

ORIGINAL ARTICLE

Open Access



A tale of two data sets: comparing German administrative and survey data using wage inequality as an example

Heiko Stüber^{1,4*} , Markus M. Grabka²  and Daniel D. Schnitzlein^{3,4,5} 

Abstract

The IAB's Sample of Integrated Labour Market Biographies (SIAB) and the Socio-Economic Panel (SOEP) are the two data sets most commonly used to analyze wage inequality in Germany. While the SIAB is based on administrative reports by employers to the social security system, the SOEP is a survey data set in which respondents self-report their wages. Both data sources have their specific advantages and disadvantages. The objective of this study is to describe and compare the evolution of wage inequality for these two types of data. For this purpose, different sample restrictions are applied. The comparison without any harmonization of the data shows different levels and trends. When the information is largely harmonized, comparable trends and similar levels emerge.

Keywords Wage inequality, SIAB, SOEP, Administrative data, Survey data, Germany

JEL Classification D31, D33, J31

1 Introduction

For years, the Socio-Economic Panel (SOEP) has been the standard data set for analyzing income and wage¹ inequality in Germany (e.g., Steiner and Wagner 1998; Biewen 2000; Gernandt and Pfeiffer 2007; Biewen and Juhasz 2012; Sommerfeld 2013). However, research based on administrative data has recently gained importance, especially in labor economics (e.g., Card et al. 2013;

¹ We use the terms wage and salary synonymously, since neither the SOEP nor the SIAB specify whether a person receives a fixed amount per pay period (salary) or is paid by the hour (wage).

*Correspondence:

Heiko Stüber
Heiko.Stueber@iab.de

¹ Institut für Arbeitsmarkt- und Berufsforschung (IAB), Regensburger Straße 10, 90478 Nürnberg, Germany

² DIW Berlin/SOEP, Berlin, Germany

³ Leibniz Universität Hannover, Hannover, Germany

⁴ IZA, Bonn, Germany

⁵ Inside Statistics, Passau, Germany



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Fitzenberger and de Lazzar (2022). Large sample sizes and the (expected) accuracy of the information included are considered advantages of this type of data. However, these data are designed for administrative purposes and may not contain all the information needed for research. At the same time, most large household surveys have begun to increase their samples, in part to include specific subpopulations.² Compared to administrative data, survey data contain much more information and cover more topics, but at the cost of much smaller sample sizes and perhaps less precision in certain quantitative variables.³

Despite these developments, there is a paucity of work comparing survey and administrative data and examining whether these data sets produce comparable results.⁴ In this article, we fill this gap for Germany by comparing trends and levels of wage inequality based on the SOEP (e.g., Goebel et al. 2019) with results based on the Sample of Integrated Labour Market Biographies (SIAB; e.g., Frodermann et al. 2021).

Both data sets are widely used in empirical labor research and have been used extensively in past research on wage inequality. Based on the SOEP, Biewen and Juhász (2012), e.g., analyze determinants of the rise in income inequality in the early 2000s and show that most of the increase is due to rising labor income inequality. Burauel et al. (2020) analyze the impact of the introduction of the minimum wage on wage inequality and find that this reform leads to a reduction in inequality.⁵ In addition, the SOEP forms the basis of many social reporting statistics in Germany. Several German research institutes publish inequality statistics based on the SOEP at regular intervals (e.g., Stockhausen and Calderón 2020; Grabka 2021), which are incorporated into policy and governmental reports (e.g., Bundesregierung 2016, 2021; OECD 2018).

Most studies on wage inequality in Germany based on administrative data have been conducted with data sets from the Research Data Center (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB). Dustmann et al. (2009), e.g., use the IAB Employment Samples (IABS) to examine the West German wage structure. They show that wage inequality in West Germany has

increased between 1975 and 2004. Klein et al. (2013) use the IAB's Linked Employer-Employee Data (LIAB) to analyze the impact of export activity on wage inequality within and across skill groups. Card et al. (2013) use the Integrated Employment Biographies (IEB) and find that increasing heterogeneity at the establishment level and increasing assertiveness in assigning workers⁶ to establishments mainly explain increasing wage inequality. Dustmann et al. (2014) use the SIAB and show that real wage growth is negative for the lower end of the wage distribution. Fitzenberger and de Lazzar (2022) also use SIAB data to examine whether changes in the selection into full-time employment among German men are a cause of the rise in wage inequality since the mid-1990s.⁷

Recent results based on the SOEP show that (hourly) wage inequality—measured as the ratio between the 90th and 10th percentiles—increased substantially in the early 2000s, remained relatively stable between 2006 and 2014, then declined until 2016, and continued to move horizontal until 2019, bringing the German labor market back to the inequality level of the early 2000s (Grabka 2021). Using the SIAB data through 2014, Fitzenberger and Seidlitz (2020) confirm the same upward trend in the early 2000s for full-time worker wage inequality, measured as the ratio between the 80th and 20th percentiles. However, in their sample, inequality continued to increase for men and decrease for women through 2014. The disparate results based on different data sets underscore the need for a thorough comparative analysis.⁸

The SOEP data have the advantage that information on individuals' wages is embedded in the full set of information available in a large household survey. This allows researchers not only to calculate individual earnings or individual wages, but also to include the household level or to distinguish between measures of income before and after transfers and taxes. Administrative data are usually produced by government agencies in the course of implementing certain rules, regulations, and laws. Statistics are a byproduct of these activities. On the one hand, administrative data can be considered very accurate in terms of the information required for the administrative procedure from which the data originate, e.g., as there are legal sanctions in case of misreporting. On the other hand, some additional information collected as part of the administrative procedure, such as educational status or occupation, may be considered less reliable.

² For example, the SOEP included families with many children in 2010, an oversampling of high-income households in 2002, migrants in 2013 and 2015, and refugees since 2016 (Britzke and Schupp 2019).

³ See, e.g., the SOEP Wave Report 2018 (Britzke and Schupp, eds 2019), the technical report for the 11th wave of the UK Household Longitudinal Study (Carpente 2021), or the PSID Main Interview User Manual for the 2021 release (Beaule et al. 2021).

⁴ In Appendix A, we briefly review some articles that compare administrative and survey data.

⁵ Further examples of papers using the SOEP are Bartolucci (2013), Selezneva and Van Kerm (2016), Grabka and Schröder (2018), Tyrowicz et al. (2018), Caliendo and Wittbrodt (2022), and Beckmannshagen and Schröder (2022).

⁶ We use the term worker to describe both, blue collar (workers) and white collar (employees).

⁷ The finding of an increase in wage inequality since the early 1990s is also confirmed by Möller (2016). Further examples of papers using the SIAB, or its predecessors—the IABS and the IEBS—are, e.g., Kopczuk et al. (2010) and Felbermayr et al. (2014).

⁸ Another data source for the analysis of wage inequality is the German Structure of Earnings Survey (VSE), see Appendix C.

There is no doubt that the set of potential control variables in the SIAB is much smaller compared to the SOEP. However, because survey costs are substantial, the coverage of surveys is usually limited (about 30,000 individuals in the last SOEP wave), while the administrative data cover almost the entire population of individuals participating in the labor market. The IEB covers, *inter alia*, all workers subject to social insurance contributions in Germany since 1975. The SIAB is a two-percent random sample of the IEB. This restriction is due to data protection purposes. Nevertheless, the SIAB covers labor market related information of around 1.78 million persons.

Administrative data have the advantage that the characteristics essential to the underlying administrative process are usually of high quality and therefore have a low number of missings and a low measurement error. However, administrative data can be affected by, e.g., processing errors (e.g., duplicate reports, data entry error) (Kapteyn and Ypma 2007; Groen 2012; Lindner and Andreasch 2014) and/or coverage errors (e.g., administrative data lack information on the shadow economy). Whereas, survey data can be subject to various measurement errors (e.g., Bound et al. 2001). Basically, a distinction can be made between sampling errors and non-sampling errors. Sampling errors occur when only a non-representative subset of the population is actually surveyed. Non-sampling errors include coverage errors, framing errors, response/non-response errors, measurement errors, and processing errors (e.g., de Leeuw et al. 2008). In summary, both survey data and administrative data are subject to different types of measurement errors and therefore neither data source should be considered the only true one.⁹ For Germany, this has been shown, e.g., by Oberski et al. (2017). They “found for official administrative data obtained from the German Federal Employment Agency that the reliability of both survey and administrative data was far from perfect.” (Oberski et al. 2017, p. 1486). Any comparison of survey and administrative data additionally faces differences in the definitions of the unit of analysis, divergent reference periods, or even censoring.

As noted above, both the SOEP and the SIAB have been used to analyze wage inequality in the past, using different analytical methods that take advantage of the unique characteristics of the two data sets. To our knowledge, however, no attempt has ever been made to derive comparable inequality estimates from these data sources.

In this article, we fill this gap by bringing in estimates of wage inequality trends for Germany based on (1) samples that exploit the strengths of each data set, (2) samples that are as comparable in composition as possible, and comparing our findings from these approaches.

2 Data

In the following, we briefly introduce the two data sets we use—SOEP and SIAB. We then define different subsamples of the two data sets that we use for the analyses. More detailed information on the data sets, data preparation, and sub-sampling can be found in Appendix B.

2.1 German Socio-Economic Panel (SOEP)

The SOEP is a representative household survey. It has been conducted annually since 1984 and in the last wave covered more than 20,000 households with more than 30,000 individuals (excluding children). The SOEP covers a wide range of topics, including detailed information on earnings and wages at both the individual and household level (Goebel et al. 2019). Due to its nature as a household survey, the SOEP also covers civil servants and self-employed persons as well as marginally employed persons in addition to employees. Wages are surveyed for the previous month, including any overtime pay, and for the previous year, broken down into one-time payments and severance payments. Wages from secondary employment are queried separately. It should be noted that the way in which information for secondary employment is collected has changed fundamentally in 2017. Since then, dependent employment can be clearly distinguished from an honorary and self-employment secondary employment. Previously, this distinction was not possible. In case of item non-response, the main imputation method is the so-called row-and-column imputation developed by Little and Su (1989). When longitudinal information for the imputation process is missing, OLS regressions are applied (see Frick and Grabka 2005).¹⁰

We use the SOEP-Core v37 for our analyses. For more details and a brief description of the data preparation steps performed see Appendix B1 and Schröder et al. (2020).

