Linking PIAAC Data to Individual Administrative Data: Insights from a German Pilot Project

Daikeler, Jessica; Gauly, Britta; Rosenthal, Matthias

Chapter 11
Linking PIAAC Data to Individual Administrative Data: Insights from a German Pilot Project

Jessica Daikeler, Britta Gauly, and Matthias Rosenthal

Abstract Linking survey data to administrative data offers researchers many opportunities. In particular, it enables them to enrich survey data with additional information without increasing the burden on respondents. German PIAAC data on individual skills, for example, can be combined with administrative data on individual employment histories. However, as the linkage of survey data with administrative data records requires the consent of respondents, there may be bias in the linked dataset if only a subsample of respondents—for example, high-educated individuals—give their consent. The present chapter provides an overview of the pilot project about linking the German PIAAC data with individual administrative data. In a first step, we illustrate characteristics of the linkable datasets and describe the linkage process and its methodological challenges. In a second step, we provide an illustrative example of the use of the linked data and investigate how the skills assessed in PIAAC are associated with the linkage decision.

11.1 The Importance of Enriching Survey Data with Administrative Data

Linking survey data to other data sources offers many opportunities, such as enriching survey data with additional information without increasing the burden on respondents (Calderwood and Lessof 2009; Sakshaug 2018; Sakshaug and Kreuter 2012). Thus, from a researcher's perspective, data linkage is a respondent-friendly, cost-effective, and quick way of generating data.

J. Daikeler (✉) · B. Gauly
GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
e-mail: jessica.daikeler@gesis.org

M. Rosenthal
University of Stuttgart, Stuttgart, Germany

D. B. Maehler, B. Rammstedt (eds.), *Large-Scale Cognitive Assessment*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-47515-4_11

In this context, linking survey data with administrative data is probably the most established method of data enrichment. Administrative data are typically provided by administrative sources, for example, notifications by employers to social security institutions or data from operational processes of employment agencies. This linkage has several benefits. For example, it provides a possibility of creating longitudinal data by linking cross-sectional survey data to administrative longitudinal data, or by linking different administrative datasets to each other. Administrative data may also contain historical records and accurate retrospective information that would be difficult or impossible to collect using traditional survey methods. And, at least in theory, administrative data contain information that provides full coverage of the population of interest (Calderwood and Lessof 2009). The data are neither affected by recall error, nor can they suffer from other deficiencies of survey data, such as social desirability bias, systematic drop-outs and item nonresponse, or panel mortality. Furthermore, the linkage of survey data with administrative data allows for a validation of survey data, for example, on earnings (Gauly et al. 2019; Sakshaug and Antoni 2017).

Despite its potential benefits, data linkage has methodological and practical challenges. The validity and usability of the linked data depend on respondents' consent to data linkage. Refusal to give this consent can result in a biased sample, especially if those who consent to the linkage differ significantly in their characteristics from those who refuse (Al Baghal et al. 2014). Moreover, these differences can create biased estimates obtained from linked data (Sakshaug and Kreuter 2012). However, the explicit consent to linkage by the respondents is necessary in order to comply with privacy rights and data protection policies.

The present chapter provides an overview on how to work with the data of the German sample of the Programme for the International Assessment of Adult Competencies (PIAAC), which was linked to administrative data held by the Institute for Employment Research (IAB), the research institute of the German Federal Employment Agency (BA). The resulting linked dataset, PIAAC-L-ADIAB, is part of a pilot project. The next section describes the linkage process and the challenges that it involves. Section 11.3 provides an illustrative example of data linkage, with a focus on the role of cognitive skills in the respondent's decision to consent to linkage. Section 11.4 concludes with practical recommendations for the linkage of PIAAC data to administrative records.

11.2 Linking PIAAC Data to IEB Data

The administrative data that are linked to the data of the German PIAAC 2012 sample are the data from the Integrated Employment Biographies (IEB) of the IAB. The IEB contain information on every individual in western Germany since 1975 and in eastern Germany since 1992 who has one of the following statuses: in employment subject to social security (recorded from 1975 onwards); in marginal part-time

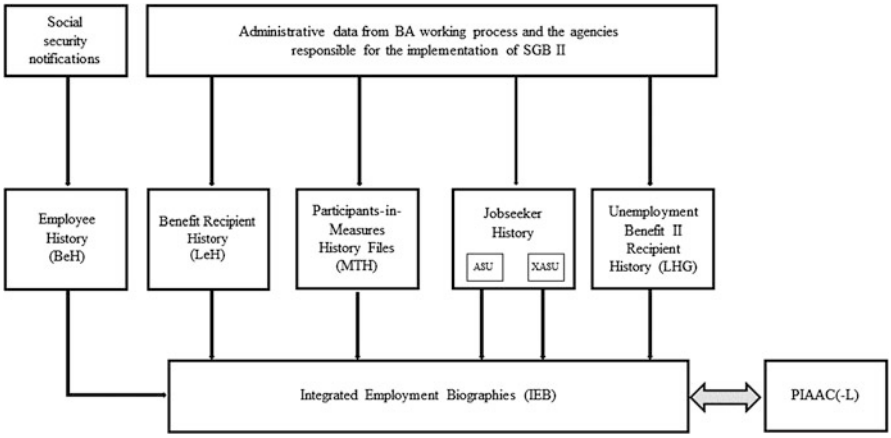


Fig. 11.1 Data sources of the Integrated Employment Biographies (Source: adapted from Antoni et al. 2017)

employment (recorded from 1999 onwards)[1]; in receipt of benefits in accordance with German Social Code[2]; registered as a jobseeker (recorded from 1997 onwards); or participating in an active labour market policy programme (recorded since 2000; for more details see Antoni et al. 2019a). The data originate from different sources within the German social security system: data on employment spells stem from compulsory notifications by employers to social security agencies; data on benefit receipt, job search spells, and participation in labour market programmes are entered mainly by caseworkers at the local employment agencies (see Fig. 11.1).

The consent question on linking survey data with administrative data was not part of the original PIAAC 2012 survey, but of the German follow-up panel study PIAAC-Longitudinal (PIAAC-L). This study followed up the German PIAAC respondents and comprised three additional waves conducted in 2014, 2015, and 2016 (for detailed information on PIAAC-L, see Rammstedt et al. 2017).

