INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

IAB-DISCUSSION PAPER

Articles on labour market issues

11|2020 Measurement error in minimum wage evaluations using survey data

Mario Bossler, Christian Westermeier

ISSN 2195-2663

Measurement error in minimum wage evaluations using survey data

Mario Bossler (IAB, LASER),
Christian Westermeier (IAB)

Contents

List of Figures

List of Tables

Abstract

We assess the role of measurement error in minimum wage evaluations when the treatment variable – the bite – is inferred from a survey wage distribution. We conduct Monte Carlo experiments on both simulated and empirical distributions of measurement error derived from a record linkage of survey wages and administrative data. On the individual-level treatment effects are downward biased by more than 30 percent. Aggregation of the treatment information at the household, firm or region level does not fully alleviate the bias. In fact, the magnitude and direction of the bias depend on the size of the aggregation units and the allocation of treated individuals to such units. In cases of a strongly segregated allocation, measurement error can cause upward biased treatment effects. Besides aggregation, we discuss two possible remedies: the use of a continuous treatment variable and dropping observations close to the minimum wage threshold.

Zusammenfassung

Wir analysieren den Einfluss von Messfehlern in Mindestlohnevaluationen, wenn die Treatmentvariable, also der "Bite", aus einer befragungsbasierten Lohnverteilung stammt. In Monte-Carlo-Simulationen überprüfen wir die Verzerrtheit der Schätzer sowohl mit simulierten als auch empirischen Verteilungen von Messfehlern. Die empirischen Messfehler stammen aus einem Link von Befragungsdaten und administrativen Daten. Auf der individuellen Beobachtungsebene werden die Treatmenteffekte über 30 Prozent unterschätzt. Eine Aggregation der Treatmentinformation auf der Ebene von Haushalten, Firmen oder Regionen löst das Problem nicht vollständig. In Fällen einer sehr stark segregierten Verteilung von betroffenen Beschäftigten auf nur wenige Firmen oder Regionen kann es sogar zu einer Überschätzung des wahren Effekts kommen. Wir diskutieren zwei Lösungsansätze: Die Verwendung einer kontinuierlichen Treatmentvariable und das Löschen von Observationen, die in der Lohnverteilung nahe der Mindestlohnschwelle liegen.

JEL

C21, C43, J38

Keywords

attenuation bias, difference-in-differences, measurement error, minimum wage, misclassification, record linkage, treatment effects, survey data, wage distribution

Acknowledgements

We thankfully acknowledge helpful comments from participants at the 2018 conference of the German Statistical Society, the 2018 Conference on Minimum Evaluation of the DIW Berlin, the 2019 Workshop on Labor Economics in Trier, the 2019 conference of the Canadian Economic Association, the 2019 conference of the European Society of Population Economics, the 2019 Conference of the European Association of Labour Economists, and the 2019 Conference of the German Economic Association. We also acknowledge particularly helpful comments and suggestions from Bernd Fitzenberger, Ingo Ipshording and Michael Oberfichtner. We acknowledge valuable research assistance from Ute Weber and the Data- and IT- Management department at the IAB.

1. Introduction

The collection of income data in surveys is prone to human error. Respondents may not recall transitory components of their income, confound months, or come up with overly crude approximations, all in good faith. Due to the sensitivity of the information, respondents might also report an incorrect value deliberately.

In empirical evaluation studies of nationwide minimum wages, assignment to a treatment group is determined by the position in the observed wage distribution. Measurement error in the wage variable might cause biased estimation of treatment effects due to misclassification of persons into treatment and control groups. One recent example is the introduction of a nationwide minimum wage in Germany: for the evaluation of its causal effects, scholars typically apply a difference-in-differences identification strategy.[1] Consequently, they have to assign individuals to treatment and control groups. Individuals are assigned to the former if the wage is below the new minimum wage threshold before the law came into force and to the latter if the wage already exceeds the forthcoming minimum wage.

In the presence of misclassification into treatment and control groups due to classical measurement error in wages, regression attenuation shrinks the estimated treatment effect towards zero. Scholars typically claim to estimate a lower bound of the true effect without elaborating the actual size of the bias. However, the size of this bias is relevant, in particular when it concerns evaluations of disruptive policy changes such as minimum wages. In this study, we find that the bias can be quite substantial when we apply both simulated and empirical distributions of measurement error in survey-based wage data in a difference-in-differences minimum wage evaluation.

Scholars typically hope to cancel out the measurement error by aggregating data in policy evaluations (Bound/Brown/Mathiowetz, 2001).[2] However, we are not aware of any study that analyses potential estimation biases of measurement error when aggregating treatment variables to higher-level units. In fact, our results do not confirm the common belief that aggregation fully alleviates attenuation bias. We also observe scenarios in which the bias changes sign, leading to an overestimation of the actual treatment effect.

To analyze the bias induced by measurement error, we present a simulation study that addresses measurement error in an artificially generated setting of a minimum wage introduc-

[1] The results in this paper, however, are not restricted to difference-in-differences estimation. All results can be generalized to OLS-based treatment effect estimations.

[2] Aggregating data is typically not done solely to alleviate attenuation bias. In difference-in-differences evaluations of uniform minimum wages, aggregated data can help to eliminate spillovers, i.e., to address violations of the SUTVA. However, we are focusing on the consequences of measurement error, ruling out any spillover effects by design.

tion. This is first and foremost a simulation study that also uses empirical measurement error distributions in later sections. In comparison to an empirical application using simulated data has some merits. First, it allows to precisely quantify the size of the bias relative to a generated effect size. Second, it yields a range of results, whereas a single application would only yield a single point estimate with large standard errors, which would have little informative value. Finally, it is not feasible to analyse measurement error in an application, simply because there is no data that is free of any measurement error in hourly wages. Even in German administrative data, the collection of gross monthly wages follows a definition that is different from the minimum wage legislation, and even more important, the reporting of working hours to the German Statutory Accident Insurance allows for estimated values and does not distinguish between contractual and actual hours, both creating a meaningful error in the respective distribution of hourly wages.

In the simulations, we start by assuming a normally distributed wage distribution and classical measurement error. We show that the treatment effect is biased downwards if we estimate the effect of a minimum wage on a generated dependent variable. Moreover, changing the error structure to include a nonclassical term hardly changes the results. We then implement a series of Monte Carlo experiments to study the bias after aggregating the data to higher-level units (e.g. households, establishments or regions). We use a synthetic allocation of individuals to such units, accounting for varying degrees of segregation of minimum wage workers across units and varying unit sizes. In the absence of measurement error in wages, the resulting grouped data regression yields unbiased estimates of the actual treatment effect (Angrist/Pischke, 2008; Prais/Aitchison, 1954). However, in the presence of fuzzy wage data, the bias depends on the size of the aggregation units and the level of segregation. Again, including a nonclassical error term has little impact on the bias compared with a classical term.

Besides the aggregation of data to higher-level units, which entails a loss of efficiency due to the loss of observations, we propose and discuss two alternative remedies. First, when building the regression model, scholars might incorporate a continuous treatment variable indicating treatment intensity instead of a binary treatment assignment. On the individual level, the results hardly improve, while for aggregated data with moderate individual-level contamination, we recommend an approach using a continuous treatment variable. Second, and somewhat more radical, we discuss deleting individuals close to the minimum wage threshold before its introduction, as those observations are the most prone to misclassification. We find that this strategy is only recommendable if we assume that the dummy variable catches the actual relationship between minimum wage and a dependent variable. When instead the relationship would necessitate a continuous treatment variable, dropping observation does not function as a remedy but yields a positive bias.

In the last part of this paper, we exploit the record linkage of survey wage data with German administrative data – including both wages and working hours – to quantify the distribution of measurement error in wages from the difference in survey-based wage information and administrative wage data. Even if the administrative data may entail measurement error, the difference of both data sources allows us to quantify the distribution of measurement error from a comparison of two independent data sources. This approach also allows us to better assess recent evaluation studies of the literature based on our new insights concerning potential biases. Even if the empirical execution is clean by conventional empirical standards, the results could be biased due to the fuzziness in existing wage data.

With regard to the generalizability of the results, it needs to be stressed that all recent introductions of a nationwide minimum wage in developed (and some developing) countries are similar compared with the German minimum wage introduction, which we study in this paper. Most importantly, the coverage of a minimum wage among the working population, as well as the level of a minimum wage (as a percentage of the median wages) are informed by similar political considerations across countries. Policy makers want aim for a minimum wage to be high enough to allow for low-wage workers to make a decent living, but low enough as to not affect middle class and minimize the overall impact on employment. In 1999, the new British National Minimum Wage was 34 percent of the median wage and covered some 8.3 percent of employees (Metcalf, 1999). One year later, Ireland introduced a minimum wage at 59 percent of the median wage that affected about 21 percent of the workforce (Nolan/O'Neill/Williams, 2002). In 2008, Croatia's general minimum wage was set at 42 percent of the average wage, covering 9.2 percent of Croatian employees (Nestić/Babić/Blažević Burić, 2018). Treatment evaluations using survey data from these countries would likely suffer from structurally similar biases, sine about 13 percent of employees where directly affected by the 2015 minimum wage introduction in Germany, which is 48 percent of the median wage (German Minimum Wage Commission, 2016).[3]

[3] The OECD keeps track of minimum wages (and ratios) of its members starting in 2000: https://stats.oecd.org/Index.aspx?DataSetCode=MIN2AVE.

2. Classical measurement error in survey wages and treatment assignment

In a hypothetical evaluation study, we are interested in estimating the relationship between an outcome variable y and an independent treatment variable t. In the case of minimum wage evaluations, t is a treatment indicator, i.e., any person with a wage below the minimum wage receives the treatment. However, in survey data, there often is no direct measure of t. Instead, we observe a survey measure of gross hourly wages (w_i) for each individual i before the treatment event occurs. For studies using observable survey data, w_i is the sum of the actual gross hourly earnings (x_i) and an error term (m_i),

$$w_i = x_i + m_i. \tag{2.1}$$

In the case of classical measurement error, mi is characterized by zero mean $E(m) = 0$ and nonzero variance $E(m^2) = \sigma_m^2$, i.e., the measurement error term has a mean of zero and is assumed to be independently and identically distributed. Classical measurement error is easier to handle than other types of error, as there is no correlation between the error term and the independent variable of actual wages x_i. However, classical measurement error is only a special case of a more general survey measure of wages:

$$w_i = \underbrace{x_i}_{\text{actual wage}} \underbrace{+\ m_i}_{\text{classical measurement error}} \underbrace{+\ c}_{\text{nonzero constant}} \underbrace{+\ \rho(x_i - \mu_x)}_{\text{nonclassical term}} \tag{2.2}$$

In this broad description of additive measurement error, the observed survey measure of gross monthly wages (w_i) is the sum of actual gross monthly earnings x_i, the iid error term m_i, possibly a constant nonzero average measurement error c and an additional term accounting for dependence between the error term and actual earnings $(\rho(x_i - \mu_x))$. The latter can account for mean reversion, i.e., above average earners understating survey wages and below average earners overestimating survey wages. This effect would result in a negative correlation ρ, i.e., $x_i - \mu_x$ and w_i are negatively correlated conditional on x_i.

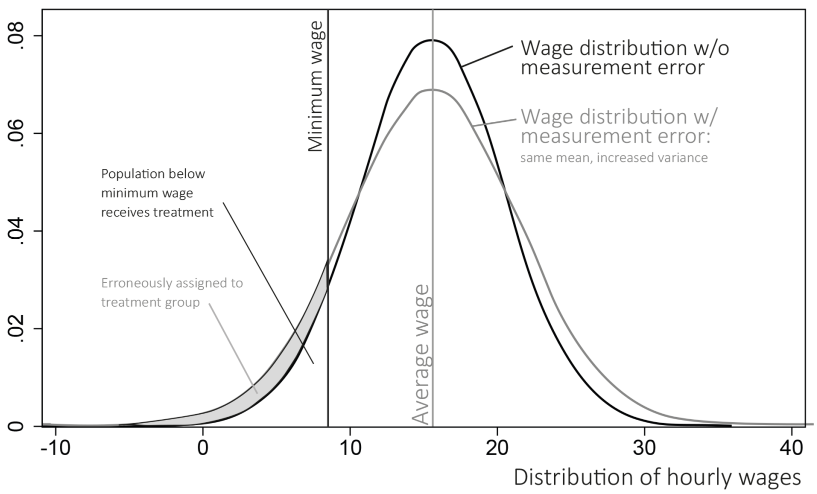
Consider the assumptions for x_i and m_i in the motivating example in the next section while excluding the presence of nonclassical measurement error components. In the first simulation, both variables are independently, identically and normally distributed random variables with $F_X \sim N(\mu_x, \sigma_x^2)$ and $F_M \sim N(0, \sigma_m^2)$. These characteristics implies that the measurement error in the survey wage variable is classical, i.e., measurement error m_i and monthly wage w_i are uncorrelated. Considering the properties of iid random variables, the observed wage is also iid normally distributed with $F_W \sim N(\mu_x, \sigma_x^2 + \sigma_m^2)$. Thus, in an additive model with classical measurement error, the observed wage distribution exhibits the same mean and an increased variance.