2.2 Sample of integrated labour market biographies (SIAB)

The SIAB is a two percent random sample drawn from the IEB of the IAB. The IEB is an administrative data set with information from various data sources. It includes,

⁹ A brief overview of articles examining the error structure in earnings data is provided in Appendix A.

¹⁰ The SOEP also asks for both agreed and actual working hours, so that hourly wages can be calculated in addition to monthly or annual wages. Since hourly wages cannot be calculated in the SIAB due to the lack of working time data, this wage concept is not presented here.

among others, all workers subject to social insurance contributions and all marginally employed individuals in Germany.¹¹

The employment information in the IEB comes from the integrated notification procedure for health, pension and unemployment insurance (Bender et al. 1996). As a part of this procedure, employers are required to submit notifications on all their employees subject to social security insurance to the relevant social insurance institutions at least once a year. Civil servants and self-employed individuals are not subject to social security insurance and are therefore not included in the data set.

Workers can be identified by an artificial individual ID and tracked over years. The data is organized by employment spells. The maximum length of a spell is one calendar year. For each employment spell, the beginning and end of employment on a daily basis and the average gross daily wage are known, among other things.

The wage data in the SIAB is very reliable. The information is used, e.g., to calculate retirement pensions and unemployment insurance benefits. However, the wage data is only relevant up to the social security contribution assessment ceiling. For this reason, the wage information in the process data is top-coded, so that we only observe wages up to the contribution assessment ceiling. Therefore, following Stüber et al. (2023), we impute top-coded wages using a 2-step imputation procedure similar to Dustmann et al. (2009) and Card et al. (2013).¹²

We use the SIAB 7519 for our analyses. For more details and a brief description of the data preparation steps performed see Appendix B2 and Frodermann et al. (2021).

2.3 Defining sub-samples for the analyses

To highlight the strengths and weaknesses of the SOEP and SIAB data and to compare the two data sources, we define different sub-samples of the data sets in Sect. 2.3.1, which we then use for the analyses. The rationale behind the sub-samples hereby is the following: the samples described in the next section play to the individual strengths of the two underlying data sets. Researchers studying wage inequality based solely on the SOEP or the SIAB will most likely end up using one of these sub-samples. These thus represent the standard use cases of the two data sets. In contrast, the samples described in

Sect. 2.3.2 follow the intention of making the two data sets as comparable as possible.

The following basic restrictions apply to all samples we draw for our analyses in this article:

- The analyses are performed for the last two decades covered by our data, i.e., 2000 to 2019.
- We consider only workers between the ages of 18 and 65.
- Wages from self-employment or wages paid as part of an internship or, e.g., a voluntary social year are not taken into account.¹³

2.3.1 Exploiting the strengths of the SOEP or SIAB

For each data set, we create two sub-samples to take advantage of the respective strengths of the data set. The basic restrictions listed in Sect. 2.3 apply here as well, of course, but are not repeated in the sketches for the individual data sets.

2.3.1.1 SOEP-Pure 1 (monthly wage) Workers considered: All workers subject to social insurance contributions and civil servants.

Sampling: All respondents who reported either positive wages in the last month or positive wages in a second job when wages from the main job were either zero or missing.

Wage for the calculation of wage inequality: Nominal gross monthly wage of person's main job or, if not available, second job in the month prior to the interview.

2.3.1.2 SOEP-Pure 2 (annual wage) Workers considered: All workers subject to social insurance contributions and civil servants.

Sampling: All respondents who reported positive individual wages for the last calendar year. Since the latest survey year available in the SOEP v37 is 2020, the latest available observation for wages in the last calendar year is for 2019.

Wage for the calculation of wage inequality: Nominal gross annual wages (sum of all labor wages excluding income from self-employment) of an individual in year preceding the survey year, including any one-time payments.

¹¹ Information on marginal part-time employed persons has been included in the IEB since 1999.

¹² We impute right-censored wages mainly to display the mean wages in the descriptive statistics in Appendix B2. Imputation does not play a role in the calculation of our inequality measures. See also Footnote 18.

¹³ A precise delineation of the groups of people included in the analyses can be found in the Appendix B1 for the SOEP and Appendix B2 for the SIAB.

2.3.1.3 SIAB-Pure 1 (average daily wage) Workers considered: All workers subject to social insurance contributions.

Sampling: Consideration of all employment spells.

Wage for the calculation of wage inequality: Average nominal gross daily wage of all jobs held by a person during the year [weighted by the duration of the employment spell; following Dustmann et al. (2009)].

2.3.1.4 SIAB-Pure 2 (daily wage as of June 30) Workers considered: All workers subject to social insurance contributions.

Sampling: Consideration of all employment spells as of June 30 of each year.

Wage for the calculation of wage inequality: Sum of all nominal gross daily wages of an individual as of June 30 of each year.¹⁴

2.3.2 Generating comparable sub-samples

The samples in the last section correspond to what researchers might use if they were analyzing only one of the two data sources. In contrast, in this section, we create sub-samples that are as comparable as possible between the two data sources.

First, we add another basic restriction to all comparable sub-samples: to ensure that the individuals considered are comparable, we exclude civil servants from the SOEP—or, more precisely, wages from civil servant employment relationships—in the comparable samples. The other basic restrictions listed in Sect. 2.3 also apply here, but are not repeated.

2.3.3 SOEP-SIAB-comparable 1: Gross monthly wage

The SOEP asks for wage information for the month preceding the survey month. Interviews are conducted in almost all months, but the vast majority of surveys take place from February to May.¹⁵ Therefore, we replicate this survey structure in the SIAB. In each year, we randomly assign an interview month to each person in the SIAB. We ensure that the random assignment of the month results in the same distribution as for individuals aged 18 to 65 in this year's SOEP. If a person in the SIAB is not employed in the month preceding the assigned interview month, this individual is not included in the analysis.

Since the SIAB does not contain the monthly wage, we calculate the gross monthly wage to the day as the gross

daily wage (henceforth daily wage) multiplied by the number of days in the employment spell in the respective month. The daily wage in the SIAB also includes bonus payments, etc. (e.g., Christmas bonus) received by the employee in the duration of the spell. This information is not included in the monthly wage in the SOEP. However, the information on bonus payments is collected retrospectively for the last survey year. To mimic the wage measure in the SIAB, we use this information as a proxy and add 1/12 of the one-time payments collected retrospectively in the SOEP to the monthly wage.

If workers are employed by more than one employer, only the information for the main employment, i.e., the job with the highest monthly wage, is used. This leads us to the following comparable samples:

2.3.3.1 SOEP-comparable 1 Workers considered: All workers subject to social insurance contributions.

Sampling: All respondents who reported either positive wages in the last month or positive wages in a second job when wages from main job were either zero or missing.

Wage for the calculation of wage inequality: Nominal gross monthly wage of person's main job or, if not available, second job in the month prior to the interview; adjusted proportionally for bonuses, etc.

2.3.3.2 SIAB-comparable 1 Workers considered: All workers subject to social insurance contributions.

Sampling: Annual random assignment of a survey month; distribution of survey months is predetermined by the actual distribution of survey months in SOEP.

Wage for the calculation of wage inequality: Nominal gross monthly wage (calculated to the day) of person's main job in the month prior to the assigned interview.

2.3.4 SOEP-SIAB-comparable 2: Gross annual wage

Since the SIAB does not include annual wages, we calculate gross annual wages by adding up all wages (employment spell duration multiplied by the gross daily wage) for each person in each year.¹⁶

2.3.4.1 SOEP-comparable 2 Workers considered: All workers subject to social insurance contributions.

Sampling: All respondents who reported positive individual wages for the last calendar year.

Wage for the calculation of wage inequality: Nominal gross annual wage of an individual in year preceding the survey year, including any one-time payments.

¹⁴ Aggregated data sets, such as the Establishment History Panel (BHP), are generated at the FDZ as of June 30. Therefore, on this cut-off date, individual-level data can be enriched with aggregated establishment data. In addition, June 30 is frequently used as the cut-off date in official statistics in Germany.

¹⁵ A small fraction of the SOEP surveys is conducted in January and therefore contains the monthly wage for December of the previous year. We have chosen to report all wages for a given survey year.

¹⁶ We apply the distribution of the interview months of the SOEP to the SIAB, which is correct for the case that we consider monthly wages. If we consider annual wages, we would actually have to use the distribution of year $t+1$. In order to be able to consider the same period throughout the article, we refrain from doing so and use the distribution of year t .

2.3.4.2 *SIAB-comparable 2* Workers considered: All workers subject to social insurance contributions.

Sampling: Consideration of all employments.

Wage for the calculation of wage inequality: Sum of the nominal gross payrolls of all employment relationships of a person during the year.

3 The development of wage inequality in Germany

3.1 Inequality measures

To analyze income inequality and illustrate the difference between survey and administrative data, we focus on wage percentiles and wage percentile ratios. Percentile ratios are widely used, e.g., by the OECD, and are an intuitive way of representing income inequality.¹⁷

Percentile ratios indicate the ratio of the wages of two individuals who are in different positions in a given distribution. For example, the P90/P10 ratio compares the wage at the 90th percentile with that at the 10th percentile. If the P90/P10 ratio has increased over a period, this indicates that inequality between the top and the bottom tails of the wage distribution has increased. The disadvantage of the P90/P10 ratio is that wage trends above the 90th percentile and below the 10th percentile are not considered. Somewhere above the 90th percentile, however, is the wage threshold above which employers no longer report detailed wage information, so that inaccuracies can occur even when wage information is estimated. Thus, using the P90/P10 ratio circumvents the uncertainty above the wage threshold.¹⁸ For the purposes of this article, we consider the following three percentile ratios: P90/P10, P90/P50, and P50/P10.

¹⁷ An alternative way to measure inequality would be to calculate, e.g. Gini indices for the individual data sets. However, while the Gini index is only an aggregate measure, percentile ratios provide information on the shape of the underlying distribution.

¹⁸ The 90th percentiles are not affected by the imputation of wages. We impute right-censored wages so that we can display the mean wages in the descriptive statistics in Appendix B2. The 90th percentiles however would be affected by imputed wages if we were to report our results differentiated, e.g., by gender and education. As Stüber et al. (2023, p. 6) note: The “wage information in the process data is top-coded, and hence we only observe wages up to the social security contribution ceiling. While this feature only affects approximately 5.2% of all employment spells for workers between 1975 and 2019, the proportion of censored observations within certain subgroups is substantial. For instance, nearly 44% of the spells of regularly employed male workers with a degree from a university or university of applied science are affected by top-coding.” However, looking at the SIAB-Pure 2 sample, e.g., and not considering wages below the marginal earnings threshold, no year has more than 90% of wages imputed above the 90th percentile.

