Documenting and Improving the Hourly Wage Measure in the Danish IDA Database*

Christian Giødesen Lund,[≠] Rune Vejlin[#]

Abstract: This paper overhauls the hourly wage measure that is most often used in Danish research, the TIMELON variable in the IDA database. Based on a replication that we have constructed, we provide a documentation of the wage variable, the first of its kind, and then continue with a performance analysis. We find four puzzles. 1) The wages of part-timers fall steeply from 1992 to 1993, 2) the wages of full-timers fall from 2003 to 2004, 3) the level of the part-timer wages is around 12.5% higher than it should be, and 4) the wages of new hires fall steeply from the first year of employment to the second year. We analyze these puzzles in depth and solve almost all of them. Finally, we propose a new hourly wage measure that incorporates all the solutions and we show that it performs much better.

Keywords: Danish hourly wages; IDA data

JEL: J00; J31

1. Introduction

Access to accurate wage statistics is essential in a modern society. Policy makers, unions and employer associations, researchers, think tanks, economists in the private sector, and wage earners in general all rely on wage statistics. In Denmark, we

* Acknowledgements: We would like to thank Søren Leth and Jørn Schmidt from Statistics Denmark for help with old pieces of code and variable definitions. We would also like to thank Annette Mortensen for proof-reading. This paper could not have been written if it was not for the help of Henning Bunzel, who was instrumental in establishing the contact with Statistics Denmark and obtaining the data. Finally, we want to thank the editor Peter Møllgaard and two anonymous referees for comments and suggestions. All errors are our own.

≠ cgl@u.northwestern.edu

Corresponding author: Department of Economics and Business, Business and Social Sciences, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V., Denmark. Email: rvejlin@econ.au.dk

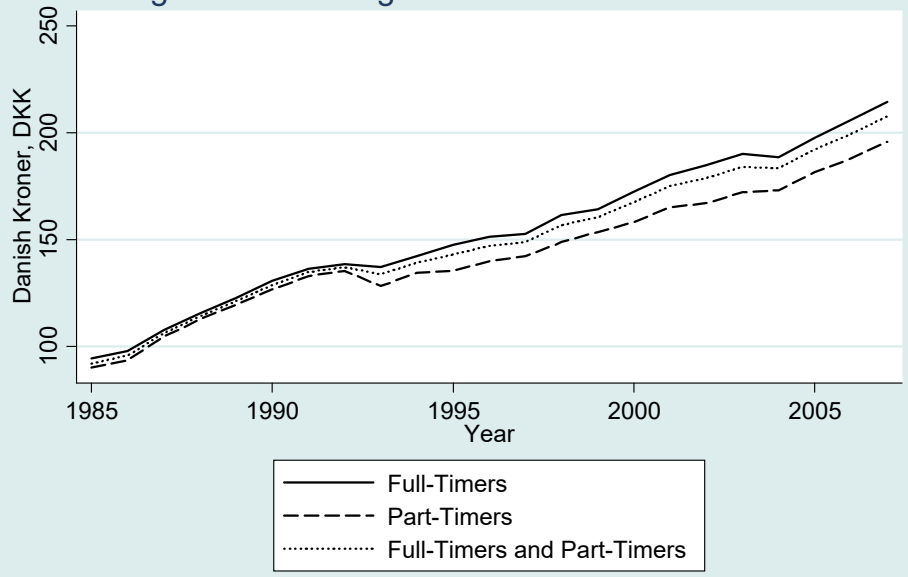
have several sources of wage information. One key source is the hourly wage measure from the IDA database ("Integrated Database for Labour Market Research") which is maintained by Statistics Denmark. This wage measure has been estimated yearly from 1980 to 2010 for practically all workers in the Danish economy and is pervasively used among researchers and economists of all kinds. Curiously, despite the widespread use, the IDA wage is only sparsely documented and has not undergone much rigorous testing. In fact, the only known public references are Statistics Denmark (1991) and the readily available but quite cursory online documentation (Online, 1). After studying these sources carefully, we still felt that the estimation method was like an impenetrable black box, and the sources were silent on its performance. In order to fully understand how the hourly wage is actually computed, what its advantages and disadvantages are, and if there is room for improvements, it is necessary to dig much deeper than what the two sources do. In this paper and in an accompanying working paper, Lund and Vejlin (2015), we take on that task and conduct a thorough evaluation of the IDA wage. For readers that are more interested in the technical details we refer to Lund and Vejlin (2015). The result is a detailed documentation in English, a performance analysis, and a new and improved wage measure to replace the IDA wage.

Our documentation is based on an accurate replication of the IDA wage. The sources for our replication are access to the registers of Statistics Denmark as well as former and retired employees of Statistics Denmark, who have generously provided us with a file of old parameter values and a PLUK and SAS programming code that in all likelihood were used by Statistics Denmark early on to compute the wages. We use the registers and code as a laboratory in which we experiment with variations of the parametric form of the code and with historical parameter values. When in doubt about an aspect of the replication, we run our own identification exercises to be completely sure that the identified parametric form and historical parameter values are correct. The result is a replication that only rarely deviates from the IDA wage, and when it does, the difference is only -1 or 1 DKK in the vast majority of cases, presumably as a result of different rounding.

We use our replication for the performance analysis. Our advantage is that all the auxiliary variables that are generated during the replication stage can be used for testing purposes. For instance, we can condition on part-time and full-time workers, which is crucial for isolating errors since their wages are estimated differently. The outcome of the analysis is four puzzles in which the hourly wage measure behaves too suspiciously to not reflect fundamental problems with the estimation method. The first puzzle concerns the time series of part-timer wages. In most years, the wages of part-timers track closely the wages of full-timers, but not in 1986-1993 where the part-timer wages approach the full-timer wages until 1992, only to take a big hit in 1993. This is a very unlikely event in the Danish labour market where collective bargaining is pervasive and the wages of different groups

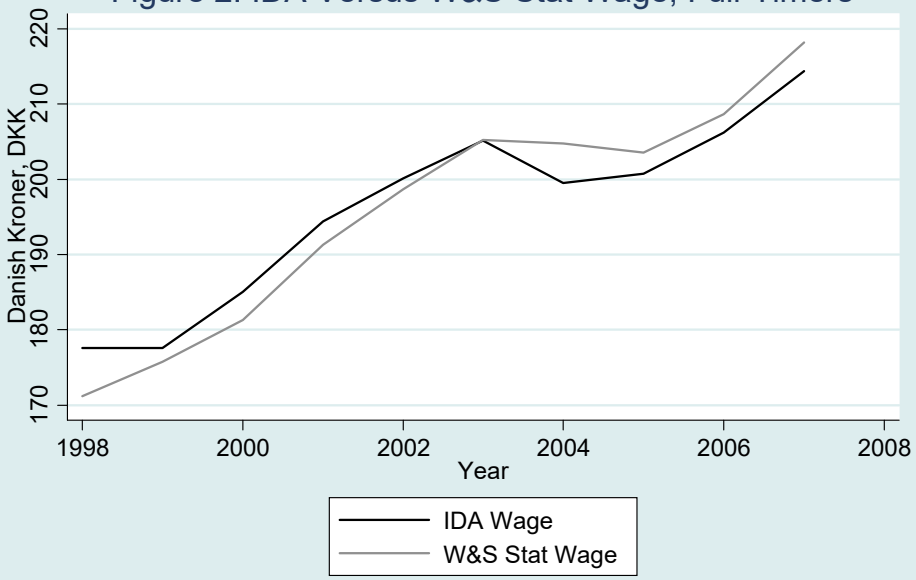
of workers exhibit roughly the same growth rates and rarely fall. The second puzzle concerns the time series of the full-timers. From 2003 to 2004, their wages fall on average whereas the wages of the part-timers rise, giving rise to a falling aggregate wage. This pattern is also unlikely to reflect what happened in the labour market in practice. Third, we assess the wage levels of the part-timers and full-timers in a comparison with a register called "Lønstatistik" (Wage and Salary Statistics/W&S Stat) produced from 1997 by Statistics Denmark. While the level of the full-timer wages is roughly right in most years, the level of the part-timer wages seems to be quite off the mark, with a difference of roughly 20 DKK or around 12.5%. Finally, tenure profiles are unrealistic. In the second year of employment at a firm, a worker's wage falls which goes against common sense as well as an overwhelming body of evidence of positive returns to tenure. The puzzles are illustrated in Figures 1 to 4. (The wage levels in Figures 2-3 are not the same as in Figure 1 – please see the online Appendix A for details.)