[1]Marginal part-time employment: (1) short-term employment with a maximum duration of 3 months or a maximum of 70 working days per calendar year; (2) employment with a monthly salary of no more than 450 euros; (3) employment in private households as a special type of marginal part-time employment.

[2]This comprises benefits according to Social Code Book III—namely, unemployment benefit, unemployment assistance and maintenance allowance (since 1975), as well as benefits in accordance with Social Code Book II, which covers both basic social security benefits (e.g. Unemployment Benefit II) and supplements to unemployment benefit or additional benefits (since 2005; Antoni et al. 2019a).

Preliminary remarks:
In the interview, various topics were discussed, including work and occupation. In order to be able to carry out detailed statistical analyses in the field of employment, we would like to include data from another database in the analysis. As already explained in our data protection sheet, these data (see below) are available from the Institute for Employment Research (IAB) of the Federal Employment Agency (BA) in Nuremberg. Further information on these BA/IAB data can be found at http://fdz.iab.de/de/FDZ_Overview_of_Data.aspx.
For data protection reasons, the IAB requires the following consent in order to merge the data, which we would like to ask you to give below. Your consent is of course voluntary. Participation in PIAAC-L is possible independently of this consent.

Consent to record linkage of selected register data with your data from the PIAAC-L/PIAAC surveys

I consent to a query being sent to the Institute for Employment Research (IAB) of the Federal Employment Agency (BA) in Nuremberg regarding data that are available about me. These data are information about my working life, employment relationships and spells of unemployment. My contact details may be sent to the IAB to request this information. The IAB data may then be merged with my data from the PIAAC-L/PIAAC surveys in order to include them anonymously in the analyses. My contact data shall be deleted after the query has been carried out. I have been informed that all data protection regulations will be complied with. I reserve the right to withdraw my consent at any time.

Fig. 11.2 English translation of consent to linkage question
(Source: see Steinacker and Wolfert 2017, for the original German-language version)

All PIAAC 2012 anchor persons[3] who participated in the second wave (2015) of the German PIAAC-L study were asked at the end of the interview to consent to the linking of their survey data to administrative data from the IAB (see Fig. 11.2 for the English translation of the linkage question). The linkage required written consent, and respondents could give this consent in two different ways—directly at the interview, or afterwards by sending the consent form to the survey institute at their discretion (Zabal et al. 2017).

In total, 2363 (72.4%) of the 3263 anchor persons in PIAAC-L 2015 gave their consent to the linkage of their survey data to administrative data. For these respondents, personal information (including name, name at birth, date of birth, gender, and place of residence) were transmitted to the IAB.[4] This information was

[3]The PIAAC anchor persons are those respondents who already participated in the original PIAAC 2012 survey. In addition, PIAAC-L also surveyed partners and household members of the PIAAC 2012 anchor persons.

[4]Linkage was performed by the staff of the German Record Linkage Center (GRLC; see Antoni and Schnell 2019).

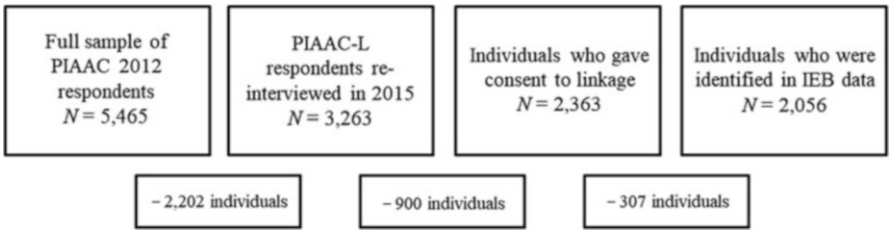


Fig. 11.3 Drop-out of respondents from PIAAC 2012 to PIAAC-L 2015 and PIAAC-L-ADIAB

subsequently used to identify the respondents in the IEB data (for more detailed information on the linkage procedure, see Braun 2016; GESIS – Leibniz Institute for the Social Sciences and Institute for Employment Research (IAB) 2017).

11.2.1 Sample Differences

Of the PIAAC-L 2015, participants who gave their consent to data linkage, 2056 (87%) could be identified in the IEB data.[5] Thus, the sample of individuals that can be used for joint analyses with PIAAC and IEB data (referred to in what follows as 'PIAAC-L-ADIAB') is significantly smaller than the full sample of individuals participating in PIAAC-L 2015 ($N = 3263$) as well as the original and representative PIAAC 2012 sample ($N = 5465$) due to sample attrition, missing consent, and missing IEB data (see Fig. 11.3).

Table 11.1 presents the (unweighted) distributions of sociodemographic characteristics in the various samples. It is clear from the table that the samples differ with regard to the distribution of these characteristics. For example, individuals in the PIAAC-L-ADIAB sample are on average older compared with all individuals who participated in PIAAC-L 2015 or compared with those who gave their consent to data linkage. In addition, the share of women and the share of individuals with a primary or lower secondary qualification (ISCED 1 and 2) are lower in PIAAC-L-ADIAB. Thus, researchers working with the linked PIAAC-L-ADIAB data have to be aware of sample selection bias. For example, if only high-educated individuals consent to linkage (as the unconditional distribution in Table 11.1 suggests), the average educational level is higher in the linked sample, and the estimated relationships between education and any other variable might be biased. Following this, results obtained from the linked sample cannot be generalised to the population at large.

[5]Civil servants, self-employed persons, and individuals doing only unpaid domestic work are not included in the IAB data unless they have previously or incidentally been in dependent employment, registered as unemployed, registered as jobseekers, or had one of the other statuses mentioned in Sect. 11.2. Thus, these respondents could not be identified and linked in the IAB data.