3.2 Using the strengths of survey and administrative data

First, we measure wage inequality in both data sets using the two most common wage measures for the respective data sets: monthly wage (SOEP-Pure 1) and annual wage (SOEP-Pure 2) for the SOEP and average daily wage (SIAB-Pure 1) and daily wage on June 30 (SIAB-Pure 2) for the SIAB. These are the native measures for the respective data set, i.e., monthly wages as well as annual wages are collected directly in the SOEP questionnaire and wage information are provided as average daily wages in the SIAB. In the following section, we then measure wage inequality using comparable sub-samples of the two data sets.

The results based on monthly and annual wages in the SOEP are shown in Fig. 1. Results based on average daily wage and daily wage at June 30 in the SIAB are shown in Fig. 2.

Looking at the trajectories of the percentile ratios in the two data sets, it is immediately apparent that the P90/P10 value is lower in the SOEP than in the SIAB. The trend over the years also shows clear differences.

In the SOEP, the monthly P90/P10 percentile ratio is relatively stable until 2016 and then declines (see Fig. 1A).¹⁹ In contrast, the annual P90/P10 percentile ratio shows a weak inverted U-shape: it rises until 2011 and then falls again starting around 2013 (see Fig. 1B).

In the SIAB, the pattern is much more rigid. However, a very weak inverted U-shape can be seen for the P90/P10 percentile ratio of average daily wages (see Fig. 2A). Comparing the P50/P10 percentile ratios, a similar pattern emerges as for the P90/P10 ratios. In contrast, the P90/P50 percentile ratios are quite similar in all four samples: a nearly constant pattern with a ratio of about two.

The fact that the P90/P50 percentile ratios are quite similar in all four samples, but there are major differences in both the level and trend of the P90/P10 and P50/P10 percentile ratios, suggests that there are differences in the data sets at the lower end of the wage distribution. The lower P90/P10 and somewhat lower P50/P10 percentile ratios in the SOEP indicate that the P10 value in the SOEP must be higher than that in the SIAB.

Of course, this difference may be because the population covered by the two data sets is not fully comparable. For example, the SOEP includes wages of civil servants, which are not included in the SIAB data. In addition, the SOEP surveys do not take place evenly throughout the year. Therefore, wages in the SOEP could be influenced by seasonal effects.

To examine whether the differences described are due to these factors, we consider the comparable sub-samples

¹⁹ The trend of a decline in monthly wage inequality even began several years earlier when using gross hourly wages (cf. Grabka 2022).

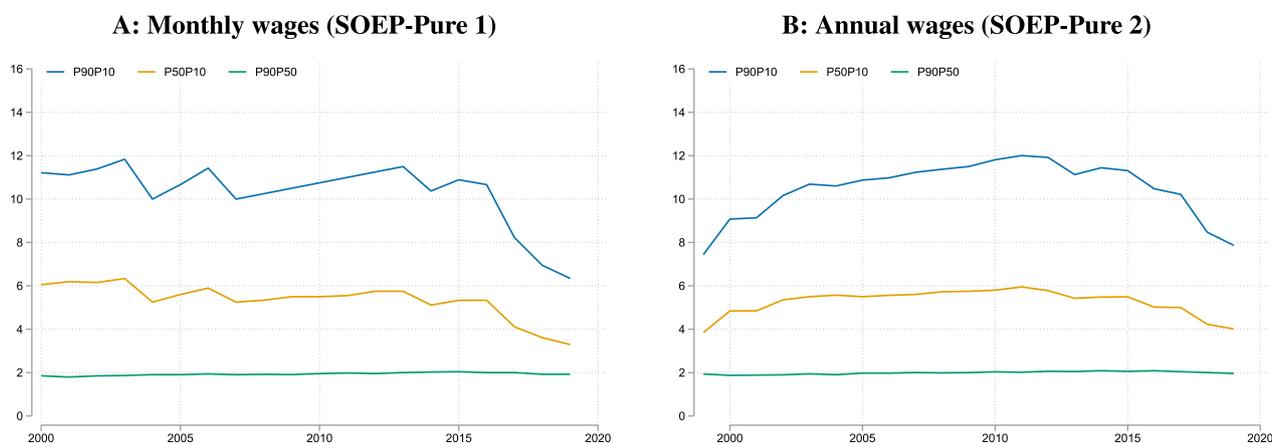


Fig. 1 Development of wage percentiles ratios in the SOEP 2000–2019. Source: own calculations, SOEP-Core.v37.EU

in the following section. However, another strategy for uncovering potential factors to explain the observed differences between SOEP and SIAB may be to use other data sources with wage information. Therefore, in further analyses, we use the Federal Statistical Office's Structure of Earnings Survey (VSE, Verdienststrukturerhebung) in Appendix C to gain further insights.

3.3 Using comparable sub-samples of survey and administrative data

In the following, we consider the four comparable sub-samples presented in Sect. 2.3.2. Figures 3 and 4 show the comparable samples using monthly and annual wages, respectively.

The percentile trajectories of monthly wages (Fig. 3A) look relatively similar in the two data sets. This is especially true for the 50th percentile. However, the P10 of the SOEP is somewhat higher than that of the SIAB, especially from 2017 onward. On the other hand, the P90 of the SIAB is consistently higher than that of the SOEP.

The percentile trajectories of annual wages (Fig. 4A) in the two data sets also look quite similar. However, here the P10 of the SOEP is significantly higher than the P10 of the SIAB, and one also sees clear differences in the two P50 percentiles.

The differences—in both monthly and annual data—translate into striking differences when percentile ratios are considered (see Figs. 3B and 4B). Here we can see that the clear differences in percentile ratios—with the exception of P50/P10—still exist.

These results suggest that wages at the lower (and upper) ends of the wage distribution are captured differently. It could be that low wages are underrepresented in the SOEP. This could be due to sampling, but it could also be due to respondents not reporting very low wages or simply forgetting about them. It could also be because

information on second jobs has been collected more precisely in the SOEP since 2017. Another explanation could be that respondents with several marginal part-time jobs are very likely to add them together and thus report higher wages as captured in the SIAB. This last aspect might be particularly relevant, given that the number of workers with more than one job more than doubled between 1999 and 2019.²⁰

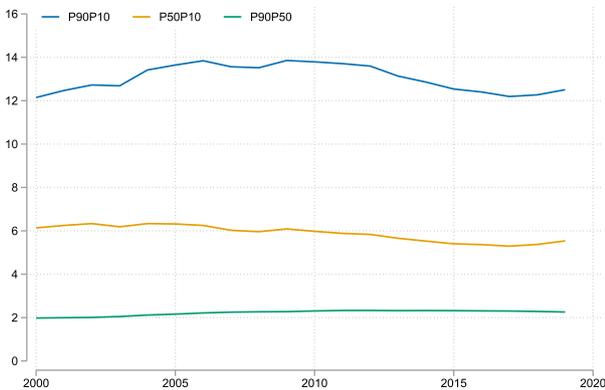
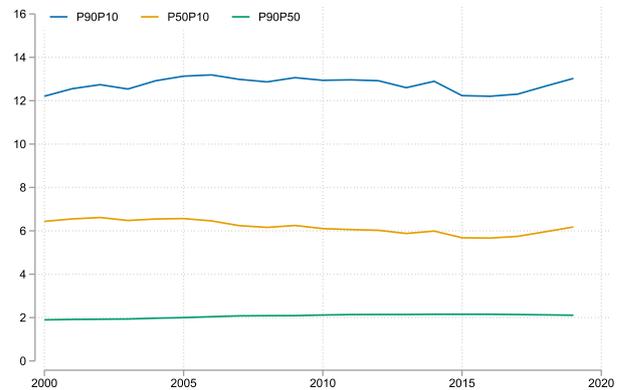
In the period under consideration, low-income wages are included in the administrative data (SIAB) because since 1999 jobs with wages below the marginal earnings threshold are subject to a lump-sum social security contribution payable by the employer. For this reason, marginal wages should be recorded relatively reliably in the SIAB. Here, too, however, it cannot be ruled out that, e.g., spells with low wages may occur due to subsequent wage declarations that do not represent actual employment. Although these are administrative data, we observe some employment spells with unrealistically low daily wages (e.g., ≤ 1 euro)²¹. This could lead to an overestimation of marginal part-time employment in the SIAB.

To further analyze this, we decided to add another restriction to our comparable sub-samples below by limiting the analyses to wages above the marginal earnings threshold.²²

²⁰ Source: <https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Qualitaet-Arbeit/Dimension-3/zweitjobl.htm>; last accessed: Oct. 23, 2022.

²¹ Less than 0.15% of workers have a daily wage of ≤ 1 euro (as of June 30); the 1% percentile is 3.12 euros and the 5% percentile is 8.22 euros.

²² To better assess the difference between SIAB and SOEP highlighted in this section, we additionally rely on the VSE in Appendix C.

A: Average daily wage over full year (SIAB-Pure 1)**B: Average daily wage at June 30 (SIAB-Pure 2)****Fig. 2** Development of wage percentiles ratios in the SIAB 2000–2019. Source: own calculations, SIAB 7519

3.4 Using the marginal earnings threshold as a lower wage floor

Using the marginal earnings threshold²³ as a wage floor, the results are no longer representative for the entire German labor force, but only for workers with wages above the marginal earnings threshold. Nevertheless, the percentile ratios can be used to examine the development of wage inequality in this group and to highlight differences in the two data sets.

The results of this exercise are striking. Looking only at wages above the marginal earnings threshold, the trend is now almost identical for all three percentiles considered (see Figs. 5A and 6A). Only in 2016 does the P90 in the SOEP decrease slightly, but the development in the following years is again in line with that of the SIAB.