In an evaluation study, the treatment status of a person t_i might be determined by a threshold such as an individual's wage being below the minimum wage before its introduction. In this setting, the measurement error in the wage variable blurs the assignment of individuals to treatment and control groups. The case distinction

$$t_i = \begin{cases} 0 & \text{if } x_i \geq w_{min} \\ 1 & \text{if } x_i < w_{min} \end{cases} \tag{2.3}$$

determines the treatment status of each individual at a point in time ahead of the treatment event. However, x_i cannot be observed, and the contamination of w_i due to measurement error invariably translates to contamination of the treatment indicator t_i. Minimum wages are a special case in this setting because only a small fraction of the working population is typically affected by its introduction (i.e., we assume that $w_{min} < \mu_x$).

Figure 1.: Stylized wage distribution with measurement error.



Note: Assuming classical measurement error, the wage distribution with measurement error exhibits the same mean with increased variance compared to the earnings distribution without error terms. As theminimumwage typically is below the average wage, the fraction of treated workers is expected to increase.
Source: own illustration.

As illustrated in Figure 1, the measurement error-induced decompression of the wage distribution causes the observed number of persons below w_{min} to be higher than that without measurement error. While persons might be falsely attributed to the treatment and control groups in both directions, more individuals are expected to be false positives. This result follows from the cumulative density functions of the observed and actual wages. The expected actual fraction of treated workers (π_x) corresponds to the value for w_{min} of the cumulative distribution function for the actual earnings distribution:

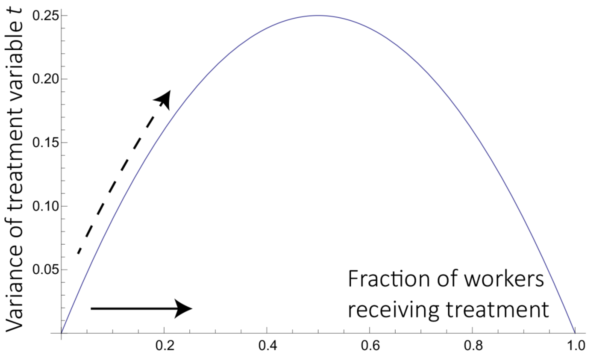
$$\pi_x := F_x(w_{min}) = \int_0^{w_{min}} \frac{1}{\sigma_x\sqrt{2\pi}} e^{-(x-\mu_x)^2/2\sigma^2} dx. \tag{2.4}$$

The measurement error-induced classification bias is then defined as

$$bias_\pi := F_W(w_{min}) - F_X(w_{min}). \tag{2.5}$$

The variance of the treatment variable $Var(t)$, which is a relevant statistic in OLS-based regression analyses, is a function of the fraction of individuals who receive the treatment. Since t is a binary variable, the variance is $Var(t) = \pi_x(1 - \pi_x)$. As positive bias is expected as long as w_{min} is smaller than μ_x, the contaminated variance of the treatment variable leads to increased variance (Figure 2). Under these conditions, as demonstrated in the next subsection, the measurement error in individual earnings causes the treatment effect $\hat{\beta}_m$ to be biased downwards.

Figure 2.: The variance of the treatment variable is a function of the fraction of the population receiving treatment.



Note: Variance of treatment variable $Var(t) = \pi_x(1-\pi_x)$. If the treatment bias $bias_\pi$ is positive, but a fraction well below 50 percent of the workers receive the treatment, the contaminated treatment variable exhibits a higher variance than before.
Source: own calculations and illustrations.

3. Consequences of treatment misclassification in person-level difference-in-differences estimation

Let us assume scholars are interested in estimating the effect of a uniform minimum wage introduction, such as the 2014-2015 German minimum wage introduction, from individual survey data. In this case, one might be interested in the effect of the treatment ($t = 1$) on an unspecified outcome variable y and might apply a difference-in-differences model with two waves of panel data[4] (before and after treatment, indexed by subscript t):

$$y_{it} = +t_i * \beta + post_t * \gamma + post_t * t_i * \delta + \epsilon_{it}, \tag{3.1}$$

where y_{it} is the outcome variable such as pay satisfaction, wages or employment. α is a constant, β captures the time-constant group effect of the treated individuals, γ captures a common time effect, and the treatment effect interaction of the treatment group and treatment time δ captures the treatment effect on the treated individuals.

An advantage of the simple two-wave difference-in-differences model is its simplification by taking first differences in equation 3.1:

$$y_i = \gamma + t_i * \delta + \epsilon_i. \tag{3.2}$$

The difference-in-differences specification in terms of first differences explains the change in the outcome between two survey waves Δy_i. The constant and time-constant group effect of specification 3.1 are canceled out by the first difference, as these are time-constant terms. The time effect (γ) then becomes the new constant, and the treatment effect δ is the coefficient of the treatment dummy. In the absence of measurement error, equation 3.2 can be estimated via OLS to obtain a consistent estimate of the true treatment effect.

For the sake of simplicity, $\Delta\epsilon_i$ is set to follow a standard normal distribution, and the effects

[4] Since the treatment indicator is typically defined at a specific point in time before the minimum wage reform, it is not influenced by the number of data waves for which the dependent variable is tracked. Hence, the simulation results remain unchanged and we can proceed with the simple case of only one wave of data before and one wave after the treatment.

sizes are fixed at $\gamma = 0$ and $\delta = 1$, i.e., the introduction of a minimum wage causes the dependent variable (e.g., pay satisfaction) to increase by 1 if $t = 1$. Thereby, we assume that the true treatment effect is homogeneous and uncorrelated with the measurement error m.

When we introduce measurement error, t_i is misclassified. In fact, the number of individuals classified as treated increases as the variance of the error term (σ_m^2) increases. An increasing fraction of individuals assigned to the treatment group implies an increasing variance of the treatment indicator (see Figure 2), which is the denominator of the OLS estimator:

$$\hat{\beta} = \frac{Cov(\Delta y_i, t_i)}{Var(t_i)} > \frac{Cov(\Delta y_i, t_i^{ME})}{Var(t_i^{ME})} = \hat{\beta}^{ME}. \tag{3.3}$$

Compared with that of the baseline, in which t_i is assigned without measurement error, in the presence of measurement error, the variance $Var(t_i^{ME})$ increases while the covariance $Cov(\Delta y_i, t_i^{ME})$ decreases. Hence, the overall effect of measurement error on the treatment effect is a bias towards zero.

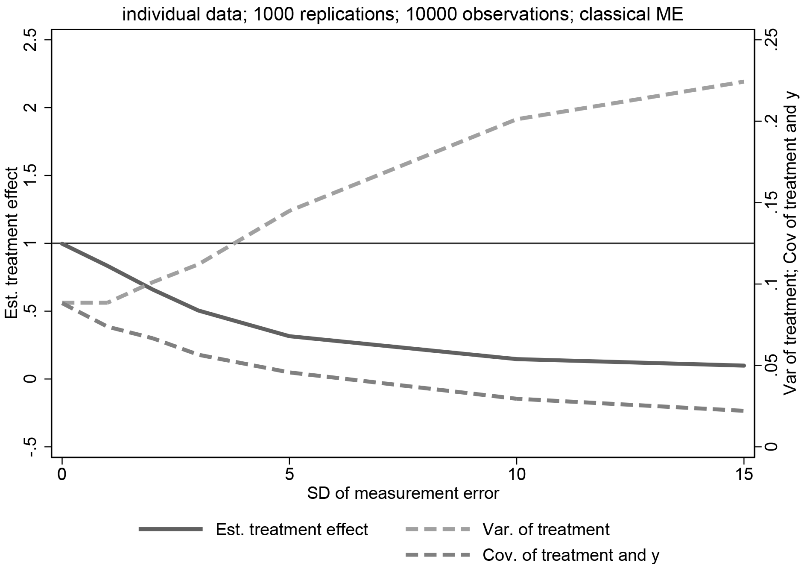
We test this conjecture in the first Monte Carlo experiment. Figure 3 depicts the median estimated treatment effects, the variance of the treatment variable and the covariance between the treatment and outcome variable for iid normally distributed wages $F_x \sim N(15, 5)$ and error terms with increasing standard deviation σ_m^2. With $F_m \sim N(0, 2.5)$, the expected fraction of workers subject to the treatment is 19.3 percent, up from 9.7 percent in the baseline (corresponding to a treatment assignment bias of 9.6 percentage points), increasing the respective variance of the treatment variable t from roughly 0.09 to 0.16.

Figure 3 demonstrates that the treatment effect decreases (solid line) as the variance of the measurement error σ_m increases, mostly due to the increasing variance $Var(t_i^{ME})$ depicted by the dark dashed line and the decreasing covariance $Cov(\Delta y_i, t_i)$ depicted by the light dashed line. This bias holds in various robustness checks of the simulation using a log normal wage distribution, an empirical wage distribution from administrative data (see section 7), different sample sizes, and different baseline values for γ and δ. Hence, we conclude that a fuzzy treatment assignment due to measurement error in wages – from which the threshold-dependent treatment status is inferred – leads to a bias towards zero.

Adding a nonclassical error term on the individual level

We repeat the simulation presented in Figure 3 and introduce an additional error term to the measurement of the individual survey wage data. The additional error term is a nonclassical term (see Eq. 2.2) that accounts for possible correlation between gross wages and the error

Figure 3.: Estimated treatment effect from individual data with increasing measurement error.

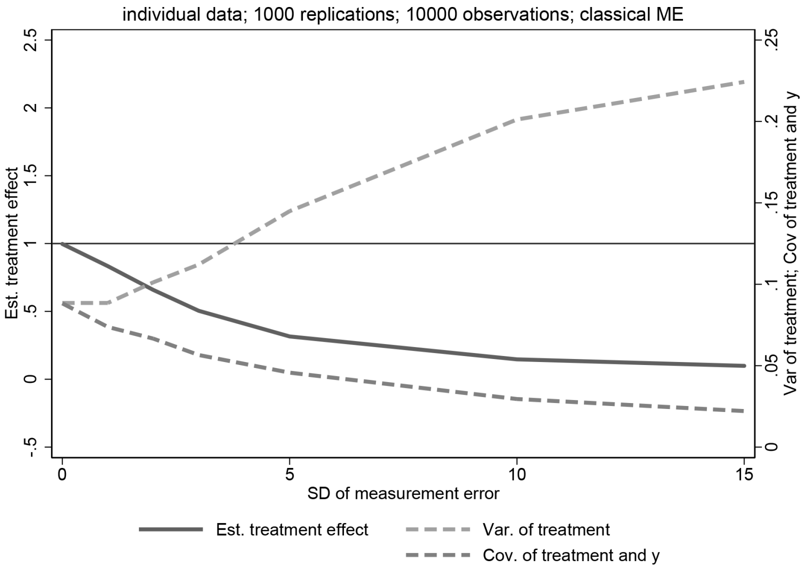


Note: Estimated treatment effect from individual data with increasing variance of the classical measurement error on the x-axis. Right axis: The variance of the contaminated treatment effect increases with increasing σ_m (standard deviation of the error term); the covariance between the treatment and unspecified outcome variable decreases with increasing σ_m. Left axis: Overall, the median estimated treatment effect for $F_x \sim N(15, 5)$ and $F_m \sim N(0, \sigma_m^2)$ decreases with increasing σ_m^2.
Source: own calculations and illustrations.

terms. While a negative correlation would have mean-reverting effect, a positive correlation would increase the total measurement error. We use different correlations ρ ranging between -0.3 and +0.3 and repeat the Monte Carlo experiment with an additional classical error term that is distributed $F_m \sim N(0, \sigma_m^2)$, where σ_m^2 ranges between zero and five.

Figure 4 depicts median estimated treatment effects after 1000 repetitions of the simulation, where the level of ρ is on the x-axis. Considering a correlation of ρ as high as ±0.2 would be extremely unusual, it is safe to say that the properties of the classical error terms are a much greater threat to the reliable estimation of treatment effects than a nonclassical term. We find that a nonclassical error structure would somewhat increase the bias of the estimation, if ρ is profoundly negative and σ_m is large. However, the overall estimation bias is driven by the standard deviation σ_m of the classical error distribution, with ρ serving to only slightly increase or decrease the bias. In summary, in an individual-level treatment effect estimation, a nonclassical error structure poses a considerably smaller threat to unbiasedness than does the variance of the error term distribution.

Figure 4.: Variance and covariance of the estimation variables.