Table 11.1 Sample statistics for PIAAC 2012, PIAAC-L 2015, the 'linkage consent sample', and PIAAC-L-ADIAB

	PIAAC 2012	PIAAC-L 2015	Linkage Consent	PIAAC-L-ADIAB
	$N = 5465$	$N = 3263$	$N = 2363$	$N = 2056$
Age (mean)	39.8	44.4	44.2	50.9
Female (%)	51.0	51.3	48.9	43.9
Education (%)	*(m = 90)*			
ISCED 1/2	17.0	8.9	8.2	8.6
ISCED 3/4	51.5	55.2	54.8	57.1
ISCED 5B	12.0	11.9	12.3	12.1
ISCED 5A/6	19.5	24.1	24.7	22.2
Native speaker (%)	89.0 *(m = 90)*	91.8	93.4	93.4
Employed (%)	75.7 *(m = 88)*	77.1	77.5	78.2
Eastern Germany (%)	20.5	21.6 *(m = 1)*	22.4 *(m = 1)*	23.1 *(m = 1)*

Notes. The percentages refer to the persons for whom valid values are available. The numbers in parentheses *(m = . . .)* indicate the number of missing values for each variable. No weighting included. There are no significant differences in any of the variables listed here between the PIAAC-L 2015 and the linkage consent sample (indicated by t-test)

11.2.2 Working with the Linked Data

In order to be able to access and use the linked data, a number of steps were required within the pilot project.[6] First, a data usage agreement with the PIAAC Research Data Center (RDC) at GESIS – Leibniz Institute for the Social Sciences for using PIAAC and PIAAC-L data is mandatory. Second, a data usage application must be submitted to the IAB's RDC. Once the application is confirmed, an agreement must be concluded between the researcher and the IAB. The IAB then issues the researcher with a user ID and a password for data access.

As the use of administrative data is subject to restrictions under data protection legislation, the data must be analysed either on-site or via remote access. The advantage of on-site use is the opportunity to view the results directly. Remote data access means that researchers submit scripts (e.g. Stata do-files) to the RDC and, after verification of compliance with data protection legislation, access the approved results.[7]

If data access is granted and the technical requirements are fulfilled, the next step is the data merging. The personal identifier (*SEQID*) from PIAAC is added to the IEB data, thereby rendering it possible to merge the two data sources via the identifier.

[6] All the following administrative steps and the list of available variables (Table 11.2) refer to data access within the pilot project. When accessing PIAAC-L linked data in the future, these steps and the variables available may be different.

[7] For more information, see https://fdz.iab.de/en/FDZ_Data_Access/FDZ_Remote_Data_Access.aspx.

```
. clear all
. version 14

. * --------- ( merge PIAAC 2012 and IEB data ) --------- *

. global data_path "..."
.
. *** load PIAAC data ***
.
.
. use "$data_path\ZA5845_v2-2-0.dta"
.
.
. sort SEQID
.
.
. *** merge with IEB data ***
.
.
. merge 1:m SEQID using "$data_path/PIAAC-L-ADIAB_7514_v1_SUF-ID.dta"
```

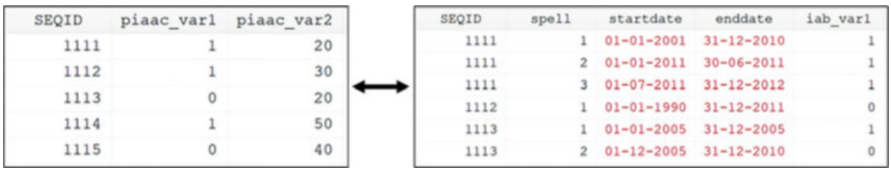
Fig. 11.4 Example of syntax to merge PIAAC 2012 and IEB data

Figure 11.4 provides a simple example of the 'merge' syntax for linking the PIAAC and IEB data in Stata.[8] As is apparent from the figure, the data are not linked 'one to one' (1:1) but rather 'one to many' (1:m). This means that one row (respondent) in PIAAC is merged with multiple rows in the IEB data. This is due to the 'spell format' of the IEB data, which means that there are several observations per person, each covering a period of time (spell) during which the person had a given employment status—for example, employed or unemployed. A new spell is created whenever an employer reports new information (e.g. change in employment status, change in establishment, or change in daily wage).

Figure 11.5 provides a fictional example of the spell structure: like any other person in PIAAC, Respondent no. 1111 has exactly one row in the PIAAC data (left side of the graph). However, as Respondent no. 1111 has three different entries in the IEB data (e.g. due to a period of unemployment between January and June 2011), he or she has three rows (spells) in the IEB data (right side of the graph). By linking the PIAAC data with the IEB data, the information from PIAAC is replicated and passed to each of the respondent's three rows in the IEB data ranges.

Unfortunately, it also happens that some of the spells in the IEB data overlap, which means that different information may be available for an individual for the same period of time (see, e.g. Figs. 11.5 and 11.6, Respondent no. 1113 in the period between December 1, 2005, and December 31, 2005). To create completely nonoverlapping periods, so-called episode splitting is performed as shown in Fig.

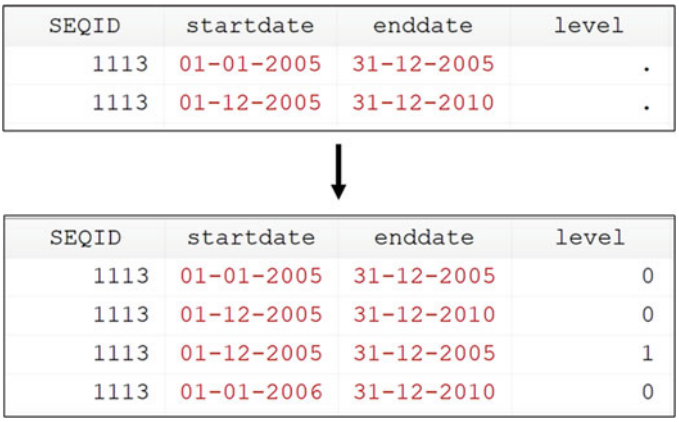
[8]This example refers only to the linkage of the administrative data with the data from PIAAC 2012. The additional waves of PIAAC-L have to be merged separately (via the personal identifier *SEQID*).