The trajectories of the percentile ratios are now also quite similar (see Figs. 5B and 6B). However, we still find a level difference in the P90/P10 ratio for annual wages (see Fig. 6B).²⁴

4 Conclusion

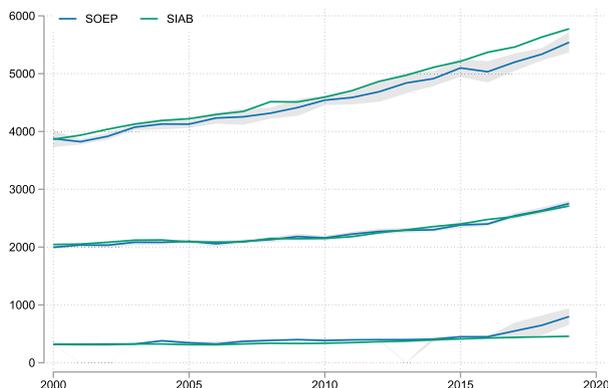
Two very different data sets are often used to describe wage inequality in Germany. On the one hand, the administrative data of the SIAB and, on the other hand, the information obtained from a population survey, the SOEP. Both data sources have their specific advantages and disadvantages. Using the full information from each data source yields comparable trends in wage inequality from 2000 to about 2015. Since then, the results diverge at first glance. After

harmonizing the analysis population, comparable trends and relatively similar levels emerge again for both data sources. However, the comparison of the two data sources also shows that there are systematic deviations between them in the area of very low wages. Various reasons may be responsible for this: On the survey side, it might be the case that respondents add up wages from several activities, so that the SOEP shows higher wage levels in the lower part of the distribution. In addition, it cannot be ruled out that money flows from the employer to the employee over and above the regular wage, e.g., to pay for short-term overtime in a mini-job, but not to exceed the official mini-job threshold. Accordingly, there are significantly more very small wage reports on the SIAB side. One of the reasons for this might be short-term registrations and cancellations of employment relationships, which are not considered relevant for respondents and therefore tend not to be reported in survey data. In addition, there is presumably a problem on the part of the SIAB, known from minimum wage research: if the minimum wage legislation is not complied with or the mini-job limit is exceeded, electronic payroll systems issue a warning message and wages or working hours might just be adjusted in the system in accordance to the legal situation (e.g., Bachmann et al. 2020, p. 26). This is likely to occur more frequently in the lower wage range and in short-term employment. Overall, however, the findings presented here show that both data sources are well suited to adequately describe wage inequality in Germany above the mini-job threshold. However, the observed differences at the bottom of the wage distribution require further analysis. The upcoming release of SOEP data linked to SIAB data by social security number (SOEP-ADIAB), planned for 2023, will provide a database helping to further investigate into these differences.

²³ The daily, monthly, and annual marginal earnings threshold are provided at http://doku.iab.de/fdz/Bemessungsgrenzen_de_en.xls; last accessed: Oct. 23, 2022.

²⁴ In addition to the marginal earnings threshold, we also calculated the respective percentiles for alternative (lower) earnings thresholds (e.g., 100 EUR and 200 EUR). However, the effect of results converging, was most pronounced at the marginal earnings threshold.

A: Development of percentiles P10, P50, and P90



B: Development of percentile ratios

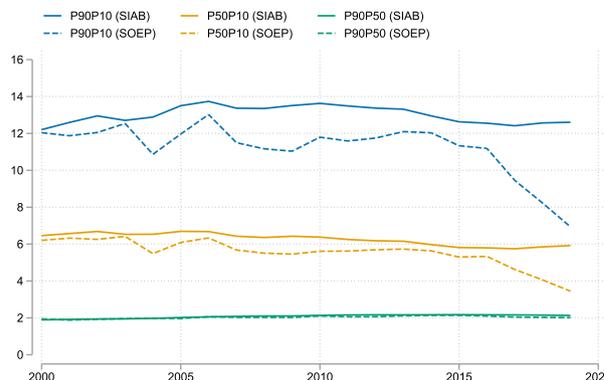


Fig. 3 Monthly wages in 2000–2019; sample: comparable sample; no lower limit. Source: own calculations, SOEP-Core.v37.EU & SIAB 7519. 95% confidence interval indicated by shaded areas (for percentiles only); 500 bootstrap replications

A: Development of percentiles P10, P50, and P90



B: Development of percentile ratios

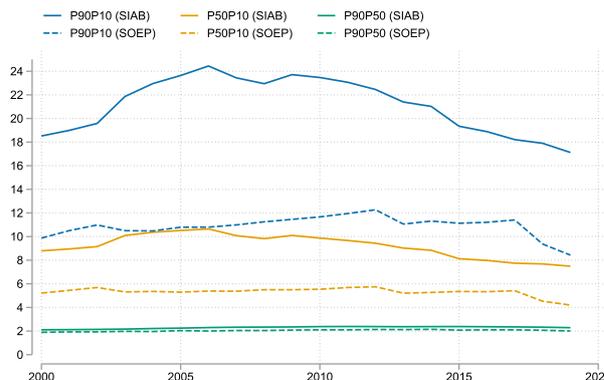


Fig. 4 Annual wages in 2000–2019; sample: comparable sample; no lower limit. Source: own calculations, SOEP-Core.v37.EU & SIAB 7519. 95% confidence interval indicated by shaded areas (for percentiles only); 500 bootstrap replications

Appendix A

Literature on the comparison of administrative and survey data and on the error structures in earnings data

This appendix provides a brief summary of the literature on comparing administrative and survey data in conjunction with a review of studies of error structures in earnings data.

There are few papers that compare survey and administrative data and examine whether these data sets provide comparable results. In particular, a number of papers compare information from administrative sources and from surveys on employment, earnings, and working hours. A first simple approach is to look at aggregates from the two sources in order to have a micro-macro

comparison. Using data for Finland, Kavonius and Törmälehto (2003) find that wages are almost identical in both data sources, while data on investment income and income from self-employment differ. The same strategy is followed by Törmälehto (2011), who reports that surveys in most countries covered by the Luxembourg Income Study (LIS) cover more than 90 percent of income. Pischke (1995) finds that information on earnings from administrative data and survey data does not differ significantly in either mean or variance. Bound and Krueger (1991) additionally consider sociodemographic variables in their comparisons and show that reliability ratios are higher for women than for men when they compare survey data from the Current Population Survey (CPS) with Social Security Administration (SSA) data. Information from the SSA is also used in the study by Bricker and Engelhardt (2008). They compare administrative information with that from the Health and Retirement Study

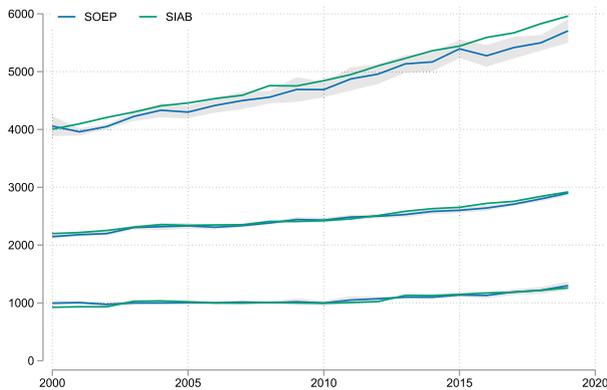
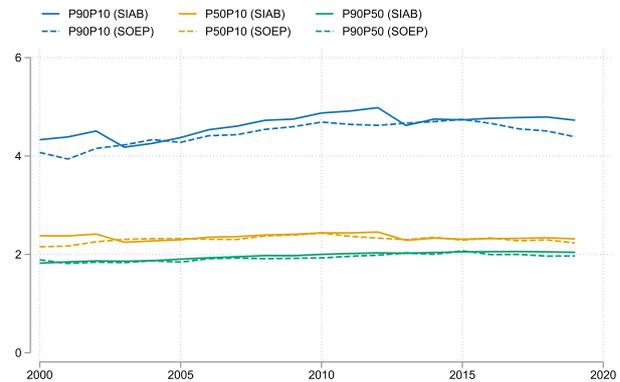
A: Development of percentiles P10, P50, and P90**B: Development of percentile ratios**

Fig. 5 Monthly wages in 2000–2019; sample: comparable sample; only wages above marginal part-time employment. Source: own calculations, SOEP-Core.v37.EU & SIAB 7519. 95% confidence interval indicated by shaded areas (for percentiles only); 500 bootstrap replications

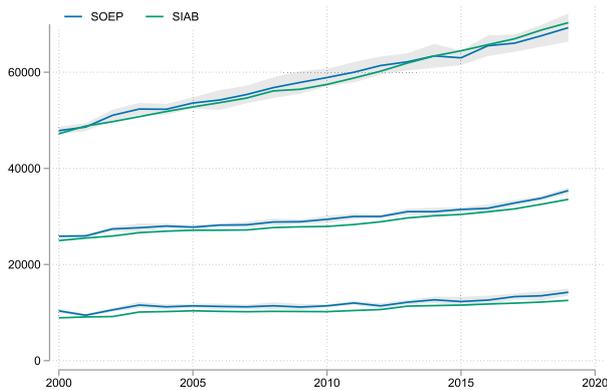
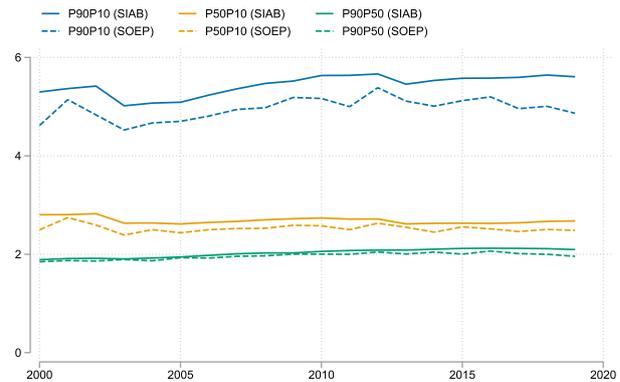
A: Development of percentiles P10, P50, and P90**B: Development of percentile ratios**

Fig. 6 Annual wages in 2000–2019; sample: comparable sample; only wages above marginal part-time employment ceiling. Source: own calculations, SOEP-Core.v37.EU & SIAB 7519. 95% confidence interval indicated by shaded areas (for percentiles only); 500 bootstrap replications

(HRS). By linking the two data sets, they identify measurement errors of about six percent in men's earnings and of approximately seven percent in women's earnings.

In addition to rather straightforward comparisons of point estimates, the shapes of distributions from different data sources are also examined. One example is Roemer (2002), who compared CPS data with data from the Survey of Income and Program Participation (SIPP) and finds that surveys accurately capture the underlying patterns of income distribution.