Figure 1: IDA Wages of Full-Timers & Part-Timers



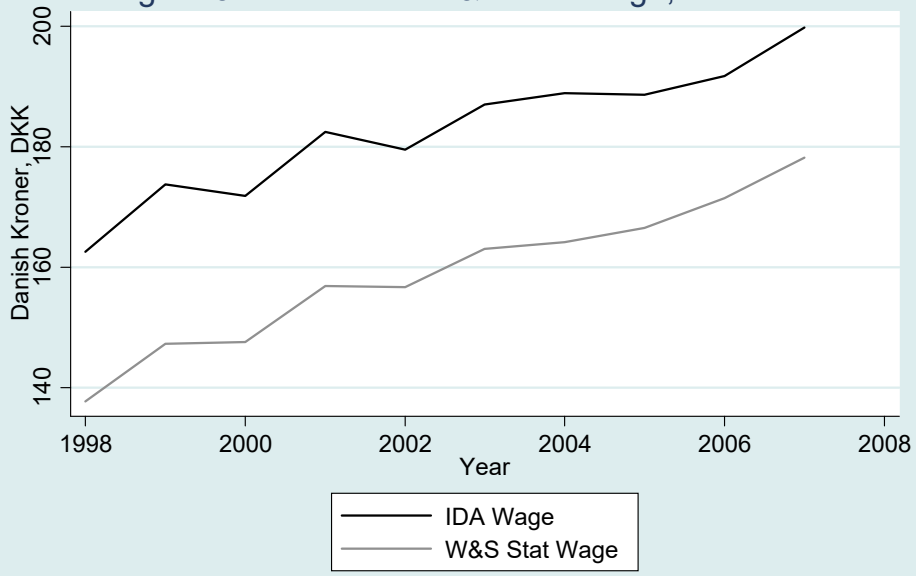
Source: Authors' calculations of wages in the Danish IDA database.

Figure 2: IDA Versus W&S Stat Wage, Full-Timers

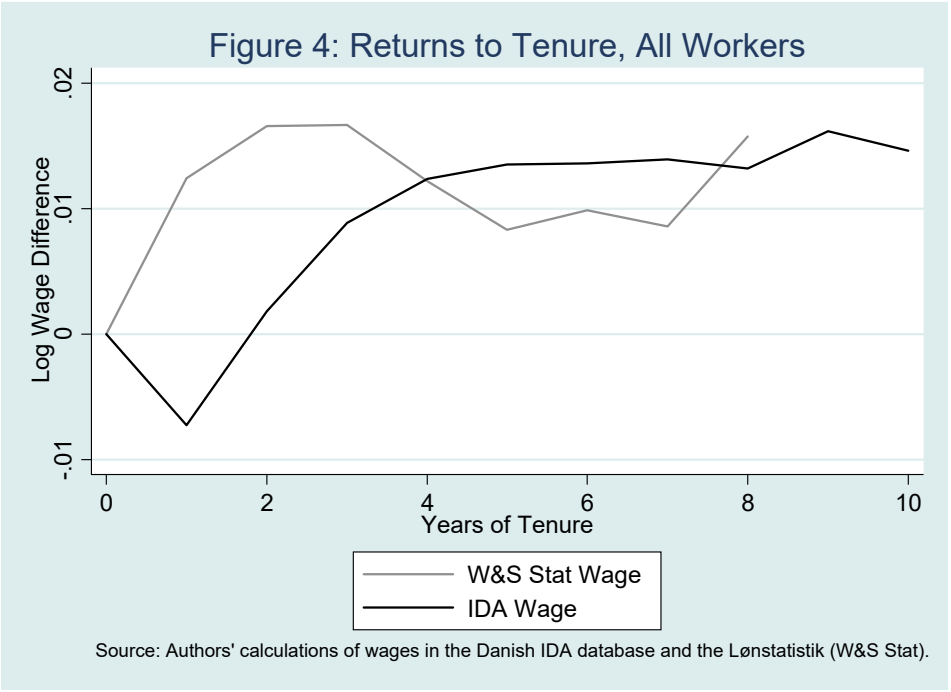


Source: Authors' calculations of wages in the Danish IDA database and the Lønstatistik (W&S Stat).

Figure 3: IDA Versus W&S Stat Wage, Part-Timers



Source: Authors' calculations of wages in the Danish IDA database and the Lønstatistik (W&S Stat).



We estimate the return to tenure by running the following regression $\log w_{it} = b_0 \cdot e_{it} + b_1 \cdot \tau_{it} + y_t + \alpha_i + \epsilon_{it}$. Here, w_{it} is the hourly wage of individual i, e_{it} is a vector of dummies capturing the overall labor market experience of worker i in year t, τ_{it} is a vector of dummies capturing the tenure of individual i in her job in year t, y_t is a year fixed effect, α_i is a worker fixed effect, and ϵ_{it} is an error term.

We are not the first to think about the performance of the IDA wage but previous efforts are too few and anecdotal to be informative. Statistics Denmark (1991) performs wage level comparisons based on a small sample from 1981. The results are ambiguous and unclear due to the few observations and insufficient wage benchmarks. DØRS (2003)[1] is an attempt to both solve the sample problems and bring in a comparable wage measure. That study also uses the Lønstatistik, but finds that the IDA part-timer wages should be closer to the part-timer wages than the IDA full-timer wages. This result is quite unlikely to be valid, since we argue in our documentation and show in our performance analysis that the full-timer wages are much better estimated than the part-timer wages.[2] Finally, Statistics Denmark (undated memo) notices a fall in wages in 1993 and ponders what drives it. Based on

1. This study does not seem to be publicly available but we are in possession of a copy that will be furnished upon request.
2. Unfortunately, DØRS (2003) is only a very brief draft, which precludes a definitive analysis of what drives the differences. Yet, in the accompanying working paper we identify a number of short-comings in their approach and argue that our wage benchmark is useful and close to optimal within the constraints of the Lønstatistik.

some descriptive statistics, a brief diagnosis is made that attributes the falling wages to a 1993-overhaul of the pension rules. However, no explicit distinction between part-timers and full-timers is made, and a full formal analysis of the estimation method and how it responds to changes in parameters is never undertaken. As a result, we do not learn what really causes the puzzling pattern, if there are ways to improve upon it, or what the time series of part-timer wages should really look like.

We investigate what the solutions to the puzzles are with formal analyses. Concerning the part-timer wages from 1986 to 1993, as it turns out the problem is not that they fall in 1993. Instead, the problem is that they grow too fast from 1986 to 1991 because the weekly working hours were reduced in this period, and the estimation method did not adapt to the changes. The overhaul of the pension rules in 1993 that Statistics Denmark (un-dated memo) pointed to as the culprit simply and accidentally restored the part-timer wages to what they should have been all along. Concerning the negative growth rate of the wages of full-timers from 2003 to 2004, we cannot claim to have solved the whole problem, but we do find that some of the fall is due to a premature rounding in the estimation code. Finally, we find that the tenure profile puzzle and the very high part-timer wage level are due to a common underlying problem, namely that true full-timers are classified as part-timers by the estimation method, with too low hours estimates as a result. The problem is particularly pervasive in the first year of employment at a firm, causing the flawed tenure profile that we see.

After solving the puzzles, we use the insights to construct a new and better hourly wage measure. In so doing, we retain the IDA wage concept and upgrade it with a number of improvements that fix the flaws. The improvements are effective yet conservative in the sense that very little or no extra assumptions are needed to justify them. They encompass correcting outright coding errors, controlling for measurement error, making the method more flexible and accommodative to changes in pension rules and working hours, fine-tuning some parameter values, and abolishing the distinction between part-timers and full-timers in favor of a continuum of degrees of full-time work. The result is a time series and a tenure profile that look much more realistic and a much improved level of the part-timer wages. The level can easily be further improved if one is ready to accept stronger assumptions. We hope that our new wage measure will be useful for researchers, statisticians, think tanks, perhaps the Danish Economic Councils, and, ultimately, anybody with an interest in wage statistics. In order to make the wage measure as useful as possible, we release it together with the auxiliary variables used in its estimation. The programming codes are also available upon request.

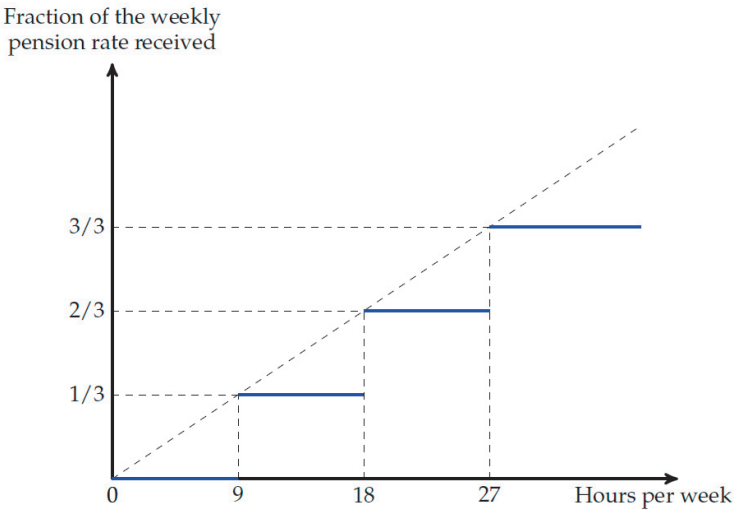
This paper is structured as follows. Section 2 explains how Statistics Denmark estimates the hourly wage in the IDA database, Section 3 solves the puzzle of the part-timer wages from 1986 to 1992, Section 4 solves the tenure profile puzzle and the wage level puzzle of the part-timers, and Section 5 combines all the changes

and shows the performance of the new wage measure. A sample description and other details of our empirical approach can be found in the online Appendix A. Space considerations have made it necessary to omit many of the details of our work. They can be found in Lund and Vejlin (2015), part of the 2015 Economic Working Papers series of the Department of Economics and Business, Aarhus University.