SEQID	piaac_var1	piaac_var2
1111	1	20
1112	1	30
1113	0	20
1114	1	50
1115	0	40

SEQID	spell	startdate	enddate	iab_var1
1111	1	01-01-2001	31-12-2010	1
1111	2	01-01-2011	30-06-2011	1
1111	3	01-07-2011	31-12-2012	1
1112	1	01-01-1990	31-12-2011	0
1113	1	01-01-2005	31-12-2005	1
1113	2	01-12-2005	31-12-2010	0

Fig. 11.5 Individual data given in PIAAC and IEB
Notes. The left-hand side shows an example of data provided in the PIAAC survey. The right-hand side shows an example of the PIAAC-L-ADIAB data, which contains information from both PIAAC and IEB data



SEQID	startdate	enddate	level
1113	01-01-2005	31-12-2005	.
1113	01-12-2005	31-12-2010	.

SEQID	startdate	enddate	level
1113	01-01-2005	31-12-2005	0
1113	01-12-2005	31-12-2010	0
1113	01-12-2005	31-12-2005	1
1113	01-01-2006	31-12-2010	0

Fig. 11.6 Example of episode splitting

11.6.[9] In this way, the episodes (period between December 1, 2005, and December 31, 2005) are replaced by artificial episodes with new start and end dates.[10]

Once the two datasets have been linked, users can access a wide range of additional labour market-related information. Table 11.2 provides a list of the variables that were available in the IEB data during the pilot project (for an overview, also see Antoni et al. 2019a). In addition to these variables, further sensitive characteristics, such as nationality and occupational subgroup, can be requested from the RDC. However, as these variables would enable the identification of particular individuals or establishments, they are disclosed in their original form only if it is necessary for the study objective and explicitly justified in the application for data access. The specific variables that are classified as sensitive are documented in Antoni et al. (2019a).

[9]The *level* variable counts how often information is repeated for the same time period. The value '0' indicates that the information is available for the first time; '1' indicates the first repetition.

[10]'Episode splitting' is part of the preparation of the IEB data and is not specifically related to PIAAC. For more details, see Antoni et al. (2019a).

Table 11.2 Variables available in IEB data

Variable name	Variable label
Persnr	Individual ID
Betrnr	Establishment ID
spell	Observation counter per person
quelle	Source of spell
begorig	Original start date of observation
endorig	Original end date of observation
begepi	Start date of split episode
endepi	End date of split episode
frau	Gender
gebjahr	Year of birth
nation_gr	Nationality, aggregated
famst	Marital status
kind	Number of children
ausbildung	Vocational training
schule	School leaving qualification
tentgelt	Daily wage, daily benefit rate
beruf	Occupation—current/most recent (KldB 1988)[a]
beruf2010_3	Occupation—current/most recent (KldB 1988)[a]
niveau	Level of requirement—current/most recent (KldB 2010)[a]
teilzeit	Part-time status
erwstat	Employment status
gleitz	Transition zone
leih	Temporary agency work
befrist	Fixed-term contract
grund	Reason of cancellation/notification/termination
estatvor	Employment status prior to job search
estatnach	Employment status after job search
profil	Client profile
art_kuend	Type of termination of last job
arbzeit	Desired working hours of the job sought
restanspruch	Residual claim/planned duration
treager	Type of institution
alo_beg	Start of date of unemployment
alo_dau	Duration of unemployment
wo_bula	Place of residence: federal state (*Bundesland*)
wo_rd	Place of residence: regional directorate

Notes.[a] KldB = Klassifikation der Berufe (German Classification of Occupations)

The linked data can then be used for various substantive and methodological research questions. Substantive research questions may focus, for example, on the relationship between earnings development (IEB data) and skill level (survey data). Methodological questions that exploit the potential of these two data sources may

deal, for example, with the evaluation of the measurement error in the survey data by assessing it against (less error-prone) administrative data (see also Antoni et al. 2019b; Gauly et al. 2019).

In the next section, we examine the role of cognitive skills in the respondent's decision to consent to data linkage.

11.3 Illustrative Example: Is Consent to Linkage Less Likely Among Low-Skilled Individuals?

In the present example, we extend existing research on the determinants of consent to linkage. In particular, we explore the role of cognitive skills. The sociodemographic correlates of linkage consent have been well researched. For example, previous research has shown that education, and thus human capital, has a strong and positive association with consent to linkage (see Table 11.3). However, education, which has been tested in a large number of studies, is only a proxy for a person's concrete ability and skills (see, e.g. Hanushek et al. 2015) and might not give sufficient insight into how abilities and skills are related to the decision to consent, or withhold consent, to data linkage. So far, comprehensive evidence on the role of skills in the respondent's decision to consent to data linkage is missing, as survey data containing objective skill measures are in short supply.

For researchers who work with the linked PIAAC(-L) data, it is important to know whether the linked sample differs significantly from the initial sample and to be aware of the mechanisms involved in the linkage consent process. As the majority of analyses with PIAAC data involve the skills assessed, our analysis focuses on the relationship between skills and consent to linkage.

As an example for such possible mechanisms, low-skilled individuals who receive public benefits are highly dependent on the decision of the institutions that allocate the benefits. Thus, these individuals may be more sceptical when asked for additional information, anticipating a potential change in their benefits compared to medium-skilled individuals who have less contact with institutions. High-skilled individuals are less dependent on the institutions' decisions, but may follow public debate on data security more closely. This may lead to a higher sensitivity for the transfer of sensitive data and a higher rate of linkage refusals compared with low- or medium-skilled persons who follow public debate less closely.

The next section provides an overview of existing literature on the determinants of consent to data linkage before we present our own analysis strategy and results.