Note: In an individual level DiD estimation with nonclassical error structure, Eq.(2), the treatment effect decreases with increasing σ_m (standard deviation of the error term), and the covariance between treatment and unspecified outcome variable decreases with increasing σ_m. Left axis: Overall, the median estimated treatment effect for $F_x \sim N(15, 5)$ and $F_m \sim N(0, \sigma_m^2)$ decreases for increasing σ_m.
Source: own calculations and illustrations.

4. Assessing the effect of data aggregation to higher-level units

In this section, we analyze whether the treatment effect dilution bias is alleviated through data aggregation to higher-level units. In applied research, such higher-level units could be households, establishments, or regions. When using aggregated data (or grouped data), one might hope that the measurement error is canceled out by averaging over the within-group measurement error. We first discuss that data aggregation requires a definition of the unit sizes and a rule for allocating individual observations to units, and we estimate how the bias evolves with different unit sizes and levels of segregation in the allocation rule (Section 4.1). In a second step, we use realistic unit size distributions and allocation rules to simulate the treatment effect for realistic cases of data aggregation. In Section 4.2 we scrutinize these results using joint distributions of unit sizes and segregation from empirical data.

4.1. Conceptual treatment effect bias of aggregation

The aggregation of data to higher-level units is highly prevalent in the literature and is based on the identification goal of limiting the potential for spillovers within these units (Card, 1992; Caliendo et al., 2018). In the absence of such spillovers, i.e., without correlation within units, aggregated regression yields an unbiased estimate for the disaggregated model of interest. Hence, population-weighted estimation of equation 4.1 yields unbiased estimates of γ and δ, as do equations 3.1 and 3.2.

$$\overline{\Delta y}_u = \gamma + t_u * \delta + \epsilon_u, \tag{4.1}$$

where u is the unit subscript that determines the level of aggregation. $\overline{\Delta y}_u$ is the average change in y among observations in unit u, and t_u is the fraction (average) of treated individuals in unit u.

However, it is not a priori clear how individuals are allocated to higher-level units; hence, we use different rules of allocation with varying extents of sorting of low-wage individuals into low-wage units. In reality, workers are not randomly allocated to households, establishments and regions. In fact, the literature documents an increasing segregation of low-wage workers to low-wage plants for Germany over the past two decades (Card/Heining/Kline, 2013).

For different degrees of segregation, we test how sorting affects potential biases in the measurement error in aggregated treatment effect estimations. Assuming no segregation, for instance, implies that workers are randomly allocated to units (e.g. firms), i.e., on average, the units themselves are similarly affected by a treatment such as the minimum wage introduction, as differences between units are random. This characteristic implies that the fraction of affected employees t_u across units is roughly the same. However, if we assume complete segregation, some firms employ solely workers who are affected by the treatment, while other firms – which constitute the majority, since w_{min} is smaller than μ_x – do not employ any affected workers. After aggregation at the establishment level the treatment variable is a function of both segregation and measurement error.

To measure segregation, we use the normalized Herfindahl-Hirschman Index (HHI):

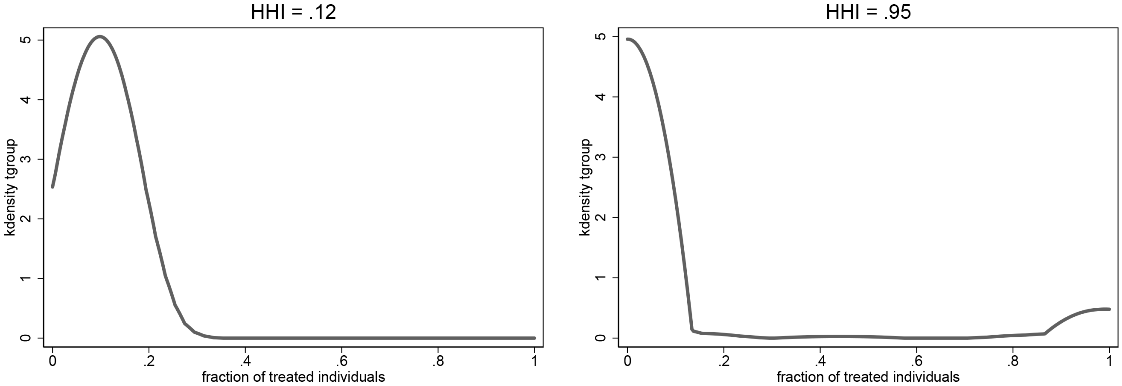
$$HHI = \frac{1}{U}\sum_{u=1}^{U}\frac{t_u^2}{t_u}, \tag{4.2}$$

where U is the total number of u units of aggregation and t_u is the fraction of treated individuals in u. We prefer the HHI over an index of dissimilarity, as it is multiplicative and hence linear in the variance and covariance, which is beneficial in terms of the components of the OLS estimator.

Figure 5 illustrates the aggregated treatment variable densities before adding measurement error to the wages for different level of segregation, as measured by the normalized HHI. Figure 5 demonstrates that the levels of segregation largely determine the variance of the aggregated treatment variable. In fully segregated markets, the aggregated treatment variable takes only the values $t_u = 1$ or $t_u = 0$, maximizing the variance $Var(t_u)$. By contrast, when individuals are randomly allocated across establishments, the treatment variable of establishments is distributed around μ_t, resulting in a very low variance $Var(t_u)$. After adding classical measurement error to the underlying individual wage data, the aggregated variances change in different ways: the variance of the aggregated treatment variable is likely to decrease in the fully segregated scenario (as it was already at its maximum in the absence of measurement error) but in the case of random allocation, the variance of the aggregated treatment variable is unlikely to change substantially. Hence, the illustration of the aggregated unit-level treatment variables indicate different effects of measurement error for different magnitudes of segregation.

To assess the treatment effect bias, we again conduct a series of Monte Carlo experiments. We estimate the aggregated treatment effect model as specified in equation 4.1. The individual data are as before; hence, the aggregated (baseline) grouped data regression without measurement error yields unbiased estimates of $\gamma = 0$ and $\delta = 1$ (Angrist/Pischke, 2008;

Figure 5.: Densities of aggregated treatment variables with 50 workers per unit.
Random allocation (le panel) and highly segregated allocation (right panel)).



Note: Densities of aggregated treatment variables, where the level of aggregation is 50 workers per unit. Since the original data included 10,000 workers, the aggregated data include 200 units. The aggregation uses different aggregation rules w.r.t. segregation. The level of segregation is calculated from the normalized Herfindahl-Hirschman Index (HHI).
Source: own calculations and illustrations.

Prais/Aitchison, 1954). To illustrate how measurement error affects these estimates, we conduct a simulation and inflict classical measurement error m_{ti} on the individual wage data, which is distributed $F_m \sim N(0, 2.5)$. To demonstrate how different aggregation rules affect the average treatment effect bias, Table 1 displays the difference between the median estimates and actual treatment effect ($\overline{\hat{\delta}_{w=x+m}} - \overline{\hat{\delta}_{w=x}}$) for different levels of segregation measured by the normalized HHI. We repeat the simulation for different unit sizes, but for simplicity, we hold the size of the units constant within each of these simulations.[5]

The results in Table 1 demonstrate a strong downward bias when the level of aggregation is small, i.e., when the observed number of employees per firm or region is small. Moreover, the bias tends to be negative when the level of segregation is low, i.e., when the allocation of low-wage workers to higher-level units is random. Most remarkably, we observe strong positive biases for highly segregated markets (indicated by high HHIs), especially when the aggregation unit size is large.

These patterns of treatment effect biases can be explained by separate simulations of the variance and the covariance, which both determine the treatment effect of the linear regression, as specified in equation 4.1, i.e., $\hat{\delta} = \frac{Cov(\Delta y_u, t_u)}{Var(t_u)}$. Figure 13 displays separate estimates of the variances ($Var(t_u)$) as solid lines and estimates of the covariances ($Cov(\Delta y_u, t_u)$) as dashed lines for varying levels of segregation and for four different unit sizes. The blue lines depict

[5] Holding the size of the units constant within each regression avoids the need to conduct weighted regression because each unit u is given the same weight.

Table 1.: Simulation results for aggregated data for different unit sizes and levels of segregation.

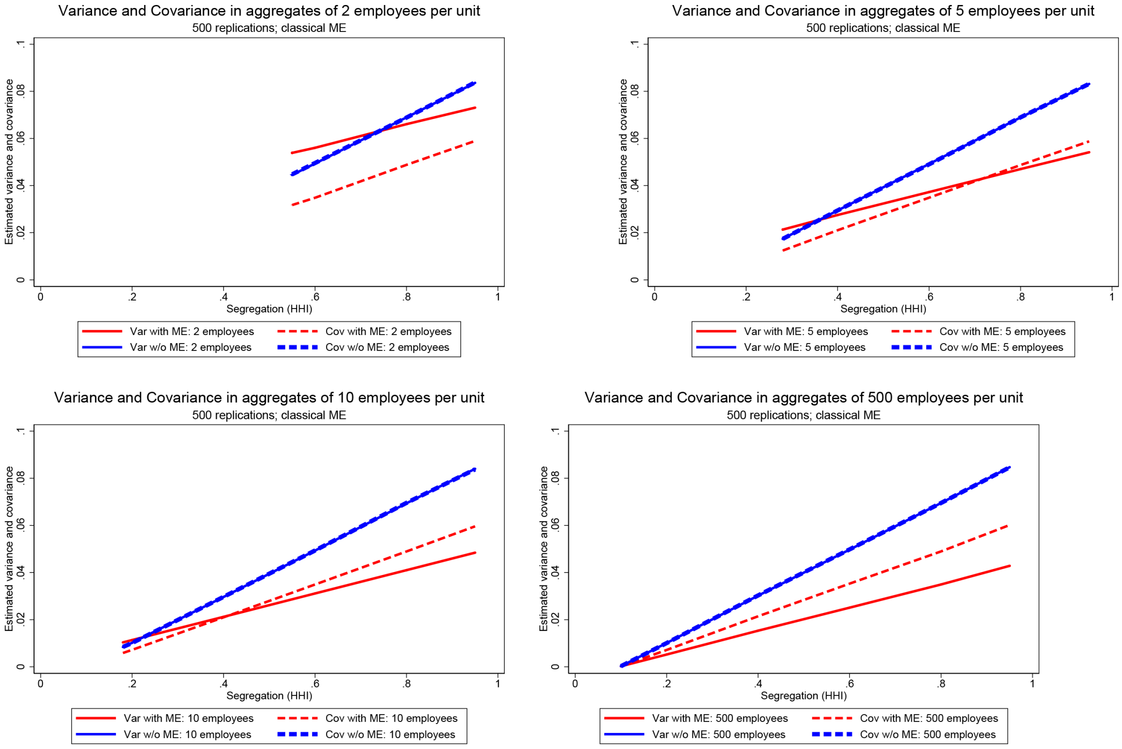
	Employees per unit					
	5		10		50	
Concentration (HHI)	Bias	p95-p5	Bias	p95-p5	Bias	p95-p5
random allocation, which yields different HHIs	-0.42	0.34	-0.44	0.44	-0.41	1.07
0.2	n/a	n/a	-0.37	0.45	0.14	0.60
0.4	-0.23	0.31	-0.01	0.34	0.29	0.40
0.6	-0.07	0.26	0.12	0.28	0.34	0.32
0.8	0.03	0.23	0.19	0.24	0.37	0.29
0.95	0.08	0.21	0.23	0.25	0.37	0.25
	Employees per unit					
	100		20		500	
Concentration (HHI)	Bias	p95-p5	Bias	p95-p5	Bias	p95-p5
random allocation, which yields different HHIs	-0.39	1.50	-0.40	1.87	-0.34	3.23
0.2	0.28	0.67	0.33	0.70	0.36	0.72
0.4	0.35	0.41	0.38	0.46	0.40	0.54
0.6	0.38	0.34	0.39	0.34	0.40	0.46
0.8	0.39	0.30	0.40	0.30	0.40	0.45
0.95	0.39	0.25	0.40	0.27	0.40	1.45

Notes: This table shows the bias of the median treatment estimates, i.e., the absolute difference between the estimated treatment effect and the predefined treatment effect (= 1). The bias is displayed for different levels of segregation and for varying firm sizes. Segregation is measured by the HHI concentration index, which is the normalized Herfindahl-Hirschman Index. Smaller values of HHI translate to less segregation. p95-p5 is the range between the 5th and 95th percentile of the estimated treatment effects.
[a] Random allocation yields an HHI=0.28 for 5 employees per unit, HHI=0.18 for 10 employees, HHI=0.12 for 50 employees, HHI=0.11 for 100 employees, HHI=0.1 for 200 employees, and HHI=0.1 for 500 employees.
[b] In very small units with as few as 5 employees, an allocation that yields an HHI=0.2 is infeasible.
Source: own calculations.

the respective baseline without measurement error, and the red lines include measurement error, as in Table 1, i.e., $F_m \sim N(0, 2.5)$.