2. Documentation

The hourly wages in the IDA database are estimated by Statistics Denmark based on accurate yearly earnings records from the tax authorities and an estimate of the number of hours worked during the year. The earnings include compensation for overtime and earnings from paid periods of absence such as vacation and, in many cases, illness. Pensions are not included. Hours are estimated based on yearly pension contribution records ("ATP") for workers who are employed in either main or second jobs in the last week of November. As we will see, vacation, overtime, periods of illness, legal holidays, and lunch breaks are excluded from the hours estimates.
The estimation exploits the fact that the accumulated pensions depend on hours in a known way, and on hours only. The relationship is illustrated in Figure 5 in the case of workers who are paid every week after 1993.

Figure 5: Weekly Hours and the Received Fraction of the Weekly Pension Rate



As is seen, the earned pensions depend positively on the working hours although not in an invertible way. Thus, instead of point-identifying the working hours, Statistics Denmark backs out a range of possible hours that is consistent with the accumulated pensions in the year. For example, a worker who has accumulated two weeks' worth of pensions cannot have worked less than 54 hours in total, regardless of how many weeks the worker has been employed. In turn, the upper bound simply adds to the lower bound the number of weeks with employment multiplied with the maximum number of hours one can work in a week without earning extra pensions. If our worker was employed for three weeks and did not exceed 27 hours in any of the weeks, then the upper bound is simply 54+3*9=81 hours. The question is then what to do with this range of working hours, or what to do if the employment period is unknown or if an employee works overtime. In order to answer these and other questions, we need to be a bit more systematic.

2.1. Defining minimum and maximum values for hours

In this sub-section we define the minimum and maximum values for hours used in the later estimation of hours.

The Lower Bound on Hours

It is convenient to begin with some notation and definitions.

Pay cycles:	The contractual frequencies of pay that exist in the labour market, i.e. weeks, fortnights, or months. The most common pay cycle is a month.
Full pay cycle:	A week, fortnight, or month that is not shortened due to legal holidays. The alternative is a calendar pay cycle. Some but not all calendar pay cycles are full.

The variables to be defined next will take on different values depending on the type of pay cycle, yet for simplicity that dependence is suppressed in the notation. This means that, whenever two variables enter together in an expression, they refer to the same frequency of pay even if the notation does not explicitly indicate it. Also notice that one easily converts the values of one pay cycle to another – from weeks to fortnights, multiply by two; from weeks to months, multiply by 4 1/3.

h:	Effective number of full-time hours in a full pay cycle, as stipulated by agreements between unions and employers. Lunch breaks are not included since no work is done during lunch.
τ_1, τ_2, τ_3:	The time thresholds defining the *hours-brackets* of a pay cycle. At the end of each pay cycle, the employers determine what hours-bracket each employee is in and chips in the corresponding pension amount. There are four brackets: The top-bracket, middle-bracket, low-bracket, and zero-bracket. If a worker works less than τ_1 hours in a cycle, she is in the zero-bracket

and earns no pensions in that cycle. If she works τ_1 hours or more but less than τ_2 hours, she is in the low-bracket and earns $1/3$ of the pension rate pertaining to the pay cycle. If she works τ_2 hours or more but less than τ_3 hours, she is in the middle-bracket and earns $2/3$ of the pension rate. If a worker works τ_3 hours or more, she is in the top-bracket and earns the full pension rate in that cycle. The following relations hold: $0 < \tau_1$, $\tau_2 = 2\tau_1$, $\tau_3 = 3\tau_1$, and $\tau_3 < h$. In Figure 5, $\tau_1 = 9$, $\tau_2 = 18$, and $\tau_3 = 27$.

A: The yearly pension rate, that is, how much an employee accumulates if she is in the top-bracket in all pay cycles of an entire year. This is measured in DKK and is usually DKK 1166.

ATP: Total pension contributions from an employer to an employee in a particular job, accumulated over the year and reported to the tax authorities. This is also measured in DKK. Notice, that $ATP \leq A$.

T: The Hours Standard ("Normaltimetallet"), that is, the effective number of hours of work in a year with full-time work, agreed on by the social partners and enshrined in collective agreements. Vacation, weekends, legal holidays, or lunch breaks are not included in T because no work is done during these periods. Note that the collective agreements offer considerable flexibility for the employer to allocate time, especially in industries where this is necessary. In other words, T is just the reference point of full-time work, and not the exact hours that all full-timers work in a year.

k: An estimate of the number of hours a worker can work in a full pay cycle without earning extra pensions. In the zero-, lower-, and middle-brackets, a worker can maximally work τ_1 hours without passing the threshold to the next bracket and earning more pensions. Since the top-bracket has no upper bound, in that bracket an assumption on how many hours a worker can maximally work in a pay cycle is necessary to obtain a value for k.

For an overview of some of the variables see Table 2.
With these definitions, the lower bound on hours, t_{min}, is computed as

$$t_{min} = \frac{ATP}{A} T \cdot (1 - k/h).$$

The intuition is that the lower bound is the earned share of the yearly pension rate multiplied by the minimum number of hours one needs to work to obtain the yearly rate.

The Upper Bound on Hours

We now proceed to explain how Statistics Denmark computes the upper bound of hours, t_{max}. The upper bound adds to the lower bound an estimate of the number of pay cycles the worker could have been employed in the year multiplied by how many hours in each of those pay cycles the worker could maximally have worked without earning extra pensions. It is convenient to define the following variables.

u_{min}: The minimum number of full pay cycles consistent with ATP and A, given by

$$u_{min} = \frac{ATP}{A} \cdot \frac{T}{h},$$

the earned share of the yearly rate multiplied by the number of full pay cycles in a year. If one always works τ_3 hours or more and receives the full rate, one reaches ATP by working in only u_{min} full pay cycles.

V: The official number of weeks of vacation as guaranteed by law. There is no information in the IDA database about held vacation so the next best alternative is to use the official number for everyone.

u_{max}: An upper bound on the number of calendar weeks of employment in which there is actual work. The bound takes into account that individuals do not work while they are on vacation or full-time unemployed or receive social security allowances for illness. (There is some information in the IDA database about unemployment and illness.) Specifically, u_{max} is computed as the number of weeks in a year over and above the number of weeks with vacation, illness allowance, or full-time unemployment insurance.

p: The period of employment as reported in (calendar) weeks by employers. They indicate if their employees were employed the entire year, in a continuous period including begin and end dates, or in more than one period.

u: A mix between a lowest available upper bound and an estimate of the number of full pay cycles within a year. With a slight abuse of notation, it is obtained as $\max\{u_{min}, \min\{p, u_{max}\}\}$ where $\min\{p, u_{max}\}$ is now measured in full pay cycles. (To make the conversion, first convert $\min\{p, u_{max}\}$ into the same pay cycle as u_{min}, then multiply by $T/(h(52 - V))$ to make it a full pay cycle.) Legal holidays are not included in u. Note that in principle $u_{min} > \min\{p, u_{max}\}$ cannot happen, but when it does in the data, then u_{min} is considered the most reliable number, and the contradiction is resolved in favour of u_{min}. Similarly, if $p > u_{max}$ happens then u_{max} wins.

Note that the maximum possible number of hours for which no ATP is earned if the worker works in u pay cycles is ku. With these definitions, the estimated upper bound on hours in a job in a year is equal to

$$t_{max} = t_{min} + ku$$

There are two things to note. Firstly, it does not matter what the pay cycle is. Secondly, strictly speaking t_{max} is not always an upper bound. If $u_{min} < p < u_{max}$ (again abusing notation slightly) and p is lower than the true period length, t_{max} may be smaller than the true hours worked. Nevertheless, we will stick to referring to t_{max} as the upper bound on hours, and, similarly, to u as the upper bound on pay cycles.

Historical Values of the Parameters

Any documentation of the IDA hourly wage measure would be incomplete without the recent and historical values of the parameters that were used to estimate the wage. The parameter values are essential for understanding the wage measure and its properties better, and for replicating or correcting it if that need should arise. Unfortunately, no complete records of the historical parameters seem to exist, so we infer them from many different sources.