Table 11.3 Overview of relationship between sociodemographic variables and consent to linkage

	Antoni (2013)	Baker et al. (2000)	Bates and Pascale (2005)	Beste (2011)	Dahlhamer and Cox (2007)	Finkelstein (2001)	Haider and Solon (2000)	Hartmann and Krug (2009)	Jenkins et al. (2006)	Knies et al. (2012)	Knies and Burton (2014)	Korbmacher and Schroeder (2013)	vPascale (2011)	Sakshaug et al. (2012)	Sala et al. (2010)	Warnke (2017)
Age	−	ns	−	ns	−	ns	ns	ns	+	−	ns	+	−	ns	−	+
Female	ns	ns	−	ns	−	ns	+	−	ns	−	−	ns	+	ns	− / ns	ns
Education	ns		+	ns	−/ns		ns	ns	ns	+	+/−	ns	+	+	+	−
Literacy	ns															
Numeracy	ns													ns		
Native speaker	ns			+	−											
Foreign citizen	ns			−	−		ns	−						ns		ns
Employed	+						+	ns			ns	ns		ns		
Health status		ns	ns		+		+		ns	+/ ns	+ / ns		ns	ns	ns	
Interview duration								ns	+			ns				
Occupation	s			ns				ns			s/ns					s/ns
Income	ns		+	+	ns	ns	+	+	−	ns	+/ ns	+	ns	ns	ns	
Eastern Germany	+							+				+				+
Country	GER	UK	USA	GER	US	USA	USA	GER	UK	UK	UK	GER	USA	USA	UK	GER

Notes. The overview primarily includes studies that simultaneously tested multiple correlations of demographic variables and other predictors with consent to linkage. We report the results of the full models. Cells containing multiple entries represent different results for different samples within the same publication. *s* significant; ns = not significant, + positive significant, − negative significant, *GER*: Germany, *USA*: United States of America, *UK*: United Kingdom

11.3.1 Previous Evidence on Consent to Linkage

As mentioned above, the linkage of survey data to administrative data offers many possibilities and advantages, not only for researchers (e.g. enhanced data variety or the creation longitudinal datasets) but also for respondents (shortening the survey). However, the rates of consent to linkage vary depending on the cultural context, survey content, and sample. For example, linking survey data from the National Educational Panel Study to administrative data, Antoni et al. (2018) obtained a consent rate of over 90%; Sakshaug and Kreuter (2014) were able to achieve consent rates of 60% in a stratified random sample in Germany, whereas in the British Household Panel Survey, Sala et al. (2010) achieved only 32% consent to linkage with administrative data.

When explaining these variations in the rate of consent to linkage, most of the literature has focused on respondents' characteristics (Sakshaug et al. 2012). Common determinants of consent include age, gender, income, foreign citizenship, health/disability status, and benefit receipt (Knies and Burton 2014; Sakshaug and Kreuter 2012; Sala et al. 2010).

Table 11.3 summarises studies that have examined the association between sociodemographic variables and the decision to consent to the linkage of survey data to other data sources. Surprisingly, the findings of previous studies vary considerably in almost all sociodemographics, and it is hard to identify specific variables that consistently influenced the decision to consent to linkage across all studies.

Age, for example, was found in seven studies to have no correlation with linkage consent (e.g. Knies and Burton 2014; Sakshaug et al. 2012) and in six studies to have a negative correlation (e.g. Antoni 2013; Sala et al. 2010). Three studies found that age had a positive correlation, suggesting that consent to linkage becomes more likely with increasing age (e.g. Jenkins et al. 2006; Warnke 2017).

Education is the only variable that was found by almost all the studies considered to have a significant association with linkage consent (e.g. Knies and Burton 2014; Sala et al. 2010). With three exceptions (Dahlhamer and Cox 2007; Knies et al. 2012; Warnke 2017), higher-educated respondents were found to be more likely than lower-educated respondents to consent to linkage.

The two studies that directly investigated the association between skills, in terms of literacy and numeracy, and linkage consent (Antoni 2013; Sakshaug et al. 2012) found no correlation between the two variables. However, both of these studies exhibit shortcomings: Antoni (2013) used only self-reported (and, thus, subjective) skills measures, and Sakshaug et al. (2012) focused only on a restricted sample (adults aged 50 years or older). Therefore, Antoni's (2013) results cannot be generalised to objective skill measures, and Sakshaug et al.'s (2012) results cannot be generalised to the population at large.

In the present study, we contribute to closing this research gap by investigating whether objective measures of individual skills are associated with respondents' willingness to consent to linkage with administrative data.

11.3.2 Estimation Strategy and Measures

Our main goal in this research was to estimate the relationship between the skills assessed in PIAAC and individuals' consent to the linkage of their survey data to administrative data. To that end, we applied logistic regression models and calculated average marginal effects (AMEs):

$$\Pr\left(consent_i = 1|\boldsymbol{X}\right) = G\left(\boldsymbol{\beta X}\right) \tag{11.1}$$

where i indicates the individual and $G(\bullet)$ is a standard logistic cumulative distribution function yielding a logit model. *Consent* is a dummy variable that equals *1* if an individual gave consent to linkage and *0* otherwise, $\boldsymbol{X}$ is a vector of covariates, and the coefficient vector $\boldsymbol{\beta}$ contains parameters to be estimated.

We conducted several different regression analyses. As our key explanatory variables, we analysed the cognitive skills assessed in PIAAC. Thus, our first three models included either a standardised (mean, 0; standard deviation, 1) measure of numeracy, literacy, or problem solving in technology-rich environments (PS-TRE) skills.[11]

In our second set of models, we focused only on numeracy skills, 'the ability to access, use, interpret, and communicate mathematical information and ideas in order to engage in and manage the mathematical demands of a range of situations in adult life' (OECD 2013). As there is a strong correlation between all three skill domains, ranging from 0.753 to 0.872, we decided to report results for numeracy skills in the main model only.[12]

Additionally, we included control variables that previous studies have identified as common predictors of consent to linkage, in order to control for spurious correlations between skills and consent to linkage: age (continuous, in years); gender (1 = *female*, 0 = *male*); education (four categories; 1 = *ISCED 1/2*; 2 = *ISCED 3/4*; 3 = *ISCED 5B*; 4 = *ISCED 5A/6*); native language (1 = *non-German*, 0 = *German*); region (1 = *eastern Germany*, 0 = *western Germany*); and employment status (three categories; 1 = *employed*; 2 = *unemployed*; 3 = *non-employed*). Furthermore, we added the total duration in minutes of the survey interview in 2015 as a proxy for respondent burden. For individuals who were employed at the time of the survey, we additionally included their occupational group (four categories: 1 = *elementary*; 2 = *semi-skilled blue-collar*; 3 = *semi-skilled white-collar*; 4 = *skilled*) as well as the quartile of their monthly net income (four categories).

We present our results in Table 11.4.

[11] Plausible values were taken into account in all models. For detailed information on the definition and assessment of skills in PIAAC, see Chap. 3 in the present volume.

[12] Sensitivity analyses showed the results for the other skills to be very similar. Results are available from the authors on request.