Another strand of papers examines the error structure in earnings data. Johansson-Tormod and Klevmarken (2022) show that both register error and survey error are negatively correlated with the true values (rich tend to underreport while poor tend to overreport). This result has also been shown in previous papers such as Bollinger (1998) or Pedace and Bates (2000). Bound et al. (1994) point out that measurement errors are not

necessarily due to misreporting in the survey data, but can be explained by errors in administrative data. Bakker (2011) argues that 24% of the variance in official Dutch hourly wages is due to random measurement error. Imputation is also a relevant source of measurement error in this context. For example, Coder (1992) argues that mean error is larger when earnings are partially or completely imputed. However, the quality of imputation also depends strongly on the imputation strategy (Blough et al. 2009).

One of the few examples for Germany is Lüthen et al. (2022). Based on a record-linkage they compare not only the amount of payments coming from public pensions between SOEP data and information from the pension register, but also a number of demographic variables. While, e.g., gender and age show an almost perfect overlap, pensions are overestimated by 12.8% at the mean in

survey data which decrease to less than 6% at the 90th percentile.

Another systematic comparison between different data sources were performed by Becker et al. (2003). Analyzing market as well as disposable household income coming from the German income and expenditure survey (Einkommens- und Verbrauchsstichprobe, EVS) from the federal statistical office and from the SOEP, they show that incomes from SOEP are significantly lower and additionally poverty rates are much higher than in the EVS. The authors argue that the results are partly driven by an insufficient coverage of foreigners in the EVS which yields to biased estimates in the latter data source.

In the context of the third Poverty and Wealth Report of the German government, Hauser et al. (2007) compared differences in disposable household income and poverty risk rates between SOEP, the German microcensus (Mikrozensus) and the German part of the EU Statistics on Income and Living Conditions (EU-SILC). They come to the conclusion that SOEP and microcensus show large overlaps, whereas the results of EU-SILC deviate from them, in some cases significantly. They explain this by the fact that the structure of the foreign population in EU-SILC does not match that of the other two data sources. In addition, EU-SILC massively underestimates the number of people in the labor force, while at the same time significantly overestimating the number of people with tertiary education.

Appendix B Data

B.1 German socio-economic panel study (SOEP)

We use the SOEP version SOEP-Core.v37.EU (DOI: <https://doi.org/10.5684/soep.core.v37eu>), which covers the survey years 1984–2020. As a first step, we restrict our sample to the active survey population (SOEP *netto-codes* 10–19) and survey years 2000–2020. Although our study focuses on the 2000–2019 time window, we need to include survey year 2020 because the information on annual wages in the SOEP corresponds to the year before the survey year. In all our results, we include individual-level survey weights that exclude the first occurrence of the individual in the data (*phrf1*) to ensure representativeness of the results on the national level. We then further restrict the sample to individuals aged 18–65 years.

Preparation of the base data set

For the analysis of monthly wages, we select dependent employees in their main job who report a non-missing wage in the current or preceding month of the survey (*pglabgro*). We excluded apprentices and trainees, those in partial retirement with working time of zero hours,

freelancers, and self-employed. Technically, this refers to SOEP *stib-codes* 10–15, 120–150, and 410–433. In addition, individuals who were exclusively engaged in dependent secondary employment were included back in the sample.

The same logic applies to the study of annual wages. We include all individuals for whom we have non-missing information on their annual earnings from the year preceding the survey year (*i11110-iself*). Again, we exclude apprentices and trainees, those in partial retirement with working time of zero hours, freelancers, and self-employed.²⁵

Preparation of the four data sets for our analyses

We create four data sets for the analyses. All of them use wage measures directly surveyed (monthly and annual wage). Two of them consist of all workers including civil servants and the two “comparable” data sets rely only on workers in employment relationships subject to social security contributions.

SOEP-Pure 1: monthly wage

The SOEP-Pure 1 data set represents the base data set described above focusing on monthly wages. For the final analysis, we only consider wages that are above the monthly marginal earnings threshold. Adding this restriction, pooled descriptive statistics for the SOEP-Pure 1 are reported in Table 1.

SOEP-Pure 2: annual wage

The SOEP-Pure 2 data set represents the base data set described above focusing on annual wages. For the final analysis, we only consider wages that are above the annual marginal earnings threshold. Adding this restriction, pooled descriptive statistics for the SOEP-Pure 2 are reported in Table 2.

Comparable sample 1: monthly wage

One important difference between SIAB and SOEP is that the SIAB does only cover employment relationships liable to social security contributions. This means that civil servants are not included in the data. Thus, for our comparable samples, we exclude civil servants from our SOEP samples. In addition, bonus payments (bonuses, holiday pay, Christmas bonus, ...) are included in the wage information in the SIAB. In the SOEP, this is usually not the case for the monthly wage observations. However, information on bonus payments are available in retrospective for the last calendar year. We use this

²⁵ It should be noted that this information technically refers to the current status of the surveyed individual and does not necessarily apply to employment relationships from the previous year.

information as proxy and add 1/12 of previous year's bonus payments (*i13ly*, *i14ly*, *ixmas*, *iholy*, *igray*, *iothy*, *itray*) to the reported monthly wage.²⁶

For the final analysis, we only consider wages that are above the monthly marginal earnings threshold. Summary statistics for this data set are provided in Table 3.

Comparable sample 2: annual wage

As for the Comparable Sample 1 we exclude civil servants from our SOEP sample. The annual wage measure needs no correction for bonus payments because these are natively included. For the final analysis, we only consider wages that are above the annual marginal earnings threshold. Summary statistics for this data set are provided in Table 4.

B.2 Sample of integrated labour market biographies (SIAB)

Only employment information from the weakly anonymized version of the SIAB 7519 (DOI: 10.5164/IAB.FDZ.2101.en.v1) is used for this article (cf. Frodermann et al. 2021). Information from the SIAB that derived from data sources other than the Employment History (Beschäftigten-Historik, BeH) is not used.²⁷

In preparing the SIAB, we are largely guided by Stüber et al. (2023). They present instructions for preparing the SIAB 7519 for scientific analyses. To this end, they provide a Stata do-file collection consisting of both do-files written by themselves and slightly adapted do-files from other authors. We made some minor adjustments and extensions to the Stata codes and omitted certain processing steps when they were not necessary for processing the BeH data or the variables needed for our analysis.

Our analysis is conducted for the years 2000 to 2019. However, we prepare the SIAB for 1999 to 2019 because we also conduct analyses where we replicate the survey pattern of the SOEP. That is, we assign an interview month to each worker and then look at the previous month's wage, as in the SOEP. Since some years in the SOEP surveys were also conducted in January, we need the December wage of the previous year for a few cases. For this reason, we also reprocess 1999 employment data from the SIAB so that we can assign 1999 December pay to January 2000 "respondents."

The remainder of this section briefly describes the SIAB data preparation process we used to generate our

base data set and how we used it to prepare the four data sets for our analyses.

Preparation of the base data set

We retain only the BeH spells and omit all periods for years prior to 1999. We also restrict the data to employment periods for workers aged 18 to 65 and drop all variables that are not essential for data compilation or analysis.

Since 2013, the number of reports with deregistration because of a "notification of a one-off wage" (deregistration reason 54; coded as `grund==154`) has increased sharply (cf. Frodermann et al. 2021, Sects. 5.5.1 and 5.5.12). It is likely that special payments that were reported with the annual spells before 2013 are now reported separately. Therefore, we add these one-off payments proportionally to spells of the same employee in the same year in the same company and daily wages is recalculated. After recalculating the daily wages, we delete all spells with the deregistration reason "notification of a one-off wage" and/or with a daily wage of zero. Then we calculate the tenure etc., undo the episode splitting and restrict our data as described in the following paragraph.

We only keep employment spells of the following person groups²⁸: Employees subject to social security with no special features [101], employees in partial retirement [103], marginal part-time employees in accordance with § 8 para. 1 No. 1 Social Code Book IV (SGB IV) [109], casual workers [118, 205], home workers [124], seamen [140], seamen in partial retirement [142], maritime pilots [143], employees in private households (reported via the "household cheque procedure") [201], artists and publicists subject to social security [203], and marginal part-time employees in private households (reported via the "household cheque procedure") [209]. Hence we drop, e.g., all trainees, apprentices, interns, family workers in agriculture, and persons completing a year of voluntary social or environmental work or Federal Voluntary Service. For an overview of person group codes see, e.g., Ganzer et al. (2021, pp. 120–121).

The wage information in the SIAB is rather reliable. It stems from the integrated notification procedure for health, pension and unemployment insurance. However, wage data are only recorded up to the social security contribution ceiling; higher wages are top-coded. Therefore we impute top-coded wages following Stüber et al. (2023) using a 2-step imputation procedure, similar to Dustmann et al. (2009) and Card et al. (2013).

²⁶ Again, it should be noted that bonus payments from the previous year can only be a proxy for payments in the current year.

²⁷ The IEB consist of all individuals, which are characterized by at least one of the following employment states: employed subject to social security or marginal part-time employed in Germany, benefit recipients according to the German Social Code III or II, individuals officially registered as job-seeking at the German Federal Employment Agency or individuals with (planned) participation in programs of active labor market policies.

²⁸ Person group code provided in square brackets.

Table 1 Summary statistics SOEP-Pure 1 sample

No. of observations	Mean wage	Median wage	Mean age
211,911	2737.77	2450.00	42.79

Source: own calculations, SOEP-Core.v37.EU, years 2000–2019

Wage: nominal monthly wage the main occupation in euro, mean over (median of) all observations of the sample. Age: mean age in years

Only wages above the marginal earnings threshold are taken into account

Right-censored wages are imputed separately for the following groups of people:

- Place of work: East/West Germany
- Three education groups²⁹

Within the imputations we control for: sex, part-time, age, age squared, days in job, days in job squared, and interaction of age and age squared with “old”, where “old” is a dummy that is 1 for individuals older than 40 years. We do not consider marginal part-time spells in the imputation, since these wages—by definition—cannot be right censored.

At the end, we obtain our SIAB base data set, which includes full-time, part-time, and marginal part-time spells and, inter alia, the (imputed) nominal gross daily wage associated with them. We use this data set to create the four data sets for our analyses.

Preparation of the four data sets for our analyses

We create four data sets for the analyses. Two of them use wage measures that take advantage of the SIAB (average daily wage and daily wage on June 30) and the two “comparable” data sets use wage measures that can also be generated with the SOEP (monthly wage and annual wage).