Table 2. Overview of Historical Values of Parameters used by Statistics Denmark, 1980-2007

Year	k_{wk}	V	h_{wk}	T	A	Year	k_{wk}	V	h_{wk}	T	A
1980	10	4.5	40	1836/1840	432	1994	10	5	37	1687	1166
1981	10	5	40	1816	432	1995	10	5	37	1676	1166
1982	10	5	40	1832	1166	1996	10	5	37	1680	1166
1983	10	5	40	1824	1166	1997	10	5	37	1672	1166
1984	10	5	40	1816	1166	1998	10	5	37	1680	1166
1985	10	5	40	1808	1166	1999	10	5	37	1695	1166
1986	10	5	40	1804	1166	2000	10	5	37	1672	1166
1987	10	5	39	1764	1166	2001	10	5	37	1665	1166
1988	10	5	38.5	1748	1166	2002	10	5	37	1676	1166
1989	10	5	38	1710	1166	2003	10	5	37	1676	1166
1990	10	5	37.5	1687	1166	2004	10	5	37	1702	1166
1991	10	5	37	1672	1166	2005	10	5	37	1687	1166
1992	10	5	37	1687	1166	2006	10	5	37	1680	1166
1993	10	5	37	1695	1166	2007	10	5	37	1672	1166

2.2 Estimation of the actual number of hours

Having established the bounds, we proceed to show how Statistics Denmark uses the bounds to compute the actual estimate of hours. In order to determine where in the interval the estimate of hours should lie, Statistics Denmark draws upon additional insights. For example, the hours of full-timers should lie close to h in an average pay cycle, whereas the hours of part-timers should be more equally distributed within and across the hours-brackets. Accordingly, Statistics Denmark distinguishes between full-timers and part-timers (part-timer does not mean half-timer; it means non-full-timer). They do so by means of the top-bracket. The idea here is that the group of workers who are in the top-bracket in all pay cycles consists mostly of full-timers, although there are exceptions – some part-timers are always in the top-bracket as well, and some full-timers are sometimes outside of the top-bracket. In order to obtain a homogenous group of full-timers, Statistics Denmark classifies as part-timers all workers who are always in the top-bracket yet satisfy at least one of four conditions indicative of part-time work. These conditions are part-time unemployment insurance, part-time unemployment, high minimum hours in a second job, or whether the job itself is a second job. In practice, to identify workers who are always in the top-bracket, Statistics Denmark uses the condition $u = u_{min}$. If it is satisfied, then the worker has earned her pension amount in the number of pay cycles it takes to earn that amount when always working in the top-bracket, so she must have been in the top-bracket in all pay cycles. Now, for the workers who are classified as full-timers, Statistics Denmark set hours equal to t_{max}. Table 1 illustrates the division between full-timers and part-timers and how frequently each criterion is applied.

It is more complicated to determine the hours of the part-timers since more types of workers comprise this group, and because indications are lacking of where in the interval of hours the true number of hours really lies. A distinction is made between workers in their main and second jobs, and here we consider only the main jobs[3]. The hours in a main job are quantified as the average of the bounds.[4]

Table 1 shows the distribution for men and women separately.

3. Chapter 1 of Lund and Vejlin (2015) describes the estimation of hours of second jobs in detail.
4. Unless the lower bound on the number of hours in a second job is so high that, in all likelihood, the hours in the main job are closer to their lower bound. See online appendix B for more details.

Table 1. Criteria for Full-Time and Part-Time Employment, and Their Frequencies in Percent

	Criterion	Men	Women
Full-time	Always top-bracket, main job, no part-time insurance, no part-time unemployment, no time reduction	51.9	42.9
Part-time	One or more pay cycles below top-bracket, main job	33.7	41.1
	Always top-bracket, part-time insured, main job	0.0	1.5
	Always top-bracket, full-time insured, part-time unemployed, main job	3.7	4.5
	Always top-bracket, full-time insured, no part-time unemployment, time reduction (i.e. $T_{red} \leq t_{min}$), main job	0.1	0.2
	Second job	10.2	9.4
Unknown	No ATP earned and period of employment unreported or zero	0.4	0.4
Total	–	100	100

Source: Authors' calculations using the sample described in the online Appendix A.
Note: Reduced hours are not only used in part-timer category 4 but also in the other part-timer categories if the appropriate condi-tions are satisfied.

Notice how the criteria for being a full-timer are very restrictive. Numerically, only half of the men are employed full-time, and less than half of the women. As a result, the hours of the full-timers should be precisely estimated, simply because care has been taken to exclude from the full-timer group any worker who in the slightest way resembles a part-timer. On the other hand, the part-timer group is much more mixed and there is no guarantee that the estimation method works well for such a heterogeneous group.

3. Solving Puzzle 1

In the Introduction, we demonstrated how the wages of part-timers grew fast in the period 1986-1991, and then dropped precipitously in 1993, see figure 1. Curiously, the abnormal growth rate pattern only applied to the part-timers, with the full-timer wages exhibiting perfectly normal growth rates throughout the period.[5] As we shall see in this section, a gradual reduction in the working hours between

5. See the online Appendix or Lund and Vejlin (2015) for the specific magnitudes.

1986 and 1990 induced a bias that drove the high growth rates in the wages of the part-timers, and an immediate 1993 reduction in the thresholds τ_1, τ_2, and $\tau_3$neutralised the bias again. The full-timers were not affected by these changes because of the presence of a second bias that for full-timers was exactly equally strong.

To begin, notice that any strange patterns over time relates to the bounds on hours, and not to how the bounds are used in the estimation. The reason is that the bounds are used in the estimation in the exact same way in all years. Accordingly, we check the expressions of the bounds by deriving true expressions taking into account the changes described below and performing a comparison. In order to derive the true expression for the lower bound on hours, it is necessary to consider a number of cases. Firstly, in 1993 the definition of the yearly rate A changed[6], making it necessary to distinguish between workers receiving the yearly rate and everybody else. Secondly, it is also necessary to distinguish between monthly compensated workers on the one hand, and weekly and bi-weekly compensated workers on the other. The reason is that the former group receives pensions even for the hours during which they are on vacation, but the latter group does not[7]. Here, we focus only on workers who are paid monthly and who are not in the top-bracket in all pay cycles of the year.[8] Define Y as the hours of a full-timer in a full year when not subtracting vacation and legal holidays, let atp denote the pension rate received by the worker each pay cycle, and let an "m"-subscript denote that the pay cycle is a month. With this notation, Chapter 2 of Lund and Vejlin (2015) derives the true lower bound to be

$$t^*_{min} = \frac{ATP}{\frac{atp_m}{\frac{T}{Y}}}\tau_{3,m}.$$

The denominator is the effective monthly pension rate, taking into account that a monthly paid worker earns pensions when on vacation or legal holiday. The right-hand side is the effective number of months it would take with the full rate to accumulate the earned pensions, multiplied with the minimum number of hours of work it takes to earn the full rate in each month. In order to see how the true expression compares with the one actually used, assume that all months are exactly equally long and equal to 4 1/3 weeks and to 1/12 of a year and rewrite t_{min} as

6. BKG 1992-09-29 nr 822 paragraph 1.
7. For 1993-present, BKG 1992-09-29 nr 822 or later promulgations.
8. The other cases are similar and presented in Chapter 2 of Lund and Vejlin (2015).

$$t_{min} = \frac{ATP}{\frac{A}{\frac{T}{h}}}(h-k) = \frac{ATP}{\frac{A}{\frac{T}{h_m}}}(h_m - k_m) = \frac{ATP}{\frac{A}{\frac{T}{Y}\frac{Y}{h_m}}} \cdot (h_m - k_m)$$
$$= \frac{ATP}{\frac{A}{12\frac{T}{Y}}} \cdot (h_m - k_m).$$

Here ATP is correctly measured and $A/12 = atp_m$ is the monthly pension rate. Monthly paid workers in the top-bracket do earn this amount in a month. It is seen that the only source of error is if the factors $(h_m - k_m)$ and $\tau_{3,m}$ are different. That was indeed the case in the years 1987-1992 where the weekly numbers were $\tau_{3,w} = 30$, $k_{wk} = 10$, and h_{wk} assumed values in the range 37-39. The lower bound on hours incorrectly decreased with the gradual fall in the working hours. Then from 1993, $\tau_{3,w} = 27$, $k_{wk} = 10$, and $h_{wk} = 37$, so $(h_m - k_m) = \tau_{3,m}$ again. The bias in the lower bound is easily corrected as $t^*_{min} = t_{min} \cdot \tau_3/(h-k)$.

Turning to the upper bound, first notice that any biases in the lower bound on hours will carry over to the upper bound as well, since

$$t_{max} = t_{min} + ku.$$

Second, the k is wrong to some degree in all years after 1987, the degree depending on how often a worker is in the top-bracket. The problem is that when the working hours and the thresholds of the hours-brackets change, the widths of the top-bracket and of the other brackets are affected differently. With different widths, the hours that an employee can work without earning extra pensions depend on the bracket. Yet only one number, k, serves as the estimate of those hours. From 1987 to 1992, the top-bracket narrowed while the other brackets remained of width k. As a result, part-timers who were often outside of the top-bracket had a ku that was mostly correct, but full-timers and other workers often in the top-bracket had a ku that was significantly too high. From 1993, lower thresholds of the hours-brackets meant that the top-bracket attained width k again, while the other brackets narrowed. As a result, ku was correct for the full-timers, but too high for the part-timers.

The combined effect of the biases in t_{min} and ku is zero for full-timers in all years. This is also obvious since we know that full-timers wages are only a function of hu. The reason is that the decline in the working hours and the fall in the thresholds affect both terms equally much in each pay cycle, but in opposite directions. Since the two terms count equally in the final hours estimate (the hours of full-timers are estimated as t_{max}), and since t_{min} and ku are based on an equal number of pay cycles (recall that $u_{min} = u$ for full-timers and that $t_{min} = u_{min}(h-k)$), the

mistakes are exactly equally big and of the opposite sign. In contrast, the combined effect of the two biases is non-zero for part-timers since the ku-term only counts with a weight of a half (the hours of part-timers are estimated as the average of the bounds), and since a part-timer is sometimes outside the top-bracket. The net effect is too low hours in the period 1986-1992, with the biggest shortfalls in 1991-1992.
The bias in ku is harder to eliminate since there is no data on how frequently each worker is in each hours-bracket. However, it is still possible to remove most of the bias in k without additional assumptions by computing bounds on the frequency in the top-bracket. To see this, notice that the upper bound on the number of pay cycles in the top-bracket is u_{min}. More pay cycles than u_{min} in the top-bracket would lead to a higher ATP. Hence, given that the worker works in u pay cycles, $u - u_{min}$ is the minimum number of pay cycles outside of the top-bracket. Moreover, the accumulated pensions also determine a lower bound on the number of pay cycles in the top-bracket. If the pension total is higher than what can be earned with work in the middle-bracket in all pay cycles, then some of the pay cycles must have been spent in the top-bracket. To determine how many, consider the following definition.

u_{top}: The smallest possible number of pay cycles in which hours worked surpass τ_3 if the worker works in u pay cycles.