In the absence of measurement error in individual wages (blue lines), the estimates of the variance and the covariance are equivalent for different levels of segregation and across unit sizes of aggregation, leading to an unbiased estimator of the true treatment effect, which is uniform by design. Hence, segregation and levels of aggregation do not matter in an ideal world without measurement error. In the presence of measurement error, the estimated covariances and variances differ from those of an ideal world. The covariance is estimated to be smaller than the unbiased variance, independently of the unit size of aggregation. The estimated variance is even smaller than the covariance when the unit size is large, leading to an upward bias in the treatment effect. However, the estimated variance is larger than the covariance when the unit size is small, leading to a downward bias in the OLS estimate. The relatively larger variance at small unit sizes is most likely due to a corner solution problem: the variance can change only slowly in the presence of measurement error simply because the aggregated treatment variable t_u can be characterized by very few distinct fractions of

Figure 6.: Estimated variances and covariances from Monte Carlo experiments for four different unit sizes of aggregation.



Note: Estimated variances and covariances from Monte Carlo experiments for four different unit sizes of aggregation: 2 employees per unit, 5 employees per unit, 10 employees per unit, and 500 employees per unit. Varying levels of segregation are on the x-axis, where the level of segregation is quantified by the Herfindahl-Hirschman Index (HHI). Solid lines display estimated variances of the treatment variable tu, and dashed lines display estimated covariances of the treatment variable t_u and Δy_u. Blue lines are estimates without measurement error, and red lines display estimates with measurement error.
Source: own calculations and illustrations.

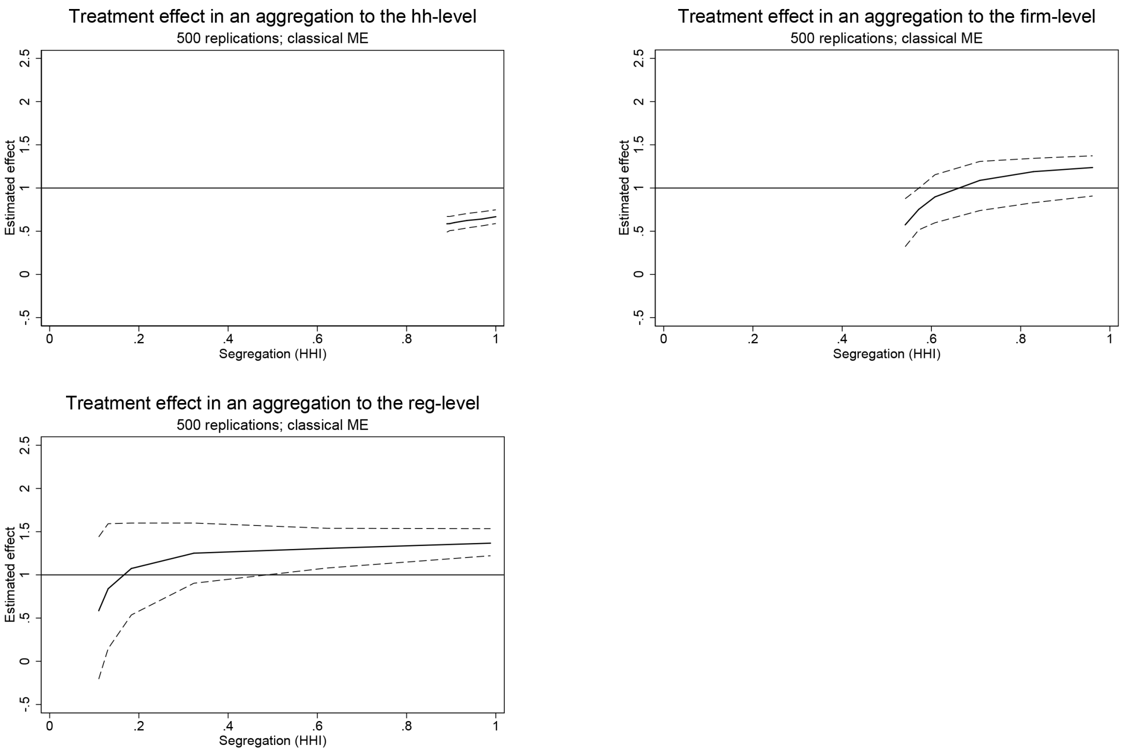
treated individuals when the aggregation units are small. By contrast, the variance $Var(t_u)$ responds more quickly to measurement error when unit sizes are large.

4.2. Aggregation to observed unit size distributions

Since we observe very heterogeneous biases in the aggregated regression when there is misclassification in the underlying wage data, we want to elaborate how such measurement error-induced biases would affect evaluation studies that use empirically relevant levels of data aggregation. In our artificially constructed data, the size and the direction of the bias heavily depend on the size of the higher-level aggregation units and the segregation of treated workers across those units. To transfer the results to actual empirical circumstances, we first infer realistic size distributions for households, establishments and regions from observable data. These unit size distributions allow us to draw conclusions about the treatment effect bias in more realistic scenarios.

The first scenario is aggregation at the household level, for which we can retrieve a realistic size distribution from the German panel study Labour Market and Social Security (PASS). The second scenario is aggregation at the establishment level, for which we retrieve a realistic firm size distribution from the German Establishment History Panel (BHP). The third scenario is aggregation at the region level. For this last scenario, we assign the underlying individual data to regions based on the region size in the BHP.[6]

Figure 7.: Estimated treatment effect bias for the scenarios summarized in Table 4.



Source: own calculations and illustrations.

Given these unit size distributions, we assess the treatment effect bias of the measurement error in the underlying wage distribution from the Monte Carlo experiments. The simulations are designed as in Table 1 to allow for different levels of segregation of treated individuals across units. The results are displayed in Figure 7. When considering the aggregated treatment effects at the household level, the difference-in-differences estimation yields a downward bias in the treatment effect irrespective of the level of segregation. However, with high levels of segregation across households, this bias decreases slightly.[7] Given households are

[6] A detailed description of how we draw the unit size distributions from the observable data sources is provided in Appendix A.

[7] Note that the HHI, a measure of segregation, does not take values below 0.89 even though the respective allocation of individuals is designed to be random. This is because the potential for an egalitarian distribution of treated workers across households is limited since households are small, on average, and many households have only one individual in the workforce.

increasingly segregated due to marital sorting (Pestel, 2017), the bias from measurement error is slightly reduced over time.

In the establishment-level analysis, the treatment effects are underestimated if individuals are randomly allocated across establishments, but we observe overestimation if the labor market is highly segregated across firms. In the region-level analysis, the treatment effect bias is positive for all levels except unrealistically low levels of segregation. Since regions are typically much less segregated than households or establishments, in practice, the bias is possibly small. However, the treatment effect bias also shows large confidence bands in cases of low segregation, demonstrating a higher degree of uncertainty when using large aggregation units.

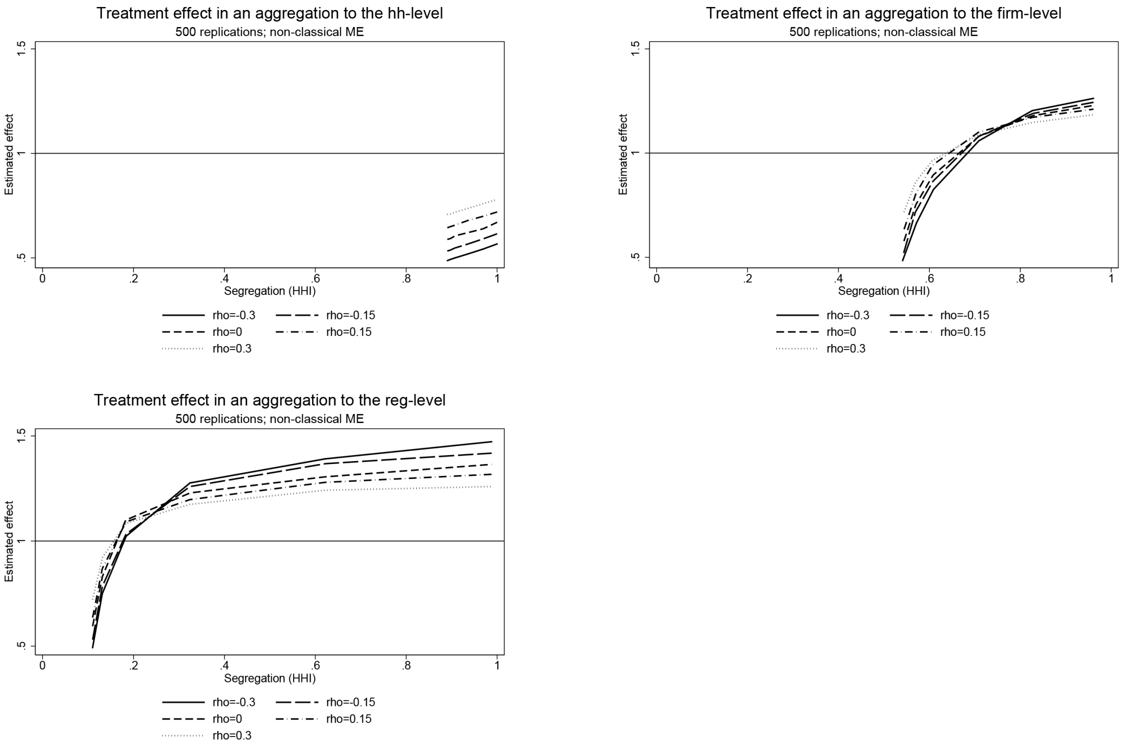
As in Section 3, we repeat the simulations presented in Figure 7 and introduce an additional nonclassical error term to the measurement of the individual survey wage data, allowing us to account for correlation between gross wages and the error term. To illustrate the additional effect of introducing a nonclassical error term determined by correlation ρ, we repeat the Monte Carlo experiment with a normally distributed error term with standard deviations of the classical term in $F_m \sim N(0, \sigma_m^2 = 2.5)$ and add the nonclassical term $\rho(x_i - \mu_x)$ for various ρ before aggregating to household, firm, and region unit sizes. As before, the x-axis shows possible values of segregation for observed unit size distributions (see Table 4).

Figure 8 depicts the median estimated treatment effects after 500 repetitions of the simulation for different values of ρ and varying levels of segregation among units. For high levels of segregation within households, the simulation without the nonclassical error term shows that the treatment effect bias after aggregation is similar to that for individual-level data. Accordingly, adding a nonclassical error term slightly increases (decreases) the median bias in the simulation if ρ is negative (positive). However, as ρ is typically small, the overall bias is driven by σ_m.

On the firm level, with much larger unit sizes, the effect of adding a nonclassical error term to the survey wages has a minimal effect for medium levels of segregation. For very high and very low levels of segregation, the bias decreases for unusually high values of ρ. Again, the real issues when estimating treatment effects with firm-level aggregation are segregation and the distribution of error terms, rather than a nonclassical error structure.

At the regional level, the impact of correlation between survey wages and error terms is small for small levels of segregation. In Table 4, using PASS data, we calculated a realistic value of 0.40 for segregation into regions of Germany. As before, a large negative ρ would slightly increase the bias, whereas a positive ρ would decrease the bias. Overall, and in accordance with

Figure 8.: Estimated treatment effect bias for various scenarios.



Note: Estimated treatment effect bias for the scenarios summarized in Table 4 with nonclassical error term ρ (Eq. 2.2).
Source: own calculations and illustrations.

individual-level survey data, after data aggregation, the effect of a nonclassical measurement error in the underlying wage data remains small compared to the impact of segregation.

5. Using a continuous treatment variable instead of a dummy

Until this section, we implicitly assumed that an increase in wages due to a reform yields a homogeneous increase (or decrease) of a dependent variable. For instance, every person affected by a newly introduced minimum wage experiences an increase in their pay satisfaction by a fixed amount. This is an implicit assumption concerning the data generating process, which results from including a treatment dummy instead of a continuous treatment indicator in the regression model. In most recent applications, researchers have chosen similar models (i.e. Bossler/Broszeit, 2017; Caliendo et al., 2019), thereby neglecting the possibility that individuals might be affected heterogeneously depending on the treatment intensity.