Table 11.4 Probability of giving consent to linkage (average marginal effects and standard errors obtained from logistic regression models)

	(1)	(2)	(3)	(4)	(5)
Numeracy (std.)	0.042*** (0.010)			0.014(0.012)	0.006(0.015)
Literacy (std.)		0.041*** (0.010)			
PS-TRE (std.)			0.031*** (0.010)		
Education (ref: ISCED 3/4)					
ISCED 1/2				0.047(0.038)	0.078(0.055)
ISCED 5B				0.076# (0.045)	0.122# (0.063)
ISCED 5A/6				0.090* (0.043)	0.141* (0.061)
Age				−0.001# (0.001)	−0.001(0.001)
Female				−0.018(0.019)	−0.022(0.025)
Non-native speaker				−0.151*** (0.040)	−0.137** (0.048)
LF status (ref: employed)					
Unemployed				0.079# (0.046)	
Not employed				0.007(0.025)	
Eastern Germany				0.024(0.023)	0.043(0.027)
Interview duration				0.002*** (0.001)	0.001(0.001)
Occupation (ref: elementary)					
Semi-skilled white-collar					0.014(0.030)
Semi-skilled blue-collar					−0.037(0.037)
Elementary					−0.008(0.052)
Monthly net income (ref:Q1)					
Q2					0.040(0.033)
Q3					0.031(0.035)
Q4					0.021(0.036)
R^2	0.010	0.010	0.004	0.026	0.026
N[a]	3256	3256	2804[b]	3256	2232

Notes.[a] Our sample size of 3256 in these analyses is based on all respondents for whom there were no missing values on any of our explanatory variables. When we focused only on employed individuals, our sample size dropped to 2232 (Column 5). [b] We had less observations in the model with PS-TRE skills, as only those persons who had computer experience and were willing to participate in the computer assessment can have values for these skills

LF status = labour force status

$p \leq 0.1$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$

11.3.3 Results: Do Cognitive Skills Influence the Linkage-Consent Decision?

Our results show that, in the baseline models (that included only literacy or numeracy or PS-TRE skills), skills had a positive association with the decision to consent to data linkage (Table 11.4, Columns 1–3). All three measures were positive and highly significant (numeracy: 0.042***; literacy: 0.041***; PS-TRE: 0.031***), which means that the higher the skills, the higher is the likelihood to consent to linkage.

Adding control variables to the model including numeracy skills, we observed that educational level had a positive association with the linkage-consent decision (Columns 4 and 5 in Table 11.4). The higher a respondent's level of education was, the more likely he or she was to agree to data linkage. Furthermore, we found a lower probability of consenting to linkage in PIAAC if German was not the respondent's first language. We also found a small significant positive correlation for the duration of the interview, which was probably due to reverse causality, whereby consenting to linkage resulted in a longer interview. After controlling for the sociodemographic variables and interview duration, the significant association with skills disappeared. Age, gender, and income did not influence the linkage decision in any of the models.

11.3.4 The Role of Cognitive Skills in Consent to Data Linkage

In the present example, we hypothesised that numeracy, literacy, and PS-TRE skills measured in PIAAC were related to respondents' decision to consent, or refuse consent, to the linkage of their PIAAC(−L) data to administrative employment history data of the IAB. Our results show that, in models without control variables, all three skill measures correlated positively with consent to linkage. This means that the higher a person's skills were, the more likely he or she was to consent to the linkage of his or her survey data to the administrative employment history records. In other words, in our baseline models, individuals with low skills were less likely to consent to data linkage.

With this knowledge, questionnaire designers could use responsive design to adapt their linkage question to low-skilled respondents. This means that, depending on the skill level achieved in the PIAAC survey, the question of consent to data linkage would be individually adjusted. However, this presupposes that the skill value of the respondent is known before the linkage question is asked. Responsive design would allow the targeted addressing of respondents' concerns. For example, for individuals with low skills, the question could be formulated in more simple language. Of course, data protection provisions would still have to be adhered to and privacy concerns addressed. It could also be emphasised that, during the linkage process, researchers are independent of institutions such as the employment agency.

We found that the decision of PIAAC(-L) respondents to consent to linkage was not affected by the sociodemographic variables gender, age, income, or employment status. However, respondents' education and native language (German/non-German) did play a particularly important role in the consent decision. These results are largely consistent with previous literature, which has identified mixed findings for sociodemographic variables and their connection with the linkage decision. However, especially the significant correlations revealed for education and native language suggest that respondents may not be able to properly understand the linkage question and its implications and that further effort should be invested in the framing of this question (e.g. Sakshaug et al. 2013).

11.4 Conclusion

The focus of this chapter was on describing the process of linking data from the German PIAAC(-L) sample to administrative data of the IAB. We focused on the technical side of data linkage and the methodological challenges involved. In addition, we provided a summary of recent findings on selective consent to data linkage and illustrated how the cognitive skills assessed in PIAAC affect the decision to consent—or withhold consent—to data linkage.

The use of linked datasets has a number of advantages: survey data can be enriched with additional information without increasing respondent burden, cross-sectional surveys can be extended with longitudinal information from other data sources, and the quality of the survey information can be evaluated against an additional (more objective) data source. Thus, linked data samples allow researchers to explore completely new fields of research.

By using the linked PIAAC-L-ADIAB sample, for instance, researchers can address questions concerning the relationship between the individual labour market history and cognitive skills. From a survey methodology perspective, the linked dataset provides many opportunities, such as research on consent to data linkage, as well as possibilities for the evaluation of survey and administrative data (Gauly et al. 2019; Sakshaug and Antoni 2017).

However, the use of linked data also involves challenges. First, when combining PIAAC and IEB data, researchers have to be aware that the latter are available in so-called spell format. This means that not only one but rather several pieces of information from the administrative data will be linked to each respondent in the survey data and that a number of steps are required before the researcher can access and use the linked data.

Second, researchers face challenges in the use of the linked PIAAC-L-ADIAB data due to the small sample size. The linkage question was included only in PIAAC-L 2015, so only those individuals who participated in both PIAAC 2012 and PIAAC-L were asked for their consent to the linkage of their survey data with administrative data. Of those respondents who were asked for their consent, only a subsample agreed to the linkage; and of those who agreed, only a subsample could

be identified within the administrative data. Thus, we were left with a total sample of only 2056 individuals in the linked PIAAC-L-ADIAB dataset, which reduces statistical power and makes subsample analyses difficult.