SIAB-Pure 1: average daily wage

To generate this data set, we use all spells from our SIAB base data set. If an individual has more than one employment spell (parallel and/or consecutive) in a year, we compute a weighted average daily wage, similar to Dustmann et al. (2009). The weighted average daily wage of person i in year t ($wdw_{i,t}$) is calculated as:

$$wdw_{i,t} = \sum_{s=1}^{S_{i,t}} \left((dw_{i,t,s} * l_{i,t,s}) / \sum_s l_{i,t,s} \right),$$

²⁹ (1) neither vocational training nor degree from university (of applied science), (2) vocational training, (3) degree from an university (of applied science).

Table 2 Summary statistics SOEP-Pure 2 sample

No. of observations	Mean wage	Median wage	Mean age
204,560	34,369	30,250	43.01

Source: own calculations, SOEP-Core.v37.EU, years 2001–2020

Wage: nominal annual wage of the main occupation in euro, mean over (median of) all observations of the sample. Age: mean age in years

Only wages above the marginal earnings threshold are taken into account

where $s = [1, S_{i,t}]$ indicates the different spells of worker i in year t , $dw_{i,t,s}$ is the nominal gross daily wage of spell s of worker i in year t and $l_{i,t,s}$ is the corresponding spell length (in days). Hence, if a worker has only one spell ($S_{i,t} = 1$), the daily wage is unchanged ($wdw_{i,t} = dw_{i,t,s=1}$). If more than one spell exists, $wdw_{i,t}$ is the spell length weighted average daily wage.

For the final analysis, we only consider wages that are above the marginal earnings threshold. Summary statistics for this data set are provided in Table 5.

SIAB-Pure 2: daily wage on June 30

To create this data set, we use all spells from our SIAB base data set that include the June 30 cutoff date. We sum up the daily wages for each worker for every year $wdw_{i,t}$:

$$dw_{i,\text{June } 30,t} = \sum_{s=1}^{S_{i,\text{June } 30,t}} (dw_{s,i,t}).$$

For the final analysis, we only consider wages that are above the marginal earnings threshold. Summary statistics for this data set are provided in Table 6.

Comparable sample 1: monthly wage

To obtain a comparable monthly wage, we transfer the survey pattern of the SOEP to the SIAB. For this purpose, the distribution of the survey months of each year, is randomly assigned to the individuals in the SIAB. The assigned interview month is then merged to our SIAB base data set. As in the SOEP samples, individuals are only included if they have earnings in the month prior to the assigned interview. This previous month’s wage is used in the analyses.

We calculate the exact monthly wage for all employment spells of the month prior to the assigned interview month $mw_{i,\text{main},t}$ as:

$$mw_{i,\text{main},t} = dw_{i,\text{main},t} * l_{i,\text{month},t},$$

where $dw_{i,\text{main},t}$ is the nominal gross daily wage of the main employment spell of the month prior to the assigned interview month and $l_{i,\text{month},t} = [1, 31]$ are the days worked in that month.

For the analysis we only keep the main employment spell (the spell with the highest monthly wage) of the month prior to the assigned interview month.

Table 3 Summary statistics SOEP: comparable sample 1

No. of observations	Wage (in euro)		Mean	Working (in %)		Education (in %)		
	mean	median	age	f-t	p-t	low	med.	high
195,168	2838.60	2485.83	42.65	70.91	29.09	10.68	68.05	21.04

Source: own calculations, SOEP-Core.v37.EU, years 2000–2019

Wage: nominal monthly wage in euro, mean over (median of) all observations of the sample. Age: mean age in years. Working: f-t = full-time; p-t = part-time. Education: low = no vocational training; medium = vocational training; high = university (of applied sciences)

Only wages above the marginal earnings threshold are taken into account

Due to missings in working-time and education, percentages do not sum up to 100%

Table 4 Summary statistics SOEP: comparable sample 2

No. of observations	Mean wage	Median wage	Mean age
187,010	33,654.76	29,400.00	42.87

Source: own calculations, SOEP-Core.v37.EU, years 2001–2020

Wage: nominal annual wage in euro, mean over (median of) all observations of the sample. Age: mean age in years

Only wages above the marginal earnings threshold are taken into account

For the final analysis, we only consider wages that are above the marginal earnings threshold. Summary statistics for this data set are provided in Table 7.

Comparable sample 2: annual wage

To obtain a comparable annual wage $w_{i,t}$, we sum up all wages of a worker in each year:

$$w_{i,t} = \sum_{s=1}^{S_{i,t}} (dw_{s,i,t} * l_{s,i,t}),$$

where $s = [1, S_{i,t}]$ indicates the different spells of worker i in year t , $dw_{s,i,t}$ is the nominal gross daily wage of spell s of worker i in year t and $l_{s,i,t}$ is the corresponding spell length (in days).

For the final analysis, we only consider wages that are above the marginal earnings threshold. Summary statistics for this data set are provided in Table 8.

Appendix C

Structure of earnings survey (VSE)

To better assess the difference between SIAB and SOEP highlighted in Sect. 3.3, we draw on the Structure of Earnings Survey (VSE, Verdienststrukturhebung) for additional comparison.

The official report on the VSE is a decentralized set of statistics and provides gross wage percentiles for the years 2014 (Statistisches Bundesamt 2016) and 2018 (Statistisches Bundesamt 2020). The data are collected from public and private sector employers via online

questionnaires. Employers are obliged to provide information in accordance with the Earnings Statistics Act. The VSE covers full-time and part-time jobs. Companies that have only marginal employees are not included in the VSE. Also self-employment is not covered. Only jobs that existed for the entire reporting month and for which wages were paid in the reporting month are recorded. The reporting period is the calendar year for some data and a representative month (April) for most data. Therefore, seasonal employment is not captured in a representative manner. The VSE covers all sectors with the exemption of employees in private households and in extraterritorial organizations and entities. We refrain from an in-depth comparison between all three data-sources because the VSE collects wage information only every fourth year and micro data are only available since 2006.

In Table 9, we compare the percentiles and percentile ratios of the VSE with the results from the unrestricted SOEP (SOEP-Pure 1 and 2) and the comparable SIAB data (SIAB-Comparable 1 and 2).

When interpreting the tables, it should be noted that the VSE does not record and depict persons, but rather employment relationships, i.e. jobs or employment contracts. For example, if an employee has two part-time jobs in two different surveyed establishments and is sampled in both establishments, he/she appears in the VSE statistics two times. The person appears only once in the VSE if the second establishment was not surveyed or he/she was not sampled in the second establishment.

The VSE considers in the gross nominal monthly wages (GMW) only employment relationships that existed during the entire reference month (i.e. April) and for which remuneration was paid in the reference month. This excludes employments that were not started or ended on a monthly basis. Therefore, short employments—which are usually associated with low wages—are not taken into account. In the SIAB, on the other hand, employment is recorded on a daily basis. Therefore, short employments—even just one-day ones—are also taken into account. In the SOEP, the wages of the previous month are queried—here, too, employment that was not started or ended on

a monthly basis is recorded. However, it is questionable whether respondents would report one-day or very short employment, especially if it is a second or irregular job. In the GMW results of the SOEP and the SIAB, only the job with the highest gross monthly wage is considered.

In the VSE, the gross nominal annual wages (GAW) of partial years were extrapolated to 12 months. Only employment relationships with 30 or more working weeks in 2014 or 2018 are considered. In the GAW of the SOEP and the SIAB, on the other hand, all jobs of a person are taken into account.

Comparing the percentiles of the GMW of the VSE with the SOEP and the SIAB, it is noticeable that the P10 and P50 in the SOEP are higher than the values in the VSE in both years. The P90, on the other hand, is higher in both years in the VSE. In the SIAB, a different pattern emerges. The P10 of the SIAB and the VSE are remarkably close to each other in both years. On the other hand, the P50 and P90 of the SIAB are always significantly higher than those of the VSE. An explanation of the higher P10 in the SOEP could be, that respondents with several marginal part-time jobs add them together and thus report higher wages than the other two sources.

Comparing the percentiles of the GAW of the VSE with those of the SOEP, we find a similar pattern as for the GMW. Again, the P10 is higher in the SOEP than in the VSE, and the P90 is higher in the VSE. But the P50 is higher in 2018 only in the SOEP. However, there are stronger deviations in the comparison of SIAB and VSE. Here, all three percentiles of the SIAB are significantly lower than those of the VSE. The large deviations in the GAW can probably be explained by the fact that the SIAB includes all employment, no matter how brief. In other words, many people who are perhaps only employed for a few days a year are taken into account here and therefore push the percentile values down. In the VSE, on the other hand, these short jobs are not taken into account at all. It also cannot be ruled out that companies with more marginal part-time employees are underrepresented due to the composition of the companies surveyed. In addition, the VSE records wages in April and therefore seasonal employment is not recorded in a representative manner. Also in the SOEP low wages, especially those due to short-time jobs, could be underrepresented. This may be due to sampling, but also because respondents did not report very low wages or simply forget about very short-term employment.

In the period we consider, marginal part-time wages are included in the SIAB, because since 1999 wages below the marginal earnings threshold jobs are subject to a lump-sum contribution payable by the employer. For this reason, marginal part-time wages should be recorded relatively reliably in the SIAB. Here, too, however, it cannot be ruled out that, e.g. due to subsequent reporting of

Table 5 Summary statistics SIAB-Pure 1 sample

No. of observations	Mean wage	Median wage	Mean age
11,492,264	90.44	76.15	41.77

Source: own calculations, SIAB 7519, years 2000–2019

Wage: nominal daily wage in euro, mean over (median of) all observations of the sample. Age: mean age in years

Only wages above the marginal earnings threshold are taken into account

Table 6 Summary statistics SIAB-Pure 2 sample

No. of observations	Mean wage	Median wage	Mean age
10,376,844	96.59	82.38	42.41

Source: own calculations, SIAB 7519, years 2000–2019

Wage: nominal daily wage in euro, mean over (median of) all observations of the sample. Age: mean age in years

Only wages above the marginal earnings threshold are taken into account

wages, spells with low wages occur that do not represent actual employment. Although these are administrative data, there are observations with unrealistically low daily wages. This could lead to an overestimation of marginal part-time employment—and hence an underestimated P10—in the SIAB.