If the effect of a treatment is assumed to be dependent on the treatment intensity, we must rewrite the model to account for a continuous treatment variable. We choose a simple model in which we incorporate the fraction of the wage that can be explained by a treatment. In this case, the treatment variable is given by

$$t_i^{continuous} = \begin{cases} 0 & \text{if } x_i \geq w_{min} \\ \frac{w_{min} - x_i}{w_{min}} & \text{if } x_i < w_{min} \end{cases}$$

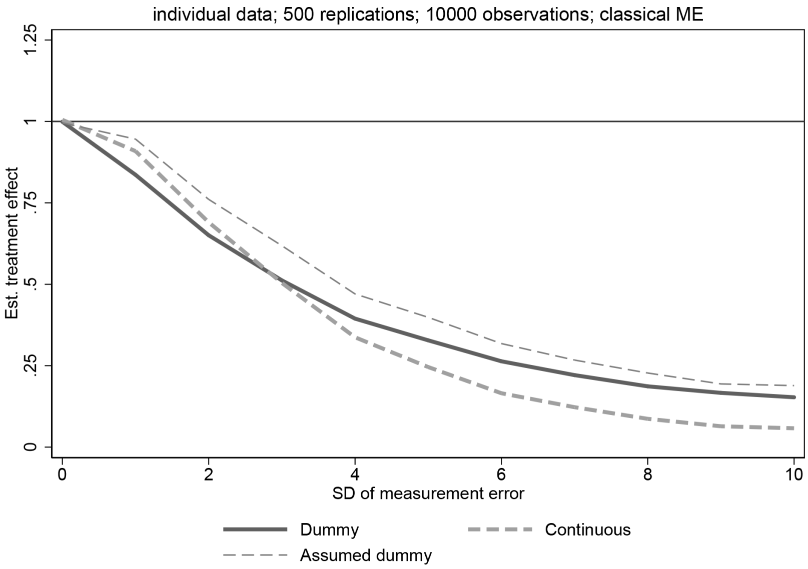
The data generating process follows from the continuous treatment variable. In our model, the change in the dependent variable is equal to the treatment intensity

$$\Delta y_i = t_i^{continuous} + \Delta\epsilon_i.$$

Figure 9 shows the known treatment dummy specification (Dummy) and the alternative continuous treatment variable specification (Continuous) for individual wage data. The use of a continuous indicator instead of a dummy for the individual-level model appears to be advantageous for moderately contaminated data (SD of the error terms less than 2 for the simulated wage data). For more extensive contamination, it is preferable to assume a uniform treatment effect and resort to a dummy variable in the model specification (if feasible).

Figure 9 includes another specification, accounting for the possibility that the assumed data generating process and the actual process are different. In this second case, we assume a

Figure 9.: The variance of the contaminated treatment effect increases with increasing σ_m.



Note: Right axis: The variance of the contaminated treatment effect increases with increasing σ_m (standard deviation of the error term), and the covariance between the treatment and unspecified outcome variable decreases with increasing σ_m. Left axis: Overall, the median estimated treatment effect for $F_x \sim N(15, 5)$ and $F_m \sim N(0, \sigma_m^2)$ decreases with increasing σ_m.
Source: own calculations and illustrations.

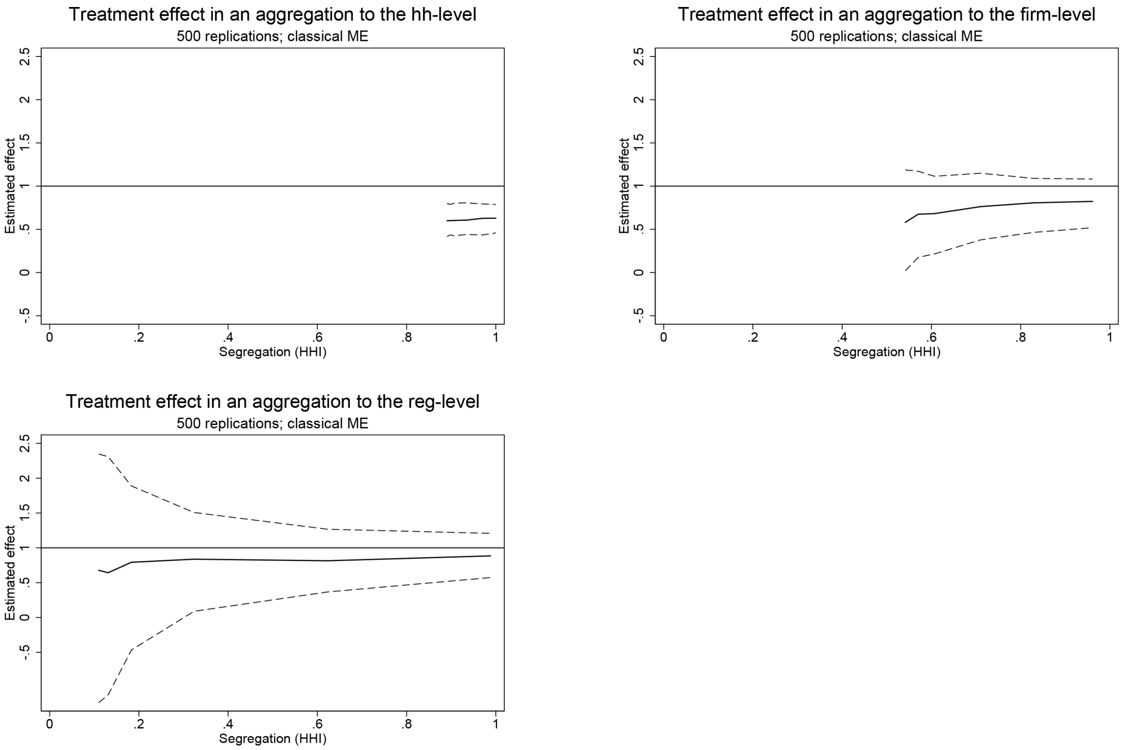
uniform effect (Assumed dummy), while the effect in the data generating process actually differs by intensity. We scale the results to match the treatment effect of one, as we would not be able to compare the results directly otherwise. Interestingly, the patterns of bias appear to be similar for all specifications.

Overall, on the individual level, the expected improvements are limited when using a continuous treatment variable. While the bias is slightly reduced for low to moderate contamination, the dummy variable estimator is under real-world conditions more efficient than the continuous variable estimator.

In Figure 10, we show the results of a simulation in which we use a continuous treatment variable with aggregated data. As before, we aggregate at the household, firm and region levels with varying levels of segregation of minimum wage workers into higher-level units. After aggregating, the size of the bias depends on the error term distribution only and not on the level of segregation. Consequently, a positive bias induced by high segregation within higher-level units is not within the realm of possibility in the specification with a continuous treatment variable.

The specifications yield different outcomes because the variances of the two treatment variables t_u and $t^{continuous}$ are affected differently. For a treatment dummy, the OLS estimator after aggregation is given by $\hat{\beta} = \frac{Cov(\Delta\overline{y}_i, t_i^{ME})}{Var(t_i^{ME})}$, with t_i^{ME} being the fraction of treated individu-

Figure 10.: Using a continuous treatment variable.



Note: Estimated treatment effect bias for the scenarios summarized in Table 4 using the continuous treatment variable instead of the dummy treatment.
Source: own calculations and illustrations.

als within unit i (e.g., the fraction of treated personnel in an establishment). Adding measurement error to individuals in highly segregated units makes those units more homogeneous as the fraction of treated individuals decreases in firms with very high treatment quotas and for firms with no or low treatment quotas, i.e., $Var(t_i^{ME})$ decreases, while $\hat{\beta}$ increases.

For a continuous treatment variable, the OLS estimator after aggregation is defined similarly as $\hat{\beta} = \frac{Cov(\Delta\bar{y}, t_i^{ME,continuous})}{Var(t_i^{ME,continuous})}$. For a continuous treatment variable, the indicator $t_i^{ME,continuous}$ is not a fraction but rather the average treatment intensity within a unit i. In this case, changes to $Var(t_i^{ME,continuous})$ induced by measurement error and its homogenization are mostly negligible, no matter the level of segregation.

Overall and independently of segregation, for aggregated data with moderate individual-level contamination, we recommend an approach using a continuous treatment variable, if this assumption is justifiable with respect to the underlying data generating process. However, as with individual-level data, aggregating at the region level might not be viable due to the loss of efficiency.

In Appendix B we briefly discuss aggregation to higher-level units with a treatment dummy specification, when the data-generating process would necessitate a continuous treatment variable specification (see Figure 14).

6. Dropping observations around the minimum wage threshold

In this section, we simulate the treatment effects using a very simple approach to address the measurement error, which is to drop observations around the minimum wage threshold. The intuition of this approach is to drop the observations that are most likely misclassified due to the measurement error in wages (Caliendo et al., 2019). Deleting observations around the threshold may therefore reduce bias but may possibly decrease the efficiency, as the number of observations decreases.

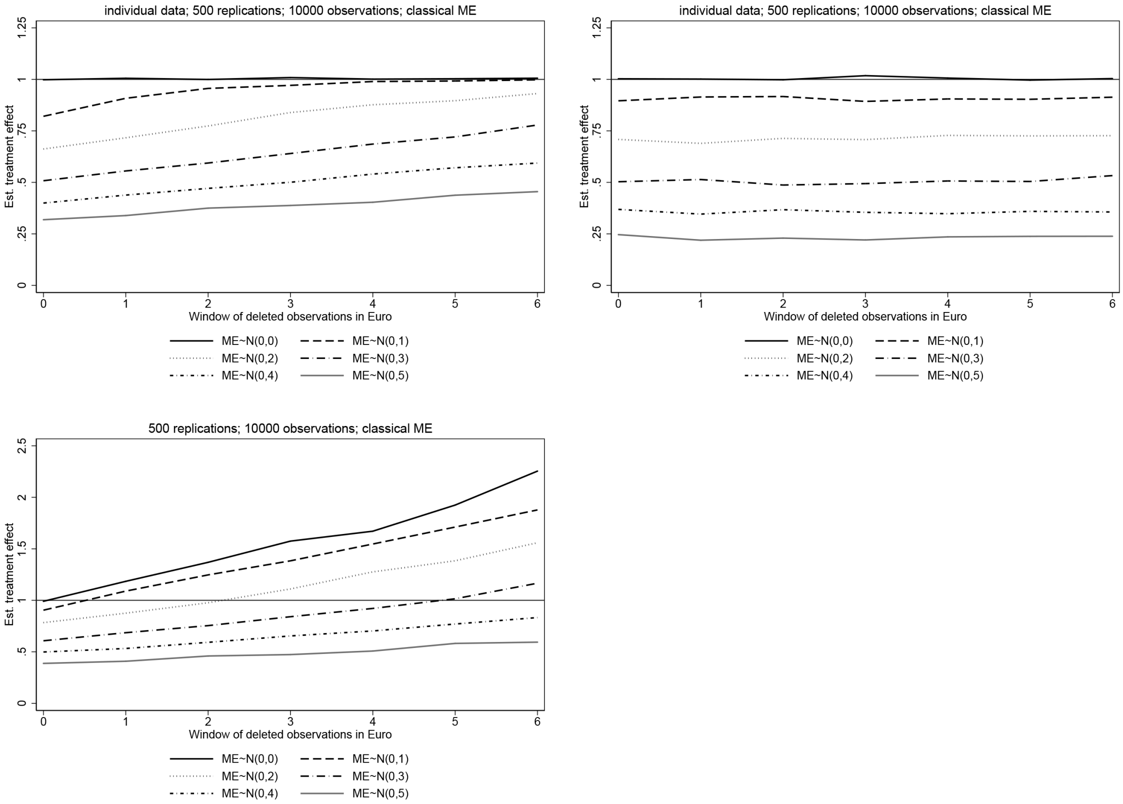
In our simulation, we delete a window of EUR 0 to EUR 6 around the minimum wage threshold and repeat the treatment assignment and treatment effect estimation. If a €6-window is deleted, all individuals with an initial wage between EUR 5.50 and EUR 11.50 are dropped. As the first step, we repeat the individual-level estimation and present the results in Figure 11. Different sizes of measurement error are depicted by different lines, and the size of the Eurowindow of deleted observations is displayed on the x-axis. The three graphs illustrate the three possibilities for defining the treatment variable along with the data generating process, as in previous Section 5: treatment dummy, continuous treatment variable and assumed dummy (i.e., we assume a dummy suffices, while the effect actually depends on the intensity).

In the case of the dummy treatment, the simulations show an unbiased treatment effect in the absence of measurement error, which does not change when individuals around the threshold are dropped. With increasing variance of the classical measurement error, the initial bias increases, as demonstrated by the treatment effect (at zero on the x-axis when no observations are deleted). When we drop individuals around the minimum wage threshold, the bias decreases as the number of dropped individuals increases. Dropping observations does not help to reduce the bias if the measurement error is very pronounced.

In the second case of a continuous treatment variable and a continuous data generating process, the bias increases with the size of the measurement error. However, dropping observations around the threshold does not help to reduce this bias. The intuition of this result is as follows: individuals further from the threshold still remain in the sample and have the same size of measurement error in the treatment variable as before. Hence, the bias remains the same size.

In the third case of an assumed dummy treatment, the bias increases as the variance of the measurement error increases. When observations around the threshold are dropped, the treat-

Figure 11.: Dropping observations with individual level data.
Dummy treatment (top right, continuous treatment (top left), and assumed dummy (bottom).