A worker reaches the pension total for instance by working u_{min} pay cycles with the full rate. Accordingly, u_{top} satisfies the following equation.

$$u_{min} = u_{top} + \frac{2}{3}\left(u - u_{top}\right) \text{ if } \frac{2}{3}u \leq u_{min}, u_{top} \geq 0.$$

The condition $2/3u \leq u_{min}$ determines if the pensions earned in the middle-bracket are lower than the accumulated total. The solution to the equation is

$$u_{top} = (3u_{min} - 2u) \cdot 1\left[u_{min} \geq \frac{2}{3}u\right].$$

Intuitively, if u is only a little larger than u_{min}, then the pensions coming from the $u - u_{min}$ pay cycles are minor and it is necessary to work in the top-bracket in many of the remaining pay cycles in order to reach the total pension amount. If u is a lot larger, then it is possible to accumulate ATP without working in the top-bracket at all.

Equipped with these bounds and nothing else, the best correction that can be made to k without additional assumptions is

$$k^* = \frac{u_{top}}{u} \cdot (h - \tau_3) + \left(1 - \frac{u_{min}}{u}\right) \cdot (\tau_3 - \tau_2) + \frac{u_{min} - u_{top}}{u} \cdot k.$$

The first term replaces k with the true number $h - \tau_3$ for the guaranteed pay cycles in the top-bracket. The second term replaces k with the true number $\tau_3 - \tau_2$ for the guaranteed pay cycles in other brackets than the top-bracket. And the third term does not make any correction for the maximum number of pay cycles in ambiguous brackets.

4. Solving Puzzles 3 and 4

Figure 4 in the Introduction demonstrated how the wages in the Danish IDA database fell significantly from the first year of employment to the second year, in contrast to the positive associations found anywhere else in the literature. This is a worrying puzzle given that the IDA wage has been used in many studies on the returns to tenure and related questions (Bagger et al, 2014; Buhai et al, 2014). The Introduction also demonstrated that the wages of part-timers seem to be much too high (Figure 3) while the wages of full-timers attain the right level (Figure 2). This is worrying as well since the composition of part-timers and full-timers could contaminate many kinds of wage comparison exercises.

Chapter 1 of Lund and Vejlin (2015) offers some clues as to what could be wrong with the wage measure. They estimate separate tenure profiles of both full-timers and part-timers and find that they look perfectly normal.[9] No fall occurs from the first year of employment to the second year. As a result, the only likely explanation for the puzzling tenure profile in Figure 4 is that the ratio of part-timers to full-timers changes systematically from the year of hiring to the second year of employment. If this is the case (and it is due to miss-classification) then workers in their first year of employment would get a higher wage, since their hours would be smaller due to them being classified as part-timers.

However, there are only two possible explanations for such a compositional change of the workforce. One is that part-timers and full-timers are unequally likely to stay in their jobs from the first year to the second, and another is that there is a transition between the two groups.[10] Many workers classified as full-timers in their second year of employment are in fact classified as part-timers in their first year, whereas only very few part-timers in their second year were full-timers in their first year. Moreover, the authors argue that the transition is mostly an unfortunate artefact of a classification method that does not reflect what happens in the

9. Each individual-year observation is either included in the full-timer or part-timer sample according to which of the two groups the observation belongs too. Thus, individual years in a job spell can be in different samples.
10. Chapter 3 of Lund and Vejlin (2015) shows that the latter explanation is the important one.

labour market in practice. In addition, the misclassification of workers into the wrong category is likely to have a huge impact on wages. For instance, the hours of full-timers classified as part-timers will be discontinuously too low since they are computed by averaging the bounds on hours instead of using the upper bound.

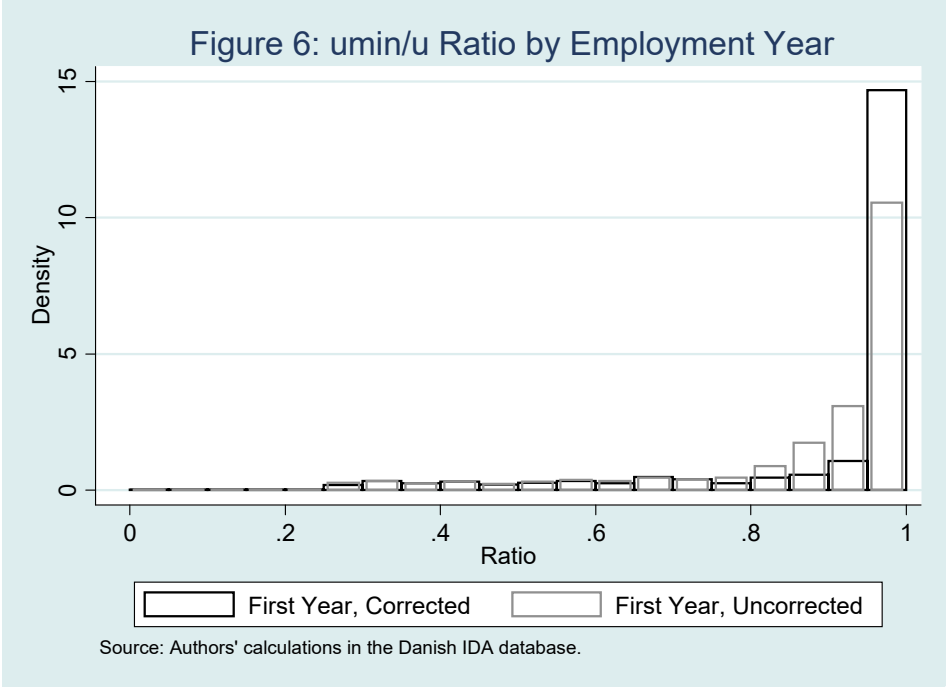
The documentation in Section 2 provides additional clues as to the direction of the misclassification. There is a fundamental asymmetry between full-timers and part-timers that make it much more likely that full-timers are classified as part-timers than the other way around. There are two reasons for this. Firstly, the condition $u_{min} = u$ is a knife-edge condition in the sense that it only takes an epsilon of a change in u to push a full-timer into the part-timer category, whereas $u_{min} < u$ is unaffected by such a change. For instance, an employee who works full-time in all pay cycles of the year except one will be treated on the same footing as part-timers who are never in the top-bracket. Worse, if the workload of the employee varies so that she compensates for few hours in one pay cycle by working overtime in other pay cycles, then she will be classified as a part-timer despite working full-time or more overall. Equally bad, simple small measurement errors in u can also push full-timers out of the top-bracket, but not the other way around. Secondly, the four conditions removing top-bracket workers from the group of full-timers are not paralleled by conditions transferring select part-timers to the group of full-timers. On a positive note, the full-timers become a quite homogenous group that is also likely to be roughly representative. However, the group of part-timers becomes a melting pot of part-timers, wrongly allocated full-timers, and workers in between part-time and full-time. In terms of the wages of the two groups, the wage level of the full-timers should be quite accurate, but the wages of the part-timers should be too high due to the workers who are pushed out of the top-bracket. As a consequence, puzzle 3 (Figures 2 and 3) should not be much of a surprise.

It remains to be explained how the misclassification of full-timers as part-timers can be so much stronger in the first year of employment than in subsequent years. There are two reasons. Most importantly, Statistics Denmark makes a mistake when computing u, the duration of employment within each year, given by

$$u = \max\left\{u_{min}, \min\{p, u_{max}\}\frac{T}{h(52-V)}\right\}.$$

As in Section 2, u_{min} is the minimum number of full pay cycles with work that is consistent with the earned pensions. u_{max} is the upper bound on the period with work, computed as one year minus vacation, detected periods of illness, and full-time unemployment and measured in calendar pay cycles. p is the employment period as reported by employers and measured in calendar pay cycles. The factor $T/(h(52-V))$ simply converts the calendar pay cycles into full pay cycles so that they are comparable with u_{min}. The mistake can be seen by the fact that p includes

vacation but u_{min} and u_{max} do not. The effect is that, for employment spells with $u_{min} < p < u_{max}$, u is off by a factor of roughly $52/47$ =1.1064, implying that u_{min}/u is off by a factor of 0.904. We can think of one case in which the condition $p < u_{max}$ is satisfied in a systematic fashion, and that is new hires in their first year of employment who by definition do not work the entire year. In these cases, even full-timers will have $p > u_{min}$, classifying them as part-timers. The effect of removing this coding error on the distribution of u_{min}/u is huge and can be seen in Figure 6. This shows the density of the u_{min}/u-ratio with and without the coding error corrected. Notice the considerable increase in the mass of full-timers (right-most bar) once we correct the error.