And finally, there is a risk of selection bias in the linked PIAAC-L-ADIAB dataset. This arose for two reasons. The first selection occurred at the transition from PIAAC 2012 to PIAAC-L 2015. Research shows that PIAAC respondents who were willing to participate in the first wave of the longitudinal PIAAC survey differed significantly in terms of educational attainment, literacy skills, and native language from the initial PIAAC 2012 sample (Martin et al. 2019). The second selection resulted from the consent to data linkage, as the distribution of the sociodemographic characteristics in the linked dataset differs from that in the PIAAC-L 2015 sample (see Table 11.1). Numerous studies have shown that individual characteristics influence the consent to link survey with administrative data (see Table 11.3). As a result, not all sociodemographic groups are adequately represented in the linked dataset, and analyses will not obtain representative results. For instance, in the PIAAC-L-ADIAB sample, higher-educated individuals are overrepresented, which can lead to bias in the estimation of the relationship between education and any other variable.

In the example of the use of the PIAAC-L-ADIAB data presented here, we showed a positive and statistically significant association of the skills assessed in PIAAC and respondents' willingness to consent to data linkage. Our results indicate that individuals with low skills are less likely to consent to linkage than their high-skilled peers. However, this finding holds only for the zero-order correlations (models without control variables), as the coefficients became statistically insignificant when we controlled for individual characteristics. Moreover, our results show no statistically significant relationship between the decision to consent to linkage and sociodemographic variables, such as gender, age, income, or employment status. These results are largely consistent with previous literature, which has shown mixed findings for sociodemographic variables and their connection with the linkage decision. In contrast, respondents' education and native language (German/non-German) seem to be associated with the consent decision, which suggests that consent is related to the comprehension of the linkage question. In the light of these findings, further research should be conducted on how linkage questions should be framed and how responsive design could be used to achieve a higher linkage rate and low linkage bias (e.g. Sakshaug et al. 2013).

However, our findings do not imply that all individual characteristics play a negligible role in the linkage decision and that all individuals have the same probability of being represented in a linked dataset. This would suggest that there were no differences between the PIAAC-L 2015 sample, the original PIAAC 2012 sample, and the subsample of individuals who consented to linkage with administrative information. Instead, as can be seen from Table 11.1, the linked dataset (PIAAC-L-ADIAB) and the PIAAC-L 2015 and PIAAC 2012 datasets differ in terms of the sociodemographic characteristics of the respective samples. The decision to participate in PIAAC-L seems to have been affected by certain characteristics (see, e.g. Martin et al. 2019), and this sample selection bias translated

also into the PIAAC-L-ADIAB sample. This suggests that future surveys would benefit from including the linkage question in the first wave of a (longitudinal) survey. In that way, panel mortality would not have a distorting effect on the sample that is asked the consent question and that could potentially be part of a linked dataset. However, we also found noticeable differences in the share of females and the average age between PIAAC-L 2015 and PIAAC-L-ADIAB when we considered unconditional sample differences. This can probably be explained by the fact that (especially older) women are less likely to be employed. Similarly, younger people are often not yet employed, which contributes to the age bias. Summing up, we want to emphasise that researchers working with the linked data need to be aware of these biases, which preclude the drawing of conclusions for the general population. These sample selection biases may lead to over- or underestimation of true effects, depending on whether the panel attrition is systematic, which would mean, for example, that lower-educated individuals who consent to data linkage are significantly different in unobserved characteristics than low-educated individuals who withhold consent to linkage.

References

Al Baghal, T., Knies, G., & Burton, J. (2014). *Linking administrative records to surveys: Differences in the correlates to consent decisions* (Understanding Society Working Paper Series No. 9). Colchester: Institute for Social and Economic Research University of Essex.

Antoni, M. (2013). *Essays on the measurement and analysis of educational and skill inequalities*. Nuremberg: Institute for Employment Research.

Antoni, M., & Schnell, R. (2019). The past, present and future of the German Record Linkage Center (GRLC). *Jahrbücher für Nationalökonomie und Statistik, 239*(2), 1–13. https://doi.org/10.1515/jbnst-2017-1004

Antoni, M., Dummert, S., & Trenkle, S. (2017). *PASS-Befragungsdaten verknüpft mit administrativen Daten des IAB (PASS-ADIAB) 1975–2015* (FDZ Datenreport, 06/2017). Nuremberg: Institute for Employment Research.

Antoni, M., Bachbauer, N., Eberle, J., & Vicari, B. (2018). *NEPS-SC6 survey data linked to administrative data of the IAB (NEPS-SC6-ADIAB 7515).* (FDZ-Datenreport, 02/2018). Nuremberg: Institute for Employment Research.

Antoni, M., Schmucker, A., Seth, S., & vom Berge, P. (2019a). *Sample of Integrated Labour Market Biographies (SIAB) 1975–2017* (FDZ Datenreport, 02/2019). Nuremberg: Institute for Employment Research.

Antoni, M., Bela, D., & Vicari, B. (2019b). Validating earnings in the German National Educational Panel Study: Determinants of measurement accuracy of survey questions on earnings. *Methods, Data, Analyses, 13*(1), 59–90. https://di.org/10.12758/mda.2018.08

Baker, R., Shiels, C., Stevenson, K., Fraser, R., & Stone, M. (2000). What proportion of patients refuse consent to data collection from their records for research purposes? *British Journal of General Practice, 50*(457), 655–656. Avaibable at http://www.amstat.org/committees/ethics/linksdir/Jsm2005Bates.pdf

Bates, N., & Pascale, J. (2005). Development and testing of informed consent questions to link survey data with administrative records. In: *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 3786–3793.

Beste, J. (2011). *Selektivitätsprozesse bei der Verknüpfung von Befragungs- mit Prozessdaten: Record Linkage mit Daten des Panels „Arbeitsmarkt und soziale Sicherung“ und adminis-*

trativen Daten der Bundesagentur für Arbeit (FDZ Methodenreport 09/2011). Nuremberg: Institute for Employment Research.