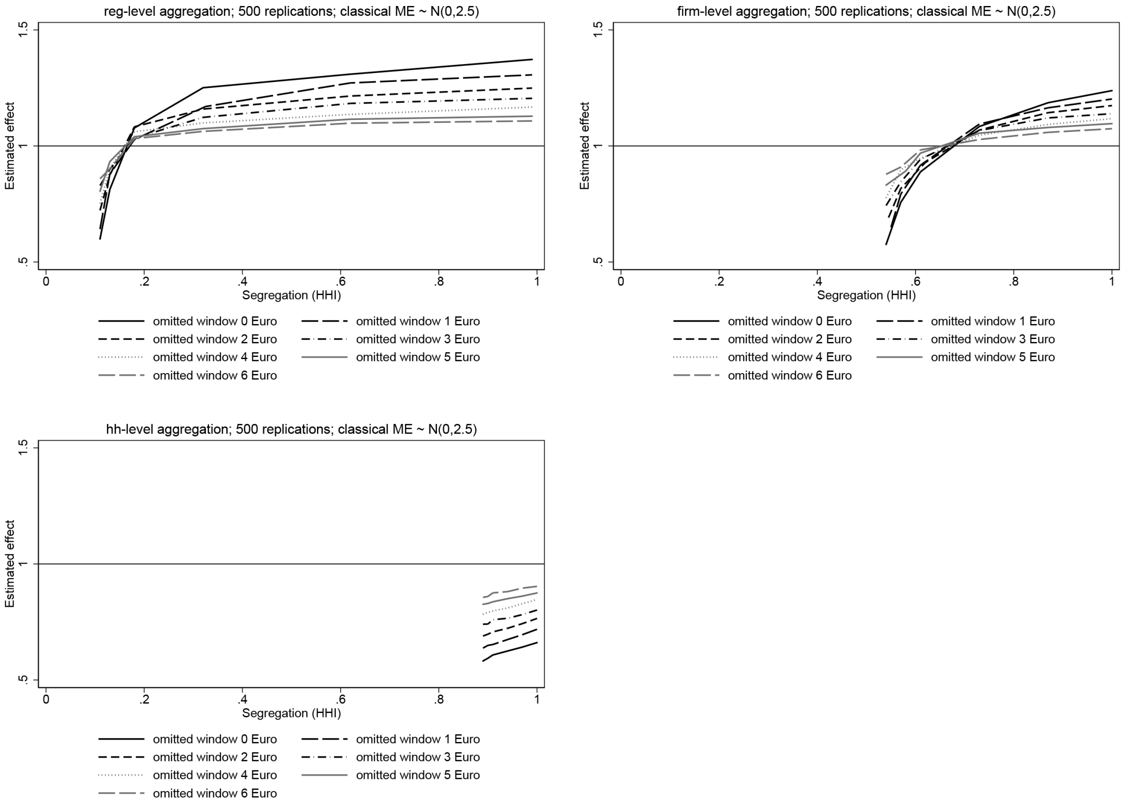
Note: Estimated individual-level treatment effect when individuals whose wages are near the threshold are dropped. The size of the Euro-window is indicated by the x-axis. Different lines illustrate various sizes of classical measurement error.
Source: own calculations and illustrations.

ment effects scale up because the group that is defined as treated comprises an increasing number of more intensely treated individuals who show a relatively stronger treatment effect. Hence, the changing composition of the treatment group scales-up the coefficient rather than providing a solution for the misclassification.

Note that in all three cases, the variances of the estimates barely increase, as the loss of efficiency is only minor. However, dropping observations may be considered to be a somewhat drastic solution in the first case, in which the data generating process and the treatment effect are assumed to be homogeneous across individuals as captured by the dummy variable. This finding also translates to the treatment effect estimation with aggregated data. Hence, we repeat the simulation for aggregated data for only this particular case using a dummy variable specification.

Figure 12 shows the estimation results when individuals around the threshold are dropped (as depicted by the different lines), and the estimation is then performed using aggregated

Figure 12.: Dropping observations with aggregated data.
Regional level (top left), firm level (top right), and household level (bottom).



Note: Estimated treatment effect of aggregated data when individuals are assigned the dummy treatment and individuals located near the threshold are dropped. The size of the Euro-window that is dropped is indicated by the x-axis. Different lines illustrate various sizes of classical measurement error.
Source: own calculations and illustrations.

data. Irrespective of the unit of aggregation, indicated by the three graphs, and the level of segregation of individuals across aggregation units, the bias generally decreases when the most contaminated observations are dropped.

Nevertheless, we urge researchers to be careful when dropping observations around the threshold. First, dropping observations is a solution only when there is a level shift at the threshold, i.e.. a homogeneous treatment effect. Second, in practice, dropping observations increases the potential of diverging trends in the difference-in-differences analysis because it yields a comparison of more dissimilar treatment and control groups by excluding observations at the point of intersection. Third, the individuals should be dropped when calculating the bite but also when running the estimation. In regressions applied to aggregated data, in principle, it is possible to calculate the bite from a sample that drops individuals around the threshold while including these individuals in the estimation. Nevertheless, such an approach can result in additional biases since the bite variable does not match the groups of comparison included in the regression.

7. Empirical distribution of wages and error terms: Validating survey data using administrative data

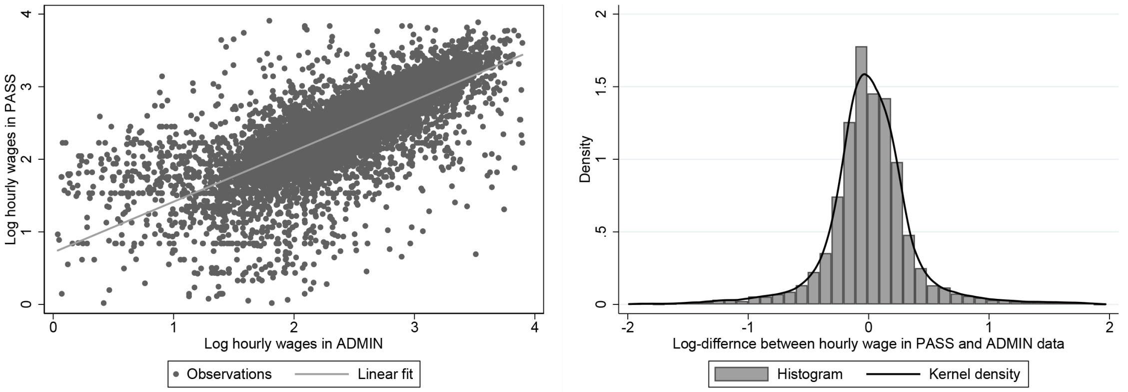
We continue our analysis by exploiting empirical distributions of wages and error terms from observable data sources. We use survey data from the PASS data that have been linked with German administrative data.[8] These data enable us to compare actual gross hourly wages (administrative data) with observed gross hourly wages (survey data), while the survey data provide information on the household composition and region for all individuals.[9] The direct link between survey and administrative data allows us to infer a realistic wage distribution and a realistic distribution of the measurement error, which we define as the difference between the two sources of data. Since the wage distribution and the distribution of measurement error may deviate from a normal distribution, this exercise serves as a check of the results under real-world conditions.

In addition to the commonly known advantage that survey data can be matched with social security data to compare gross monthly wages, we observe the respective distribution of (contractual) working hours from both sources, which enables us to eliminate possible bias from incorrect collection of working hours in the survey data. The administrative data source on monthly wages is the social security data collected by the German Institute for Employment Research (IAB), a source that is commonly known as the integrated employment biographies (IEB). The distribution of administrative contractual working hours stems from the German Statutory Accident Insurance, which is mandatory for all dependent workers in German firms and which we can link to the IEB data for the years 2010 to 2014.[10]

The PASS itself is a survey data set focused on the labor market, poverty and means-tested income support in Germany. Established by the Institute for Employment Research in 2007, this annual panel survey consists of a mixed sample of both a special sample with households that receive social benefits and a regular sample with households registered as residents of Germany. Initially, a personal interview is conducted with the heads of all selected households. Then, all members of the household aged 15 year and older are interviewed. The sampled low-income households, which include persons who might have a job in poorly paying industries, make this survey data set particularly suitable for this simulation study.

8 For a comprehensive description of the PASS survey, see Trappmann et al. (2013).
9 As the survey is missing information at the establishment level, this exercise had to be skipped.
10 In Germany, the Statutory Accident Insurance is part of the social security net. It is a mandatory insurance scheme that provides compensation for accidents and illnesses suffered by insured employees during their insured working time.

Figure 13.: Measurement error in German PASS survey data.
Scatter plot of survey and administrative wages (left panel) and observed distribution of measurement error (right panel).



Note: Measurement error in German PASS survey data, assuming that the gross hourly wages from the administrative data do not contain any measurement error ($w_{admin} = x$). The left panel is a scatter plot of log survey wages versus log administrative wages with a linear fit obtained via least squares regression. The right panel shows a histogram of the log difference between the two measures and the kernel density for comparison.
Source: own calculations and illustrations.

Observations from PASS data have been record-linked with the IEB, which includes exhaustive administrative information on past employment spells and the respective gross daily wages. The employment spells are mandatory reports from the employers for each employee provided at least once a year. Spells are day-specific in the sense that they include exact hiring dates and job termination dates, which allows us to link the individual's job at the date of the interview with the respective gross wage. The same logic applies to the administrative working hours, which are reported in the course of mandatory employer reports to the German Statutory Accident Insurance. These data on hours were mandatory in the years 2010 to 2014, and by means of the identical social insurance numbers, the job-specific information on hours can be merged with the employment spells of the IEB. Thus, we can merge the PASS survey information for the years 2011 to 2014, which is before the German minimum wage was introduced in January of 2015. Hence, this time period is ideal to infer the individual treatment status for a difference-in-differences evaluation study, as in the simulation exercises. In total, we observe 11,461 exact matches of individual observations in the PASS survey data and the administrative employment spells.

We define the measurement error in survey wages as the difference between wages in PASS and administrative data, i.e., $m_{PASS} = w_{PASS} - w_{admin}$, where we assume that the administrative wage is a good measure of the true wage ($w_{admin} = x$). Both the graphical illustration of the measurement error in Figure 13 and the additional descriptive assessments demonstrate that the measurement error in the PASS survey is similar in magnitude to some recent validation studies of other survey data (Gauly et al., 2018).[11] The variance in the measurement error is minimally dependent on the actual value, supporting our initial analyses that assume classical measurement error in survey wages (Sections 2-4). While the modus of the distribution of the PASS measurement errors is approximately zero, the distribution has a slightly negative skew.

7.1. Applying the empirical error distribution to individual data

As before, in this simulation, we set the effect of a minimum wage treatment on an outcome variable y (e.g., pay satisfaction) to be 1, and the minimum wage threshold is set to an hourly wage of EUR 8.50. The distributions of hourly wages and the error terms are derived from the PASS-ADMIN merge, and, as shown in Table 2, we gradually include more moments of uncertainty in our estimations.

Table 2.: Individual level results: Median bias of treatment effect estimates with PASS measurement error distributions after 1000 repetitions.

	PASS survey data, N=11,461 $\epsilon_{survey} = \epsilon_{PASS}$ $c_{survey} = c_{PASS}$	
	Median bias	p95-p05
Treatment dummy		
(1) $w_s = x_{admin}$	0.001	0.089
(2) $w_s = x_{PASS} = x_{admin} + c_s + \epsilon_s$	-0.366	0.095
(3) $w_s = x_{admin} + \bar{c}_s + \epsilon_s^*$	-0.371	0.094
Continuous treatment variable		
(4) $w_s = x_{admin}$	-0.001	0.240
(5) $w_s = x_{PASS} = x_{admin} + \bar{c}_s + \epsilon_s$	-0.339	0.267
(6) $w_s = x_{admin} + \bar{c}_s + \epsilon_s^*$	-0.369	0.206
Assumed dummy		
(7) $w_s = x_{admin}$	0.001	0.332
(8) $w_s = x_{PASS} = x_{admin} + c_s + \epsilon_s$	-0.199	0.351
(9) $w_s = x_{admin} + \bar{c}_s + \epsilon_s^*$	-0.227	0.333

Notes: This table shows the median bias of the estimates from the actual treatment effect (= 1) under several different assumptions: in (1), (4) and (7) – for comparison – we report the median treatment effect bias and its confidence interval for a model without measurement error (from administrative wages), in (2), (5) and (8) we use the observed error terms from survey data (deterministic approach), and in (3), (6) and (9) we randomly (*) assign errors terms in combination with a deterministic level shift.
Source: own calculations.