Secondly, there is a big potential for erroneous reporting of the period of employment of new hires, but not for more senior workers. The commencement date of new hires can be any day of the year, and there is ample room for the reported date to be far from the correct one. In contrast, by definition the workers in their second year and beyond work from the beginning of the year, so for them we have a very precise indication of the employment period. Moreover, the estimation method detects and corrects under-reporting of the employment period of the full-timers ($u_{min} > u$ cannot happen). So for full-timers, there is a lot of over-reporting in the first year, but no under-reporting. For part-timers, however, there is both under- and over-reporting in the first year. As mentioned, full-timers and part-timers are affected asymmetrically by erroneous reporting, and the result is too many part-timers in the first year.

The significant impacts of the misclassification of full-time workers call for a major revision of the notions of part-timers and full-timers. The goal here is to find an estimation method that not only works well for the subgroups of workers who fit perfectly into the two categories, but is also appealing for workers who fit best in between or who fall in the wrong category. The fundamental problem with the estimation method is the classification of workers into a small number of groups. Any such classification will suffer from trade-offs between how homogenous and representative some of the groups are and how mixed and biased the others are as a result. There will always be a significant number of workers with seriously biased wages.

It turns out that there is a straightforward way of working with a continuum of worker types. The key step is to estimate the *degree* to which a worker is a part-timer or full-timer, and there is a very good criterion for that, namely how *often* a worker works full-time. Letting $f \in [0,1]$ denote this frequency, we estimate hours as a convex combination of the current estimation methods for part-timers and full-timers, using f as the weight

$$(1-f)\frac{1}{2}(t_{min}+t_{max})+ft_{max}=t_{min}+\frac{1}{2}(f+1)ku.$$

This approach should work well given that $1/2(t_{min}+t_{max})$ and t_{max} are good estimates for true part-timers and full-timers, which we argued earlier to be the case. The difficulty, of course, is to estimate$f$. A first pass at a proxy is the frequency with which a worker is in the top-bracket. Using this proxy boils down to assuming that workers work full-time hours when and only when they are in the top-bracket. This assumption works well for workers who are either always or never in the top-bracket, and in fact Statistics Denmark uses it implicitly for such workers (Section 2). By extension, it also seems a good assumption for workers who are in the top-bracket often or rarely[11]. However, one might be concerned that some "true" part-timers randomly end up in the top-bracket in some pay cycles, and are incorrectly assigned the same hours in those pay cycles as if they had been full-timers. On those occasions, it might make more sense to assume hours to be $1/2(\tau_3+h)$ rather than h. There are two answers to this concern. As it happens, good criteria to distinguish true part-timers in the top-bracket exist and are already used by Statistics Denmark. Part-time insurance, part-time unemployment at some point in the year, and reduced hours are deemed sufficiently strong indicators of part-time work that workers always in the top-bracket satisfying these conditions are considered part-timers. We use the same criteria, for workers always in the top-bracket as well as for everybody else, to distinguish "true" part-timers. We put full weight on

11. The only reason for why this is not already in use seems to be that the technique to estimate the frequency in the top-bracket is only developed in this paper.

$1/2(t_{min} + t_{max})$ for the identified workers, that is, we copy Statistics Denmark and do not correct at all when a worker is deemed a true part-timer. For everybody else we use the frequency of pay cycles in the top-bracket as the weight on t_{max}.

We cannot observe the frequency in the top-bracket directly, but we can bound it using u_{min}/u and u_{top}/u as shown in Section 3. Another option is to use any convex combination of the bounds. We do both. For workers in their second year of employment or above, we estimate f with the estimator $f()$, a function of u_{min}/u and defined as

$$f\left(\frac{u_{min}}{u}\right) = \frac{u_{top}}{u}\left(\frac{u_{min}}{u}\right) \cdot 1[G].$$

Here, the dependency of u_{top}/u on u_{min}/u has been written out explicitly. $1[G]$ is an indicator function taking the value one if a worker is not part-time insured, is not part-time unemployed in the year, and does not work reduced hours due to second jobs. For workers in their first year of employment, we use a convex combination instead,

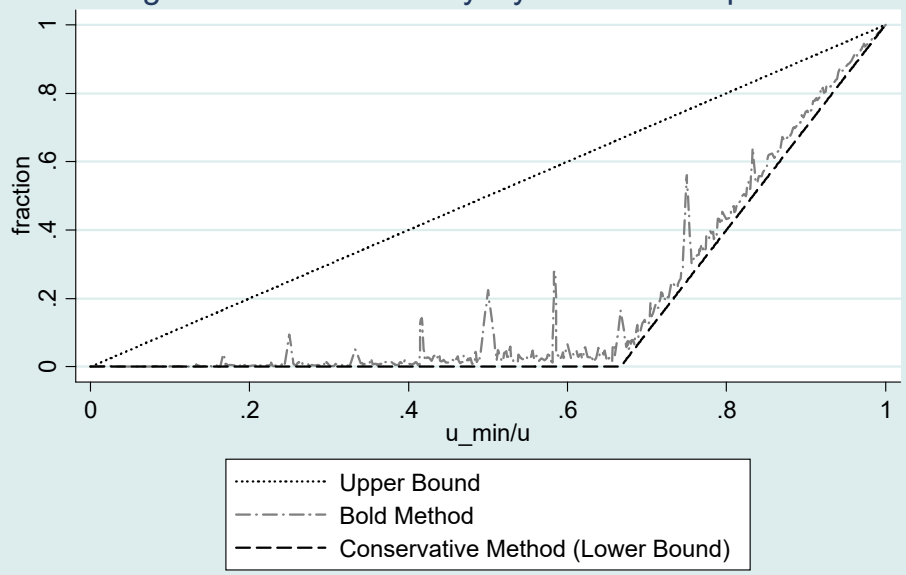
$$f\left(\frac{u_{min}}{u}\right) = \left(\left(1 - p\left(\frac{u_{min}}{u}\right)\right)\frac{u_{top}}{u}\left(\frac{u_{min}}{u}\right) + p\left(\frac{u_{min}}{u}\right)\frac{u_{min}}{u}\right) \cdot 1[G].$$

Here, $p()$ is a weighting function that accounts for over-reporting of u among new full-time hires. The idea is that the lower and upper bounds u_{top}/u and u_{min}/u can be interpreted as maximal and minimal smoothing of working hours across pay cycles. The lower bound u_{top}/u was obtained (Section 3) by counting the maximum amount of pensions earned outside of the top-bracket and comparing this maximum with the total accumulated pensions. When $u_{top}/u > 0$, that maximum was attained if a worker worked in the middle-bracket in all other pay cycles than those in the top-bracket. It follows that, if $u_{top}/u > 0$, the fraction of pay cycles in the top-bracket equals u_{top}/u if and only if there is maximal smoothing of working hours across pay cycles. On the other hand, if the fraction of pay cycles in the top-bracket equals u_{min}/u, then all work has been lumped together in a few intense pay cycles, and minimal smoothing obtains. Now, over-reporting means that no work is done between the incorrect and correct commencement dates. In other words, the smoothing of the working hours of full-timers is minimal when there is over-reporting of their employment durations, and the upper bound u_{min}/u should be used in these cases. $p()$ estimates for any given value of u_{min}/u a lower bound on the fraction of over-reported full-timers in their first year. The idea behind the estimation is to find a set of workers who are thought to work full-time in their first year of employment and observe their reported u_{min}/u-distribution. The conditional probabilities of reporting u_{min}/u given full-time work are obtained in this

way. Then Bayes' Rule can be used to infer the probability of full-time work given a reported value of u_{min}/u. The trick to identify the set of true full-time workers is to observe what happens in the second year of employment in which the spell lengths are in all likelihood correctly reported, and then assess the fraction of this group who also work full-time in the first year.

The bounds u_{min}/u and u_{top}/u on the fraction of pay cycles in the top-bracket as well as their convex combination are illustrated in Figure 7. The probabilities of over-reporting, or at least their lower bounds, are in fact quite low, since the convex combination is only a little above u_{top}/u. The exception is nine roughly equidistant spikes.[12]

Figure 7: Fraction of Pay Cycles in the Top-Bracket



Source: Authors' calculations of wages in the Danish IDA database.

Our way of estimating the frequency of pay cycles in the top-bracket is quite conservative. We prefer to not rely on a shaky foundation of strong assumptions. Instead, we change the hours estimates of Statistics Denmark only when we can explicitly justify it. For instance, we use the lower bound on the fraction of pay cycles in the top-bracket for all workers in their second year of employment and above, and for all new hires who cannot be explicitly shown to be over-reported full-timers. While our repeated use of lower bounds means that sharper and more accurate

12. Chapter 3 of Lund and Vejlin (2015) argues that the spikes are due to a specific, common pattern of over-reporting, namely that employers for simplicity report that all employees, including new hires, have worked the entire year.

estimates are likely to exist, we consider the lack of extra assumptions a significant plus. Moreover, in terms of the big picture, the workers who are most often in the top-bracket and have the most biased IDA wages will receive the biggest wage corrections, and measurement errors in u or variation in working hours from pay cycle to pay cycle will have close to zero effect on the hours estimates.