To shed more light on the impact of the VSE constraints just mentioned, we show in Table 10 what happens when we apply these constraints to the SIAB. The gross nominal monthly wage in April (GMW) for the VSE and the gross nominal wage (GMW, SIAB-Comparable 1) correspond to the values in Table 9; they are only listed again for the sake of clarity. GMW2 to GMW5 show what happens when restrictions of the VSE are applied to the SIAB.

GMW2 corresponds to GMW of the SIAB, except for the fact that the workers must be employed for the entire month—as in the VSE—in order to be taken into account. As a result, the P10, the P50 and the P90 increase. In addition, restricting the SIAB to the month April—as in the VSE—causes all percentiles to fall again (see GMW3). However, with the exception of the P10 in 2018, they are above GMW's values. GMW4 and GMW5 correspond to GMW2 and GMW3, respectively, except that we now exclude employment in sectors T and U—as in the VSE. As can be seen from Table 10, this additional constraint has no effect on P10, but P50 and P90 increase slightly compared to GMW2 and GMW3. However, the general pattern remains: The P10 percentiles are close, but the P50 and P90 percentiles are higher in the SIAB, and this difference has increased over time.

It should be noted that it is evident that the VSE restrictions have an impact on the wage percentile(s) (ratios). However, they do not explain the larger difference, of

Table 7 Summary statistics SIAB: comparable sample 1

No. of observations	Wage (in euro)		Mean	Working (in %)		Education (in %)		
	mean	median	age	f-t	p-t	low	medium	high
9,859,452	2905.86	2478.76	42.65	76.34	23.66	7.85	73.82	16.52

Source: own calculations, SIAB 7519, years 2000–2019

Wage: nominal monthly wage in euro, mean over (median of) all observations of the sample. Age: mean age in years. Working: f-t = full-time; p-t = part-time.

Education: low = no vocational training; medium = vocational training; high = university (of applied sciences)

Only wages above the marginal earnings threshold are taken into account

Due to missings in working-time and education, percentages do not sum up to 100%

Table 8 Summary statistics SIAB, comparable sample 2

No. of observations	Mean wage	Median wage	Mean age
10,808,103	33,353.38	28,406.06	42.13

Source: own calculations, SIAB 7519, years 2000–2019

Wage: nominal annual wage in euro, mean over (median of) all observations of the sample. Age: mean age in years

Only wages above the marginal earnings threshold are taken into account

about 10%, in the P90 between VSE and SIAB. This is noteworthy in that both wage information are employer declarations where a high degree of accuracy should be expected. What could still be investigated is whether or to what extent drawing a company sample has an effect on the wage inequality found. Since such an evaluation is not possible with SIAB, we leave this to future research.³⁰

Table 9 Selected 2014 and 2018 wage percentiles and wage percentile ratios for VSE, SOEP, and SIAB

Data set	VSE		SOEP		SIAB	
	GMW	GAW	GMW	GAW	GMW	GAW
<i>Percentiles</i>						
	Year 2014					
P10	396	4800	450	5400	395	2787
P50	2269	29,697	2301	29,600	2355	24,636
P90	4707	62,939	4668	61,800	5109	58,597
<i>Percentile ratios</i>						
	Year 2014					
P90/P10	11.89	13.11	10.37	11.44	12.95	21.02
P50/P10	5.73	6.19	5.11	5.48	5.97	8.84
P90/P50	2.07	2.12	2.03	2.09	2.17	2.38
<i>Percentiles</i>						
	Year 2018					
P10	438	5337	720	7800	448	3614
P50	2500	32,714	2600	33,000	2621	27,766
P90	5151	68,508	5000	66,097	5633	64,686
<i>Percentile ratios</i>						
	Year 2018					
P90/P10	11.76	12.84	6.94	8.47	12.57	17.90
P50/P10	5.71	6.13	3.61	4.23	5.85	7.68
P90/P50	2.06	2.09	1.92	2.00	2.15	2.33

VSE Structure of Earnings Survey. GMW: gross nominal monthly wage in April. GAW: gross nominal annual wage. Total without apprentices. Economic sections A-S of the Classification of economic activities. Thus, not considered is employment in sectors T (Private households with employed persons) and U (Extraterritorial organizations and bodies). Source: Statistisches Bundesamt (2016, 2020)

SOEP Socio-Economic Panel (SOEP-Core.v37.EU, own calculations). GMW: gross nominal monthly wage (SOEP-Pure 1). GAW: gross nominal annual wage (SOEP-Pure 2)

SIAB Sample of Integrated Labour Market Biographies (SIAB 7519, own calculations). GMW: gross nominal wage (SIAB-Comparable 1); GAW: gross nominal annual wage (SIAB-Comparable 2)

All wages in euro

³⁰ It would be conceivable, e.g., to use the SIEED—a 1.5% sample of all establishments in Germany (see Schmidlein et al. 2021). For these establishments, the entire employment histories of the persons employed in these establishments is available. However, the SIEED would have the disadvantage that the data are only available at mid-year (reference date 30.6). In principle, however, the SIEED would make it possible to show what happens when different establishments are drawn and the various measures of inequality are then calculated on the basis of the associated employment relationships.

Table 10 Selected 2014 and 2018 wage percentiles and wage percentile ratios for VSE and SIAB

Data set	VSE	SIAB				
	GMW	GMW	GMW2	GMW3	GMW4	GMW5
<i>Percentiles</i>	Year 2014					
P10	396	395	407	397	407	397
P50	2269	2355	2400	2388	2402	2390
P90	4707	5109	5153	5127	5157	5130
<i>Percentile ratios</i>	Year 2014					
P90/P10	11.89	12.95	12.66	12.92	12.67	12.92
P50/P10	5.73	5.97	5.90	6.02	5.90	6.02
P90/P50	2.07	2.17	2.15	2.15	2.15	2.15
<i>Percentiles</i>	Year 2018					
P10	438	448	458	444	458	444
P50	2500	2621	2670	2659	2672	2661
P90	5151	5633	5680	5648	5683	5651
<i>Percentile ratios</i>	Year 2018					
P90/P10	11.76	12.57	12.39	12.73	12.40	12.74
P50/P10	5.71	5.85	5.82	5.99	5.83	6.00
P90/P50	2.06	2.15	2.24	2.12	2.13	2.12

VSE: Structure of Earnings Survey. GMW: gross nominal monthly wage in April. Total without apprentices. Economic sections A-S of the Classification of economic activities. Thus, not considered is employment in sectors T (Private households with employed persons) and U (Extraterritorial organizations and bodies). Source: Statistisches Bundesamt (2016, 2020)

SIAB: Sample of Integrated Labour Market Biographies (SIAB 7519, own calculations). GMW: gross nominal wage (SIAB-Comparable 1); GMW2: gross nominal wage (SIAB-Comparable 1) but workers must work entire month (as in VSE); GMW3: gross nominal wage in April and workers must work entire month (as in VSE); GMW4: gross nominal wage (SIAB-Comparable 1) but workers must work entire month (as in VSE), not considering sectors T and U (as in VSE); GMW5: gross nominal wage in April and workers must work entire month (as in VSE), not considering sectors T and U (as in VSE)

All wages in euro

Acknowledgements

We would like to thank Mattis Beckmannshagen and Joachim Möller for very useful comments and suggestions. We would also like to thank Bernd Fitzenberger and participants of the 2016 IAB-IWH Workshop and the 2014 SOEP User Conference for useful comments on an earlier version of the project. All errors remain our own.

Author contributions

The authors read and approved the final manuscript.

Funding

This article was written when Heiko Stüber was an employee of the IAB, Markus M. Grabka was an employee of the DIW Berlin/SOEP, and Daniel D. Schnitzlein was an employee of the Leibniz Universität Hannover. The authors did not receive any special funding for this publication.

Availability of data and materials

The SIAB 7519 is not publicly accessible for data protection reasons (social security data). However, the Research Data Center (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) makes this data set available to the scientific community. The SOEP-Core.v37. EU can be ordered for scientific purposes via the research data center of the SOEP at DIW Berlin. Details on the procedure can be found at https://www.diw.de/en/diw_01.c.601584.en/data_access.html.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 23 December 2022 Accepted: 19 January 2023

Published online: 21 February 2023

References

- Bachmann, R., Boockmann, B., Gonschor, M., Kalweit, R., Klauser, R., Laub, N., Rulff, C., Vonnahme, C., Zimpelmann, C.: Auswirkungen des gesetzlichen Mindestlohns auf Löhne und Arbeitszeiten, IZA Research Report 96 (2020)
- Bakker, B.F.: Estimating the validity of administrative variables. *Stat. Neerl.* **66**(1), 8–17 (2011)
- Bartolucci, C.: Gender wage gaps reconsidered a structural approach using matched employer-employee data. *J. Hum. Resour.* **48**(4), 998–1034 (2013)
- Beaule, A., Campbell, F., Mary, D., Insolera, N., Johnson, D., Juska, P., McGonagle, K., Simmert, B., Warra, J.: PSID main interview user manual: release 2021, Institute for Social Research, University of Michigan (2021)
- Becker, I., Frick, J.R., Grabka, M.M., Hauser, R., Krause, P., Wagner, G.G.: A comparison of the main household income surveys for Germany: EVS and SOEP. In: Hauser, R., Becker, I. (eds.) Reporting on income distribution and poverty, pp. 55–90. Springer, Berlin (2003)
- Beckmannshagen, M., Schröder, C.: Earnings inequality and working hours mismatch. *Labour Econ.* **76**, 102184 (2022)
- Bender, S., Hilzendege, J., Rohwer, G., Rudolph, H.: Die IAB-Beschäftigtenstichprobe 1975–1990, Beiträge zur Arbeitsmarkt- und Berufsforschung 197 (1996)
- Biewen, M.: Income inequality in Germany during the 1980s and 1990s. *Rev. Income Wealth* **46**(1), 1–19 (2000)
- Biewen, M., Juhasz, A.: Understanding rising income inequality in Germany, 1999/2000–2005/2006. *Rev. Income Wealth* **58**, 622–647 (2012)
- Blough, D.K., Ramsey, S., Sullivan, S.D., Yusen, R.: The impact of using different imputation methods for missing quality of life scores on the estimation