In the first line of Table 2, for comparison, we show the median treatment effect bias and percentiles for an individual-level model without measurement error. In line (2), we add deterministic error terms derived from the PASS survey and the merged administrative measures

[11] Separate scatter plots for log monthly wages and log working hours are presented in Figure 15 in the appendix.

of gross wages and working hours. Under these conditions, the downward bias of the treatment effect is expected to be roughly 37 percent and, thus, quite substantial. In line (3), we randomly assign survey errors, include the deterministic level shift and show bootstrapped results. Overall, the median treatment effect is downward biased by more than 37 percent. However, the bias of the median treatment effect (and its variance) do not change by a large magnitude when we include more moments of uncertainty in our simulation. Regarding the alternative specification with a continuous treatment instead of dummies in lines (4)-(6), as expected, the estimates are less precise and the bias is slightly smaller.

In lines (7)-(9) we present the results of a treatment dummy specification, when the relationship would actually necessitate a continuous treatment variable (assumed dummy). As predicted by the simulation, the overall downward bias is smaller than in both other specifications, but it is the least efficient specification. However, under empirical circumstances, falsely specifying a treatment dummy model does yield the least biased outcome, as this specification appears to be the most robust to measurement error in survey wages.

7.2. Applying the empirical error distribution to aggregated data

In this section, we again augment the Monte Carlo experiments using aggregated data. The individuals in the PASS survey can be aggregated to households and labor market regions. This step enables us to repeat the simulations under realistic aggregation conditions. However, it is not possible to aggregate individuals in the PASS at the establishment level since the sampling of the PASS is not clustered at the workplace level. Hence, for this particular level of aggregation, we have to rely on the results based on assumed distributions of measurement error, as in Section 4.

In Table 3, for comparison, we show the bias of the median treatment effect and the p05−p95 percentile range of the results for a model with aggregation but without measurement error in (1a) and (2a) for a specification with treatment dummies and in (3a) and (4a) for a specification with a continuous treatment variable. The results corroborate the conjecture that different units of aggregation would not bias the estimated treatment effects in the absence of measurement error (Angrist/Pischke, 2008). Aggregation does, however, affect the variance of the estimator. Smaller unit sizes for households result in a more efficient point estimator for the treatment effect compared to the larger unit sizes of regions, i.e., the $p95 - p05$ intervals are smaller for models based on households.

In terms of the median estimates for households or regions after aggregation and including measurement error for individuals, Table 3 shows the biases and p95-p05 intervals in lines

(1b) and (2b), assuming that the error terms are not correlated between members of the same household or region (see Table 2 for details of the Monte Carlo experiments at the individual level). At -0.283, the treatment effect is expected to be severely biased for aggregated households. The direction and magnitude of the bias are completely in line with the simulation in the last section (see Figure 7) and are the result of the small unit sizes of households and very high levels of segregation.

Table 3.: Aggregation results: Median bias of estimated treatment effects with PASS measurement error distributions and data aggregation after 1000 repetitions.

	Aggregated PASS survey data, N=11,461 $\epsilon_{survey} = E_{PASS}$ $c_{survey} = c_{PASS}$ Median bias	p95-p05
Treatment dummy		
(1a) Households w/o measurement error	0.001	0.104
(1b) Households w/ measurement error	-0.283	0.111
(2a) Regions w/o measurement error	0.009	0.381
(2b) Regions w/ measurement error	-0.204	0.357
Continuous treatment variable		
(3a) Households w/o measurement error	0.002	0.273
(3b) Households w/ measurement error	-0.268	0.313
(4a) Regions w/o measurement error	-0.003	1.095
(4b) Regions w/ measurement error	-0.064	1.108
Assumed dummy		
(5a) Households w/o measurement error	0.059	0.384
(5b) Households w/ measurement error	-0.083	0.394
(6a) Regions w/o measurement error	0.018	1.386
(6b) Regions w/ measurement error	-0.021	1.184

This table shows the median bias of the estimates from the actual treatment effect (= 1) after data aggregation at the household level (1a and 1b) and region level (2a and 2b). For comparison, we show the median treatment effect and the p05-p95 percentile range for a model with aggregation but without measurement error in (1a) and (2a). The error terms in the underlying individual data in (1b) and (2b) are derived from PASS survey data, as in line (2) of Table 2. Bootstrap samples are used to generate variation in the simulation of treatment effects. Source: own calculations.

The results for regional aggregation in line (2b) are possibly more informative. The treatment effect is estimated to be one-fifth less than its actual value in this setting, which is an improvement over the individual-level estimation (also compare Table 2). Note that from the simulation, as shown in Table 1, we observed that for higher levels of segregation, the estimated treatment effect is biased upwards. However, segregation of minimum wage workers across regions is low in reality; hence, we expect a downward bias and an inflated estimator variance.

When using a continuous treatment variable instead of a dummy, the bias of the estimated treatment effect after aggregating at the household level is similar, whereas the efficiency of the estimator is greatly reduced (3b). For aggregation at the region level (4b), the bias is comparatively small. Due to the loss of precision after aggregating to large unit sizes, this approach might not be feasible for some research agendas.

For the sake of completeness, lines (5) and (6) present the results for aggregating a specification with a dummy treatment variable, while the actual relationship follows a continuous variable. As with individual-level data, aggregating data does improve the estimates of the treatment effect, while precision of the estimates is lower at the household level and possibly unfeasible at the regional level. Interestingly, without measurement error this specification leads to a small upward bias at the household level (5a).

7.3. Implications for existing evaluations of the recent minimum wage introduction in Germany

A large number of evaluation studies recently emerged after the introduction of a nationwide minimum wage in Germany. In this literature, different levels of data aggregation are applied. Most prominently, we observe evaluations of employment effects that exploit variation at the establishment level (Bossler, 2017; Bossler/Gerner, 2019) and the region level (Ahlfeldt/Roth/Seidel, 2018; Caliendo et al., 2018; Garloff, 2019; Schmitz, 2019). However, we also observe various studies that use data on the individual level or aggregated data on the household level.

Individual-level data are analyzed in Bossler (2017) to estimate the effects on pay satisfaction and the work engagement of treated workers. While the effect on satisfaction is substantial, the effect on work engagement (which is interpreted as a proxy for individual motivation and productivity) is small and nonsignificant. Another study that applies individual variation is Caliendo et al. (2019), who analyze individual hours of work. The results indicate a significant reduction in working hours in the first year after the minimum wage introduction. Finally, Hafner (2019) analyzes the effects on the individual (self-assessed) health of affected individuals, and the results suggest a remarkably large positive effect.

While only the last study explicitly relies on data from the PASS survey, the first two apply the Linked Personnel Panel (LPP) and the Socio-economic Panel (SOEP), respectively. Assuming that all these data sources are collected from similar individual interviews and hence contain similar distributions of measurement error, we can conclude that these estimates are underestimated, with a bias between 25 and 45 percent. Hence, the negative effect on working hours and the positive effects on health and pay satisfaction might be larger than suggested in the respective studies.

We are aware of only one study that analyzes the effect of the minimum wage at the household level by analyzing the effect on poverty based on PASS data (Bruckmeier/Becker, 2018). While the authors detect hardly any effects of the minimum wage, this lack of significance could be due to measurement error, as our results suggest an underestimation by approximately 28 percent.

Again, assuming error distributions as observed in the PASS data, our results suggest a smaller downward bias (approximately 20 percent) for all the studies that apply regional variation (Ahlfeldt/Roth/Seidel, 2018; Caliendo et al., 2018; Garloff, 2019; Schmitz, 2019). When we assume a continuous treatment variable, as in Bonin et al. (2019), the bias is even smaller if the assumption holds. However, our simulations also suggest that the bias has a large variation when using aggregated data at the region level, which could explain the differences in meaningful negative effects observed in Caliendo et al. (2018) and Schmitz (2019) and effects that are virtually zero, as in Ahlfeldt/Roth/Seidel (2018) or Garloff (2019).

For the results at the establishment level, we do not observe an error distribution from the PASS data. Nevertheless, we can draw some cautious conclusions from the simulations that assume a classical error distribution applied to an observed establishment size distribution. Given the observed level of segregation in Germany, which is approximately 0.6 in the establishment data applied in Bossler (2017); Bossler/Gerner (2019), the median bias is slightly negative. However, with increasingly segregated labor markets (see Card/Heining/Kline, 2013), we can expect an increasing likelihood of overestimating the treatment effects of the minimum wage at the firm level in the future.

8. Summary and conclusion

In this simulation study, we assess the role of measurement error in a situation where the treatment variable is inferred from a survey wage distribution that is contaminated with measurement error. Econometric theory predicts attenuation bias towards zero in the presence of measurement error in the independent variable. This regression dilution transfers to our Monte Carlo experiment, in which the treatment variable for an evaluation of a minimum wage introduction is inferred from a wage threshold. This treatment assignment is fuzzy when individual wages contain measurement error. The respective treatment effects from a simple two-period difference-in-differences specification are then downward biased; moreover, the size of the negative bias increases with increasing variance of the measurement error. The inclusion of nonclassical error characterized by mean reversion in the survey responses has little impact compared to the classical error term, in particular, the variance of the error term.

The second issue of this study concerns the aggregation of potentially misreported data. Researchers typically aggregate individual-level data to higher-level units, such as households, firms, or regions, to alleviate the bias. Applying such data aggregation in another series of Monte Carlo experiments, we find that the magnitude and direction of the bias depend on the size of the aggregation unit and the allocation of treated individuals to such units. In cases of strongly segregated allocation, measurement error may even cause an upward bias in the estimated treatment effect. However, using empirical distributions of wages, working hours and error terms – derived from a record linkage of survey and administrative data – we find that the treatment effect is biased towards zero in the presence of classical measurement error under empirical conditions. However, the results also show that aggregation of the treatment information from PASS survey data to the household or region level does not fully alleviate the bias.

In addition to aggregation of data to higher-level units, we propose two alternative methods to reduce the bias. First, when building the regression model, scholars might incorporate a continuous treatment variable instead of a treatment dummy (Bonin et al., 2019). A continuous treatment variable implies that the effect of the treatment depends on its intensity, while a treatment dummy captures only a level shift. On the individual level, the results barely improve. For aggregated data with moderate contamination, we recommend using a continuous treatment variable, as the results are more robust to measurement error. Second, as a more drastic measure, we propose deleting observations close to the minimum wage threshold before its introduction, as those observations are affected the most by misclassification. However, this strategy is only recommended for a model with a dummy treatment variable (for both individual-level and aggregated data). If we falsely assume that the dummy variable captures the relationship between minimum wage and a homogeneously shifting dependent variable, while the relationship would necessitate a continuous model specification, dropping observations does yield a positive bias.

For scholars and decision makers, the consequences of this study are possibly substantial. Increases and changes to policies (such as the minimum wage) are often justified by previous ex post evaluations that concentrate on estimates of the effects of such an intervention. If policy evaluations potentially yield unreliable results, the decision-making process might lead to false conclusions. In the case of minimum wage introductions, or subsequent increases, the effects of such policies are distorted downwards, if not trivialized.

References

Ahlfeldt, Gabriel M; Roth, Duncan; Seidel, Tobias (2018): The regional effects of Germany's national minimum wage. In: Economics Letters, Vol. 172, p. 127–130.

Angrist, Joshua D; Pischke, Jörn-Steffen (2008): Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Bonin, Holger; Isphording, Ingo E.; Krause-Pilatus, Annabelle; Lichter, Andreas; Pestel, Nico; Rinne, Ulf (2019): The German Statutory Minimum Wage and Its Effects on Regional Employment and Unemployment. In: Journal of Economics and Statistics.

Bossler, Mario (2017): Employment expectations and uncertainties ahead of the new German minimum wage. In: Scottish Journal of Political Economy, Vol. 64, No. 4, p. 327–348.

Bossler, Mario; Broszeit, Sandra (2017): Do minimum wages increase job satisfaction? Micro-data evidence from the new German minimum wage. In: Labour, Vol. 31, No. 4, p. 480–493.

Bossler, Mario; Gerner, Hans-Dieter (2019): Employment effects of the new German minimum wage: Evidence from establishment-level micro data. In: Industrial and Labor Relations Review, forthcoming.

Bound, John; Brown, Charles; Mathiowetz, Nancy (2001): Measurement error in survey data. In: Handbook of Econometrics, Vol. 5, Elsevier, p. 3705–3843.

Bruckmeier, Kerstin; Becker, Sebastian (2018): Auswirkungen des Mindestlohns auf die Armutsgefaehrdung und die Lage von erwerbstaetigen ALG II-Bezieherinnen und -Beziehern. In: Studie im Auftrag der Mindestlohnkommission.

Caliendo, Marco; Fedorets, Alexandra; Preuss, Malte; Schröder, Carsten; Wittbrodt, Linda (2018): The short-run employment effects of the German minimum wage reform. In: Labour Economics, Vol. 53, p. 46–62.

Caliendo, Marco; Grabka, Markus; Obst, Cosima; Preuss, Malte; Schröder, Carsten (2019): The Impact of the Minimum Wage on Working Hours. In: Journal of Economics and Statistics, forthcoming.