5. The New Wage Measure

Sections 3 and 4 dealt with three of the puzzles from the Introduction. In this section, we improve on the remaining time series puzzle concerning the period 2003-2004 as well as put forward some additional important enhancements of the programming code. We then combine all the changes into a new wage measure and analyse its performance.

To begin with the puzzle, notice that Statistics Denmark rounds several expressions prematurely in their programming code[13]. For instance, rounding is performed to parameter values in the very beginning and to numerators and denominators before their division. One can see why rounding might have a small but visible impact on the growth rates of wages. The reason is that some of the parameter values such as T change slightly from year to year, and the changes apply to all workers. The rounding accentuates the changes in some cases, the most conspicuous of which are 2003 and 2004. Once we remove the rounding, the fall in the aggregate wage from 2003 to 2004 documented in the Introduction disappears (see Figure 8). However, the average wage of full-timers still does not grow, which indicates that we have solved only part of the problem.

In addition to removing the rounding, we remove vacation from p as mentioned in Section 4 and overhaul the criterion for reduced hours completely.[14] We then combine these changes with the ones made in Sections 3 and 4, to arrive at the following estimator of hours for workers in their second year or above ($t^*_{min} = t_{min} \cdot \tau_3/(h-k)$):

$$t^* = t^*_{min} + \frac{1}{2}\Big(\frac{u_{top}}{u}(1+1[G])\frac{h-\tau_3}{k} + \Big(1-\frac{u_{min}}{u}\Big)\frac{\tau_3-\tau_2}{k} + \frac{u_{min}-u_{top}}{u} \cdot 1\Big)ku.$$

13. Back in the day, the rounding was probably carried out for data management and computational reasons.
14. As demonstrated in Chapter 4 of Lund and Vejlin (2015), the criterion used by Statistics Denmark selects the wrong workers to give reduced hours, with clear impacts on the wage level (around 5 DKK for part-timers and around 2 DKK for full-timers).

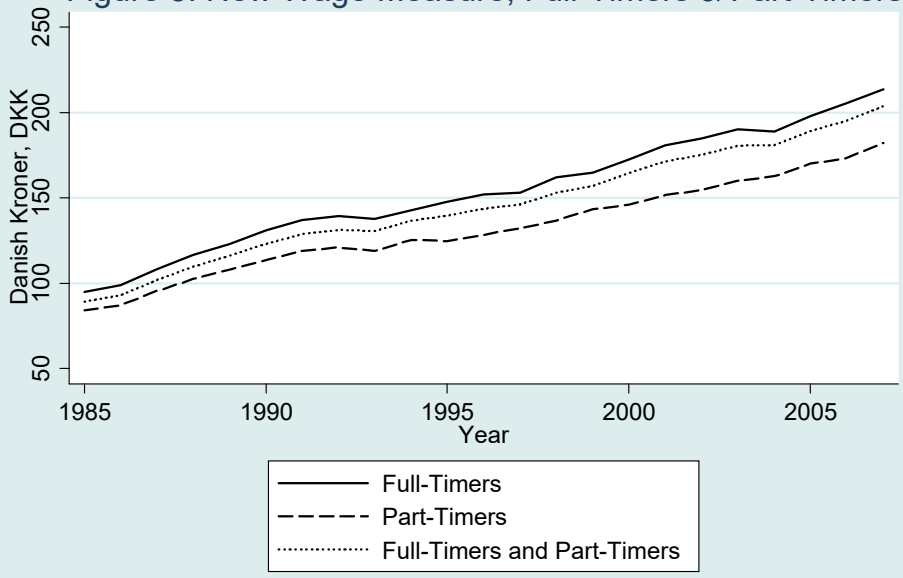
For workers in their first year:

$$t^* = t^*_{min} + \frac{1}{2}\begin{pmatrix} \left(p\left(\frac{u_{min}}{u}\right)\frac{u_{min}}{u} + \left(1 - p\left(\frac{u_{min}}{u}\right)\right)\frac{u_{top}}{u}\right)(1 + 1[G])\frac{h - \tau_3}{k} \\ + \left(1 - \frac{u_{min}}{u}\right)\frac{\tau_3 - \tau_2}{k} + \\ \left(\frac{u_{min}}{u} - \left(p\left(\frac{u_{min}}{u}\right)\frac{u_{min}}{u} + \left(1 - p\left(\frac{u_{min}}{u}\right)\right)\frac{u_{top}}{u}\right)\right) \cdot 1 \end{pmatrix} ku.$$

The performance can be gauged in Figures 8 to 11. In figure 8 we plot average wages for the full population and for full-timers and part-timers separately. A natural way compared the new wage measure and the IDA wage measure is to compare them both to Lønstatistiskken. This is done in Figure 9, 10A and 10B. The part-timer wage level has been reduced considerably but is still above the benchmark from Lønstatistiskken, among other reasons because the new criterion for reduced hours raises the level a bit. However, wages for full-timers correspond closely to those measured by Lønstatistikken. This is very promising, since Lønstatistikken only exist for workers employed in firms with more than 10 employees or in public employment.[15] Thus, at least for full-timers we can use the full population covered by our new wage measure instead of the restricted population from Lønstatistikken. In Figure 11 we estimate the tenure profiles and compare them across the wage measures. In general, the performance of the new wage measure is much better than that of the IDA wage, since there is no dip from the first to the second year.

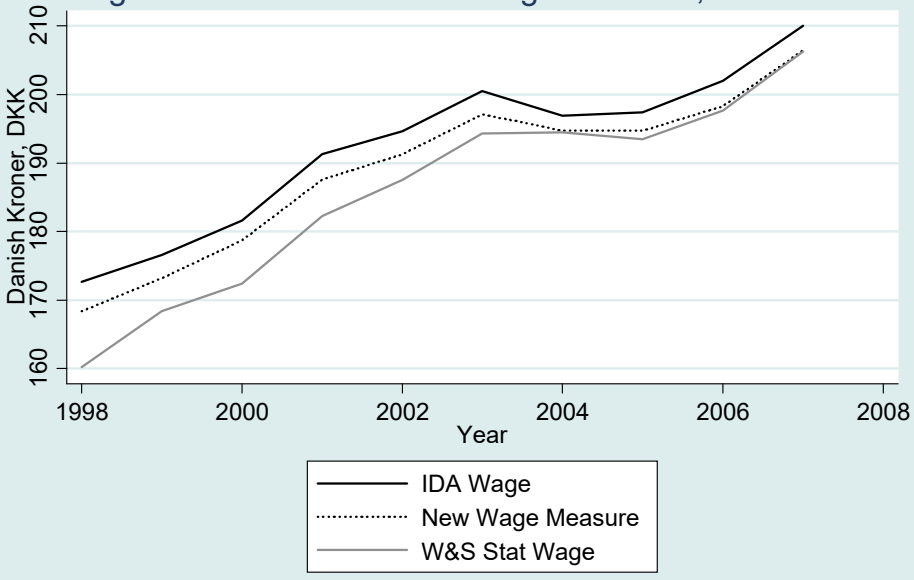
15. In our experience this criteria is somewhat lose, since one can find many violations in the data.

Figure 8: New Wage Measure, Full-Timers & Part-Timers



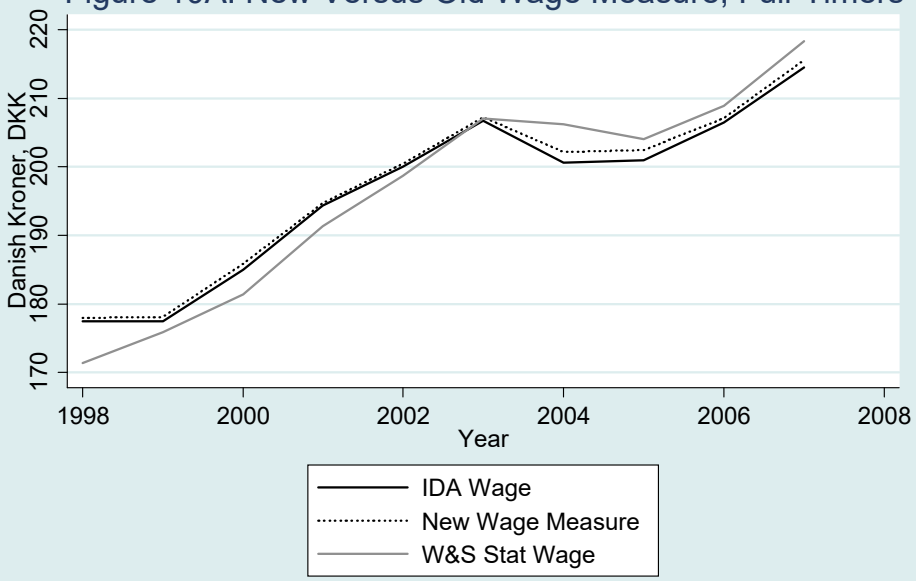
Source: Authors' calculations of wages in the Danish IDA database.