Braun, D. (2016). *Dokumentation zur Erstellung einer Verlinkungsdatei von PIAAC/PIAAC-L Befragungsdaten mit administrativen Daten der Bundesagentur für Arbeit*. Nuremberg: Institute for Employment Research. Unpublished manuscript.

Calderwood, L., & Lessof, C. (2009). *Enhancing longitudinal surveys by linking to administrative data*. University of Essex, Methodology of Longitudinal Surveys. Available at https://www.iser.essex.ac.uk/files/survey/ulsc/methodological-research/mols-2006/scientific-social-programme/papers/Calderwood.pdf

Dahlhamer, J. M., & Cox, C. S. (2007, November). *Respondent consent to link survey data with administrative records: Results from a split-ballot field test with the 2007 National Health Interview Survey.* Paper presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.

Finkelstein, M. M. (2001). Do factors other than need determine utilization of physicians' services in Ontario? *CMAJ, 165*(5), 565–570.

Gauly, B., Daikeler, J., Gummer, T., & Rammstedt, B. (2019). What's my wage again? Comparing survey and administrative data to validate earnings measures. *International Journal of Social Research Methodology, 23*(2), 125–228. https://doi.org/10.1080/13645579.2019.1657691

GESIS – Leibniz Institute for the Social Sciences & Institute for Employment Research (IAB). (2017). *PIAAC-L-ADIAB* [Data file; Version 2]. Unpublished data. Nuremberg: Institute for Employment Research.

Haider, S., & Solon, G. (2000). *Nonrandom selection in the HRS social security earnings sample* (Working Paper No. 00-01). RAND Labor and Population Program.

Hanushek, E., Schwerdt, G., Woessmann, L., & Wiederhold, S. (2015). Returns to skills around the world: Evidence from PIAAC. *European Economic Review, 73*, 103–130. https://doi.org/10.1016/j.euroecorev.2014.10.006

Hartmann, J., & Krug, G. (2009). Verknüpfung von personenbezogenen Prozess-und Befragungsdaten–Selektivität durch fehlende Zustimmung der Befragten? *Zeitschrift für ArbeitsmarktForschung, 42*(2), 121–139. https://doi.org/10.1007/s12651-009-0013-y

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, 169*(4), 701–722. https://doi.org/10.1111/j.1467-985X.2006.00417.x

Knies, G., & Burton, J. (2014). Analysis of four studies in a comparative framework reveals: health linkage consent rates on British cohort studies higher than on UK household panel surveys. *BMC Medical Research Methodology, 14*(1), 125–136. https://doi.org/10.1186/1471-2288-14-125

Knies, G., Burton, J., & Sala, E. (2012). Consenting to health record linkage: Evidence from a multi-purpose longitudinal survey of a general population. *BMC Health Services Research, 12*(1), 52–57. https://doi.org/10.1186/1472-6963-12-52

Korbmacher, J. M., & Schroeder, M. (2013). Consent when linking survey data with administrative records: The role of the interviewer. *Survey Research Methods, 7*(2), 115–131. https://doi.org/10.18148/srm/2013.v7i2.5067

Martin, S., Lechner, C., Kleinert, C., & Rammstedt, B. (2020). Literacy skills predict probability of refusal in follow-up wave: Evidence from two longitudinal assessment surveys. *International Journal of Social Research Methodology*. Advance online publication.

OECD. (2013). *OECD skills outlook 2013: First results from the Survey of Adult Skills*. Paris: OECD Publishing.

Pascale, J. (2011). Requesting consent to link survey data to administrative records: Results from a split-ballot experiment in the Survey of Health Insurance and Program Participation (SHIPP). *Survey Methodology, 03*.

Rammstedt, B., Martin, S., Zabal, A., Carstensen, C., & Schupp, J. (2017). The PIAAC longitudinal study in Germany: Rationale and design. *Large-Scale Assessments in Education, 5*(1), 4. https://doi.org/10.1186/s40536-017-0040-z

Sakshaug, J. W. (2018). Methods of linking survey data to official records. In D. L. Vannette & J. A. Krosnik (Eds.), *The Palgrave handbook of survey research* (pp. 257–261). Cham: Palgrave Macmillan.

Sakshaug, J. W., & Antoni, M. (2017). Errors in linking survey and administrative data. In: P. P. Biemer et al. (Eds.), *Total Survey Error in Practice* (pp. 557–573), Hoboken: John Wiley & Sons.

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. https://doi.org/10.18148/srm/2012.v6i2.509

Sakshaug, J. W., & Kreuter, F. (2014). The effect of benefit wording on consent to link survey and administrative records in a web survey. *Public Opinion Quarterly, 78*(1), 166–176. https://doi.org/10.1093/poq/nfu001

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. https://doi.org/10.1177/0049124112460381

Sakshaug, J. W., Tutz, V., & Kreuter, F. (2013). Placement, wording, and interviewers: Identifying correlates of consent to link survey and administrative data. *Survey Research Methods, 7*(2), 133–144. https://doi.org/10.18148/srm/2013.v7i2.5395

Sala, E., Burton, J., & Knies, G. (2010). Correlates of obtaining informed consent to data linkage: Respondent, interview, and interviewer characteristics. *Sociological Methods & Research, 41*(3), 414–439. https://doi.org/10.1177/0049124112457330

Steinacker, G., & Wolfert, S. (2017). *Durchführung der 2. Erhebungswelle von PIAAC-L (Kooperative längsschnittliche Weiterverfolgung der PIAAC-Studie in Deutschland): Feldbericht zur Erhebung 2015* (GESIS Papers 2017|04). Mannheim: GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.21241/ssoar.50488

Warnke, A. J. (2017). *An investigation of record linkage refusal and its implications for empirical research* (ZEW Discussion Paper 17-031). Mannheim: ZEW - Leibniz Centre for European Economic Research.

Zabal, A., Martin, S., & Rammstedt, B. (2017). *PIAAC-L data collection 2015: Technical report* (GESIS Papers 2017|29). Mannheim: GESIS – Leibniz-Institute for the Social Sciences. https://doi.org/10.21241/ssoar.55155