Card, David (1992): Using regional variation in wages to measure the effects of the federal minimum wage. In: Industrial and Labor Relations Review, Vol. 46, No. 1, p. 22–37.

Card, David; Heining, Jörg; Kline, Patrick (2013): Workplace heterogeneity and the rise of West German wage inequality. In: The Quarterly Journal of Economics, Vol. 128, No. 3, p. 967–1015.

Garloff, A. (2019): Side effects of the new German minimum wage on (un-) employment: First evidence from regional data. In: German Economic Review, forthcoming.

Gauly, Britta; Daikeler, Jessica; Gummer, Tobias; Rammstedt, Beatrice (2018): What's your Wage? Comparing Survey and Administrative Data to Validate Earning Information. In: Perspectives on (Un-) Employment, Nuremberg, Jan 18th 2018.

German Minimum Wage Commission (2016): Erster Bericht zu den Auswirkungen des gesetzlichen Mindestlohns. In: Bericht der Mindestlohnkommission an die Bundesregierung nach §9 Abs. 4 Mindestlohngesetz, Berlin.

Hafner, Lucas (2019): Do minimum wages improve self-rated health? Evidence from a natural experiment. Tech. Rep., FAU Discussion Papers in Economics.

Kosfeld, Reinhold; Werner, Alexander (2012): Deutsche Arbeitsmarktregionen-Neuabgrenzung nach den Kreisgebietsreformen 2007–2011. In: Raumforschung und Raumordnung, Vol. 70, No. 1, p. 49–64.

Metcalf, David (1999): The British National Minimum Wage. In: Centre for Economic Performance, London School of Economics and Political Science.

Nestić, Danijel; Babić, Zdenko; Blažević Burić, Sanja (2018): Minimum wage in Croatia: sectoral and regional perspectives. In: Economic Research – Ekonomska Istraživanja.

Nolan, Brian; O'Neill, Donal; Williams, James (2002): The impact of the minimum wage on Irish firms. In: .

Pestel, Nico (2017): Marital Sorting, Inequality and the Role of Female Labour Supply: Evidence from East and West Germany. In: Economica, Vol. 84, No. 333, p. 104–127.

Prais, Sig J; Aitchison, John (1954): The grouping of observations in regression analysis. In: Revue de l'Institut International de Statistique, Vol. 22, No. 1/3, p. 1–22.

Schmitz, Sebastian (2019): The Effects of Germany's Statutory Minimum Wage on Employment and Welfare Dependency. In: German Economic Review, forthcoming.

Schmucker, Alexandra; Seth, Stefan; Ludsteck, Johannes; Eberle, Johanna; Ganzer, Andreas; et al. (2016): Establishment History Panel 1975-2014. In: FDZ-Datenreport, Vol. 3, p. 2016.

Trappmann, Mark; Beste, Jonas; Bethmann, Arne; Müller, Gerrit (2013): The PASS panel survey after six waves. In: Journal for Labour Market Research, Vol. 46, No. 4, p. 275–281.

Appendix

A. Observed unit size distributions

In this appendix, we describe the construction of realistic unit size distributions for observable data, as applied in Section 4.2. Table 4 summarizes the three scenarios of our simulations, including the data source from which we infer the unit sizes and the observed level of segregation.

Table 4.: Three scenarios of aggregation in applied research.

Aggregation scenarios	Aggregation level	Source of the size distribution	Level of segregation (Measured by the HHI)
(1)	household	PASS	HHI=0.94
(2)	establishment	BHP	HHI=0.59
(3)	region	BHP	HHI=0.40

Notes: The HHI for the household level is calculated from the PASS data, the HHI for the establishment level is calculated from the IAB-Establishment Panel, and the HHI for the region level is calculated from the PASS data.
Source: own calculations.

For the household level, we use the PASS, which is a household survey that collects data on each household member via a personal interview (see Trappmann et al., 2013). Hence, we observe the number of household members for each household in the data. From this information, we draw household sizes until the total number of individuals in these households sums to the 10,000 individuals in our original sample.[12] The average number of employees per household is 1.26 across all bootstrap samples, where the minimum is 1 and the maximum is 4. On average, the 10,000 individuals in each data set are assigned to 7,936 households.

For the establishment level, we use the BHP of 2014, which is an establishment-level data source that covers all establishments in Germany with at least one legal employee (see Schmucker et al., 2016). Since the information on the number of employees is included for all establishment-level observations, we randomly draw establishment sizes until the 10,000 individuals of our original data sample are assigned to one of the establishments. This sampling of realistic establishment size distributions yields an average establishment size of 11.73 employees per establishment, where the minimum is 1 and the maximum ranges between 281 and 7964 employees, depending on the bootstrap sample.

For the regional aggregation, we use a slightly different approach since most regions in Germany have more that 10,000 individuals, and it is quite unlikely that all individuals of a region would participate in a survey. Instead, we fix the number of regions to 141, which is the number of distinct labor market regions suggested by Kosfeld/Werner (2012). We assign the

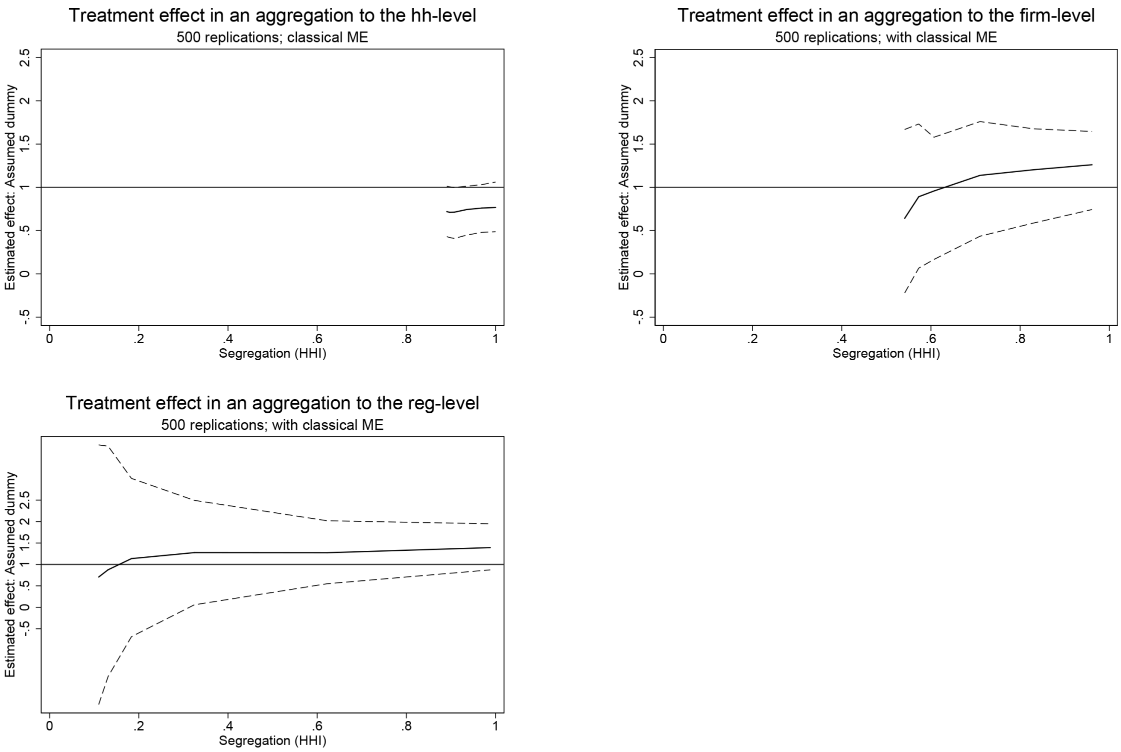
[12] In some cases, there is a residual household if the last household draw does not exactly sum up to 10,000 individuals but instead exceeds 10,000 individuals. For example, if 998 individuals are already assigned to U household and household U+1 with 3 household members is drawn, this residual household would instead be assigned 2 household members such that the total number of individuals is exactly 10,000.

10,000 individuals of our initial data sample to these regions according to the regions' true size distribution observed from the BHP, resulting in an average number of 71 individuals per region.

B. Aggregating to higher-level units: Assumed dummy

In Figure 14, we show the results of a simulation in which we use a dummy treatment variable, while setting the data generating process to a continuous relationship between dependent variable and treatment (which we call assumed dummy). As before, we aggregate at the household, firm and region levels with varying levels of segregation of minimum wage workers into higher-level units. After aggregation, the size of the bias depends on both the error term distribution and on the level of segregation. It is possible that high levels of segregation within regions and firms yield an upward bias in the treatment effect. For small unit sizes, such as households, such an outcome is unlikely.

Figure 14.: Using a dummy treatment variable when a continuous treatment variable is appropriate.

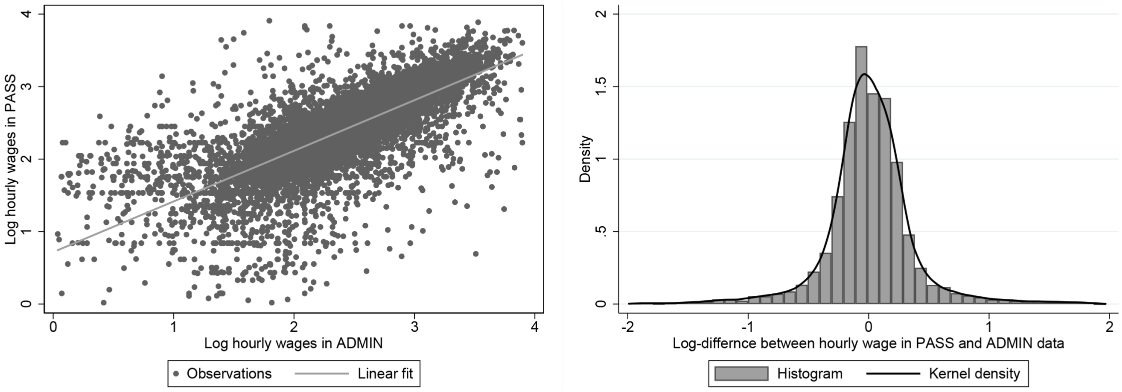


Note: Estimated treatment effect bias for the scenarios summarized in Table 4 using the dummy treatment variable while a continuous treatment variable would be appropriate.
Source: own calculations and illustrations.

C. Measurement error in monthly wages and working hours

While the Integrated Employment Biographies are commonly used among researchers both from Germany and internationally, the novel data set of reported working hours from the German Statutory Accident Insurance are widely unknown. In Germany, the Statutory Accident Insurance is part of the social security net. It is a mandatory insurance scheme that provides compensation for accidents and illnesses suffered by insured employees during their insured working time. We can link the data to the IEB data for the years 2010 to 2014 via identical social security numbers in both data sources. Similar to the Integrated Employment Biographies that is reported to German Federal Employment Agency, the spells of the administrative working hours are reported in the course of mandatory employer reports to the German Statutory Accident Insurance. These data on hours were mandatory only in the years 2010 to 2014, and by means of the identical social insurance numbers, the job-specific information on hours can be merged with the employment spells of the IEB. Before 2010 and after 2014, firms typically reported working hours directly to the German Statutory Accident Insurance.

Figure 15.: Measurement error in monthly wages and working hours in the Integrated Employment Biographies.
Scatter plot of monthly wages (left panel) and scatter plot of working hours (right panel).



Note: Scatter plot of log monthly wages (left panel) and log working hours (right panel), where the PASS survey data is on the vertical axis and ADMIN (social security data) is on the horizontal axis.
Source: own calculations and illustrations.

While this is the first time, researchers have an administratively collected estimate for individual working hours linked to the Integrated Employment Biographies and, thus, can compute hourly wages, the measurement of this variable will still be far from perfect. Most importantly, there are still inherent differences between contractual and actual hours worked in the data, that we are not able to quantify. Moreover, firms are allowed to submit an estimate of the hours worked as well, as many firms have no way of recording the hours of their workforce.

This means there is still measurement error in the hourly wages that we can not control for. There is, however, more measurement error if we only take employees working full-time and assume a 40-hour work week. Hence, we expect the empirical distribution of hourly wages to be a significant improvement over past contributions.

Imprint

IAB-Discussion Paper 11|2020EN

Publication Date
6. April 2020

Publisher
Institute for Employment Research
of the Federal Employment Agency
Regensburger Straße 104
90478 Nürnberg
Germany

Download
http://doku.iab.de/discussionpapers/2020/dp1920.pdf

All publications in the series “IAB-Discusssion Paper” can be downloaded from
https://www.iab.de/en/publikationen/discussionpaper.aspx

Website
www.iab.de/en

Corresponding author
Mario Bossler
mario.bossler@iab.de
Christian Westermeier
christian.westermeier@iab.de