- of the cost-effectiveness of lung-volume-reduction surgery. *Health Econ.* **18**(1), 91–101 (2009)
- Bollinger, C.R.: Measurement error in the current population survey: a nonparametric look. *J. Labor Econ.* **16**(4), 576–594 (1998)
- Bound, J., Brown, C., Duncan, G.J., Rodgers, W.L.: Evidence on the validity of cross-sectional and longitudinal labor market data. *J. Labor Econ.* **12**(3), 345–368 (1994)
- Bound, J., Krueger, A. B.: The extent of measurement error in longitudinal earnings data: do two wrongs make a right? *J. Labor Econ.* **9**(1), 1–24 (1991)
- Bound, J., Brown, C., Mathiowetz, N.: Chapter 59—measurement error in survey data. In: Heckman, J.J., Leamer, E.E. (eds.) *Handbook of econometrics*, vol. 5, 1st edn., pp. 3705–3843. Elsevier, Amsterdam (2001)
- Bricker, J., Engelhardt, G.V.: Measurement error in earnings data in the health and retirement study. *J. Econ. Soc. Meas.* **33**(1), 39–61 (2008)
- Britzke, J., Schupp, J.: SOEP wave report 2018. SOEP, Berlin (2019)
- Bundesregierung: Bericht der Bundesregierung zur Lebensqualität in Deutschland. Bundesregierung, Berlin (2016)
- Bundesregierung: Lebenslagen in Deutschland - Der Sechste Armuts- und Reichtumsbericht der Bundesregierung. Bundesregierung, Berlin (2021)
- Burauel, P., Caliendo, M., Grabka, M.M., Obst, C., Preuss, M., Schröder, C., Shupe, C.: The impact of the german minimum wage on individual wages and monthly earnings. *Jahrb. Natl. ökon. Stat.* **240**(2–3), 201–231 (2020)
- Caliendo, M., Wittbrodt, L.: Did the minimum wage reduce the gender wage gap in Germany? *Labour Econ.* **78**, 102228 (2022)
- Card, D., Heining, J., Kline, P.: Workplace heterogeneity and the rise of West German wage inequality. *Quart. J. Econ.* **128**(3), 967–1015 (2013)
- Carpente, H.: UK household longitudinal study—wave 11 technical report. The Institute for Social and Economic Research, University of Essex, Colchester (2021)
- Coder, J.: Using administrative record information to evaluate the quality of the income data collected in the survey of income and program participation. *Proc. Stat. Can. Symp.* **92**, 295–306 (1992)
- de Leeuw, E.D., Hox, J.J., Dillman, D.A.: Chapter 1—the cornerstones of survey research. In: Hox, J., Leeuw, E.D., Dillman, D. (eds.) *International Handbook of Survey Methodology*, pp. 1–17. European Association of Methodology, Frankfurt (2008)
- Dustmann, C., Fitzenberger, B., Schönberg, Uta, Spitz-Oener, A.: From sick man of Europe to economic superstar: Germany's resurgent economy. *J. Econ. Perspect.* **28**(1), 167–188 (2014)
- Dustmann, C., Fitzenberger, B., Schönberg, U., Spitz-Oener, A., Ludsteck, J., Schönberg, U.: Revisiting the German wage structure. *Quart. J. Econ.* **124**(2), 843–881 (2009)
- Felbermayr, G., Baumgarten, D., Lehwald, S.: *Increasing wage inequality in Germany — what role does global trade play?* Bertelsmann Stiftung (2014)
- Fitzenberger, B., Seidlitz, A.: The 2011 break in the part-time indicator and the evolution of wage inequality in Germany. *J. Labour Mark. Res.* **54**, 1 (2020)
- Fitzenberger, B., Seidlitz, A., de Lazer, J.: Changing selection into full-time work and its effect on wage inequality in Germany. *Empir. Econ.* **62**, 247–277 (2022)
- Frick, J.R., Grabka, M.M.: Item nonresponse on income questions in panel surveys: incidence, imputation and the impact on inequality and mobility. *Allg. Stat. Arch.* **89**, 49–61 (2005)
- Frodermann, C., Schmucker, A., Seth, S., vom Berge, P.: Sample of Integrated Labour Market Biographies (SIAB) 1975–2019. FDZ-Datenreport 01/2021 (en) (2021)
- Ganzer, A., Lisa S., Jens S., Wolter S.: Establishment History Panel 1975–2018. FDZ-Datenreport 01/2020 (en), revised version (v2) from April 2021 (2020)
- Gernandt, J., Pfeiffer, F.: Rising wage inequality in Germany. *Jahrb. Natl. ökon. Stat.* **227**(4), 358–380 (2007)
- Goebel, J., Grabka, M.M., Liebig, S., Kroh, M., Richter, D., Schröder, C., Schupp, Jürgen.: The German Socio-Economic Panel (SOEP). *Jahrb. Natl. ökon. Stat.* **239**(2), 345–360 (2019)
- Grabka, M.M.: Income inequality in Germany stagnating over the long term, but decreasing slightly during the coronavirus pandemic. *DIW Wkly. Rep.* **17**(18), 125–133 (2021)
- Grabka, M.M.: Löhne, Renten und Haushaltseinkommen sind in den vergangenen 25 Jahren real gestiegen. *DIW Wochenber.* **89**(23), 329–337 (2022)
- Grabka, M.M., Schröder, C.: Inequality in Germany: decrease in gap for gross hourly wages since 2014, but monthly and annual wages remain on plateau. *DIW Wkly. Rep.* **8**(9), 83–92 (2018)
- Groen, J.A.: Sources of error in survey and administrative data: the importance of reporting procedures. *J. Off. Stat.* **28**(2), 173–198 (2012)
- Hauser, R., Becker, I., Grabka, M.M., Westerheide, P.: Integrierte Analyse der Einkommens- und Vermögensverteilung: Abschlussbericht zur Studie im Auftrag des Bundesministeriums für Arbeit und Soziales, Bonn, Vol. A369 of Forschungsbericht. Bundesministerium für Arbeit und Soziales, Berlin (2007)
- Johansson, F., Klevmar, A.: Comparing register and survey wealth data. *Int. J. Microsimulation* **15**(1), 1 (2022)
- Kapteyn, A., Ypma, J.Y.: Measurement error and misclassification: a comparison of survey and administrative data. *J. Labor Econ.* **25**(3), 513–551 (2007)
- Kavonius, I.K., Törmälehto, V.M.: Household income aggregates in micro and macro statistics. *Stat. J. U. N. Econ. Comm. Eur.* **20**(1), 9–25 (2003)
- Klein, M.W., Moser, C., Urban, D.M.: Exporting, skills and wage inequality. *Labour Econ.* **25**, 76–85 (2013)
- Kopczuk, W., Saez, E., Song, J.: Earnings inequality and mobility in the United States: evidence from social security data since 1937. *Quart. J. Econ.* **125**(1), 91–128 (2010)
- Lindner, P., Andreasch, M.: Micro and macro data: a comparison of the household finance and consumption survey with financial accounts in Austria. *J. Off. Stat.* (2014). <https://doi.org/10.1515/jos-2016-0001>
- Little, R., Su, H.L.: Item non-response in panel survey. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (eds.) *Panel surveys*, pp. 400–425. Wiley, New York (1989)
- Lüthen, H., Schröder, C., Grabka, M.M., Goebel, J., Mika, J., Brüggmann, D., Ellert, S., Penz, H.: SOEP-RV: linking German Socio-Economic Panel data to pension records. *Jahrb. Natl. ökon. Stat.* **242**(2), 291–307 (2022)
- Möller, J.: Lohnungleichheit—Gibt es eine Trendwende? *Wirtschaftsdienst* **96**(13), 38–44 (2016)
- Oberski, D. L., Kirchner, A., Eckman, S., Kreuter, F.: Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *J. Am. Stat. Assoc.* **112**(520), 1477–1489 (2017)
- OECD: A broken social elevator? How to promote social mobility. OECD Publishing, Paris (2018)
- Pedace, R., Bates, N.: Using administrative records to assess earnings reporting error in the survey of income and program participation. *J. Econ. Soc. Meas.* **26**(3–4), 173–192 (2000)
- Pischke, J.S.: Individual income, incomplete information, and aggregate consumption. *Econometrica* **63**(4), 805–840 (1995)
- Roemer, M.: Using administrative earnings records to assess wage data quality in the march current population survey and the survey of income and program participation. Longitudinal employer-household dynamics technical papers, Center for Economic Studies. U.S. Census Bureau (2002)
- Schmidtlein, L., Seth, S., vom Berge, P.: Sample of Integrated Employer-Employee Data 1975–2018. FDZ-Datenreport 14/2020 (en) (2021)
- Schröder, C., König, J., Fedorets, A., Goebel, J., Grabka, M.M., Lüthen, H., Metzger, M., Schikora, F., Liebig, S.: The economic research potentials of the German Socio-Economic Panel study. *German Econ. Rev.* **21**(3), 335–371 (2020)
- Selezneva, E., Kerm, P.V.: A distribution-sensitive examination of the gender wage gap in Germany. *J. Econ. Inequal.* **14**(1), 21–40 (2016)
- Sommerfeld, K.: Higher and higher? Performance pay and wage inequality in Germany. *Appl. Econ.* **45**(30), 4236–4247 (2013)
- Statistisches Bundesamt. Verdienststrukturerhebung 2014. Fachserie 16, Heft 1 (published on 14.09.2016, corrected on 10.03.2017) (2016)
- Statistisches Bundesamt. Verdienststrukturerhebung 2018. Fachserie 16, Heft 1 (published on 14.09.2020, updated on 06.10.2020) (2020)
- Steiner, V., Wagner, K.: Has earnings inequality in Germany changed in the 1980's? *Zeitschrift für Wirtschafts- und Sozialwissenschaft* **118**, 29–54 (1998)
- Stockhausen, M., Calderón, M.: IW-Verteilungsreport: Stabile Verhältnisse trotz gewachsener gesellschaftlicher Herausforderungen. IW-Report 8 (2020)
- Stüber H., Dauth W., Eppelsheimer J.: A guide to preparing the Sample of Integrated Labour Market Biographies (SIAB, version 7519 v1) for scientific analysis. *J. Labour Mark. Res.* **57**, 7 (2023). <https://doi.org/10.1186/s12651-023-00335-w>
- Törmälehto, V.M.: LIS and national accounts comparison. LIS Technical Working Paper Series **2** (2011)
- Tyrowicz, J., van der Velde, L., van Staveren, I.: Does age exacerbate the gender-wage gap? New method and evidence from Germany, 1984–2014. *Feminist Econ.* **24**(4), 108–130 (2018)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.