeLL*-establishment. This information is, however, directly accessible from social security data. Moreover, in Steffes and Warnke (2016) we excluded individuals who left a given firm. Secondly, variables such as age are anonymised in the survey data (four categories) whilst the social security data provides exact information regarding the year in which respondents were born. We cannot therefore exclude individuals aged below 21 or above 64, nor use age continuously for the regression analyses as was the case in our earlier paper. Furthermore, in Steffes and Warnke (2016) we excluded individuals whose social security entries were missing or who received very low wages.

As expected, the replication of our results derived from the survey data gives us results similar to those seen in Steffes and Warnke (2016). There are some notable differences between the intercept and the time effects which is most probably due to the inclusion of workers who leave an establishment. In order to acquire the new skills, individuals tend to participate more in training when they have recently begun a new job.