Figure 9: New Versus Old Wage Measure, All Workers



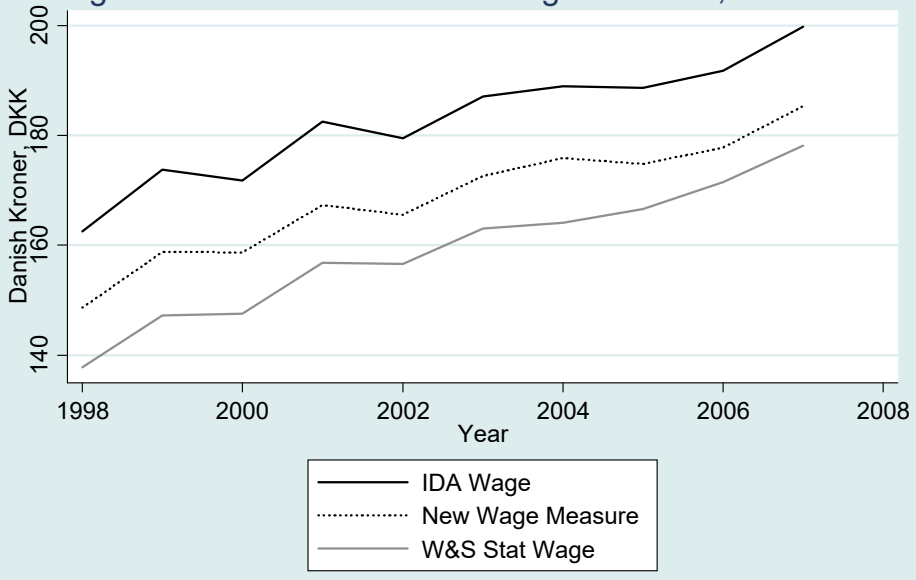
Source: Authors' calculations of wages in the Danish IDA database and the Lønstatistik (W&S Stat).

Figure 10A: New Versus Old Wage Measure, Full-Timers

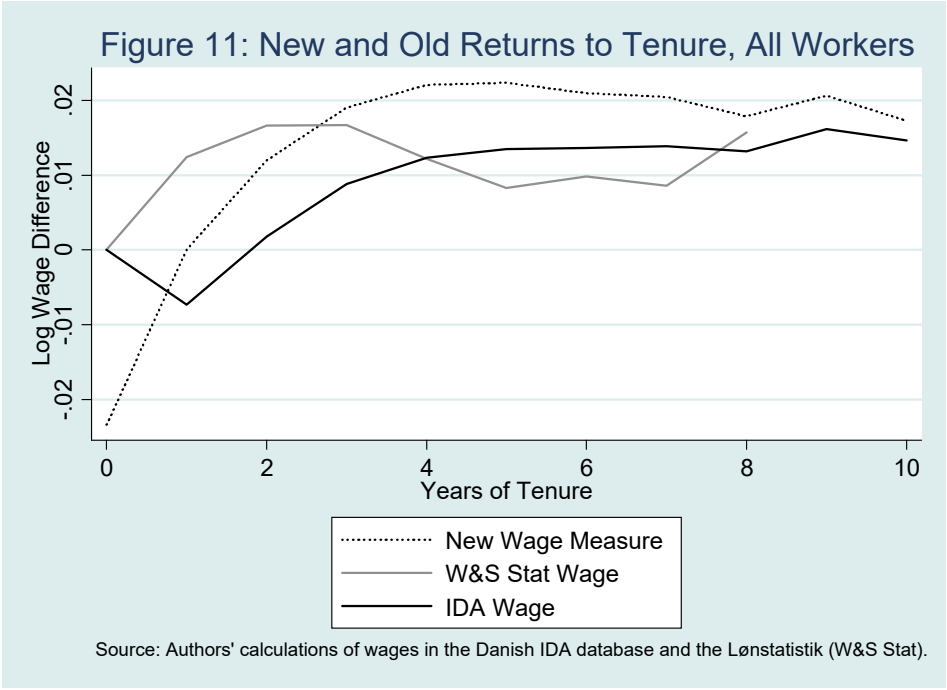


Source: Authors' calculations of wages in the Danish IDA database and the Lønstatistik (W&S Stat).

Figure 10B: New Versus Old Wage Measure, Part-Timers

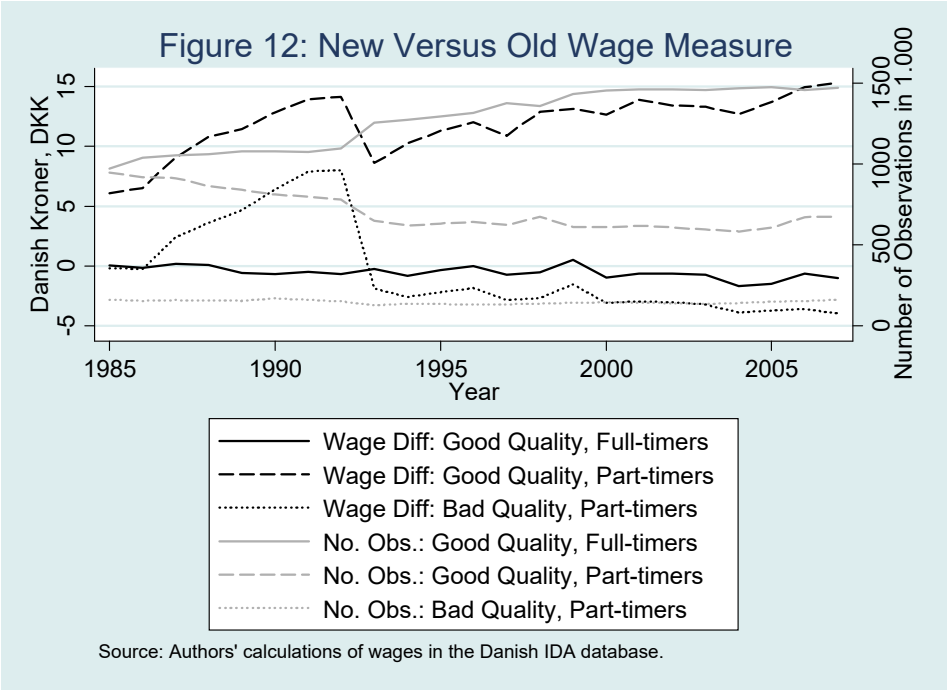


Source: Authors' calculations of wages in the Danish IDA database and the Lønstatistik (W&S Stat).



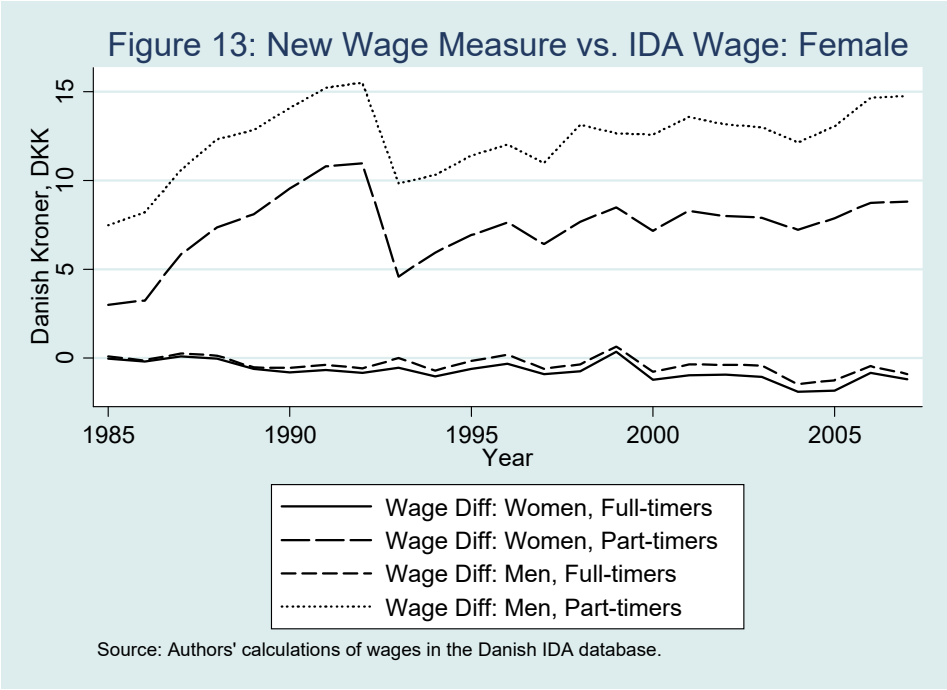
Note: We estimate the return to tenure by running the following regression $\log w_{it} = b_0 \cdot e_{it} + b_1 \cdot \tau_{it} + y_t + \alpha_i + \epsilon_{it}$. Here, w_{it} is the hourly wage of individual i, e_{it} is a vector of dummies capturing the overall labor market experience of worker i in year t, τ_{it} is a vector of dummies capturing the tenure of individual i in her job in year t, y_t is a year fixed effect, α_i is a worker fixed effect, and ϵ_{it} is an error term.

We now compare the difference between our new wage measure and the IDA wage across quality groups. Statistics Denmark has developed a quality indicator (TLONKVAL), which indicates if the wage is trustworthy. In Figure 12 we plot the difference between our wage measure and the IDA wage by full-timer/part-timer and quality status. If we first look at the wage differences we can see that the main differences in the two measures are coming from part-timers, but of both good and bad quality. Turning to the number of observations we can see that most of the part-timers are actually of good quality. Thus, selecting a sample of good quality from Statistics Denmarks wage measure will not make the results much more trustworthy.

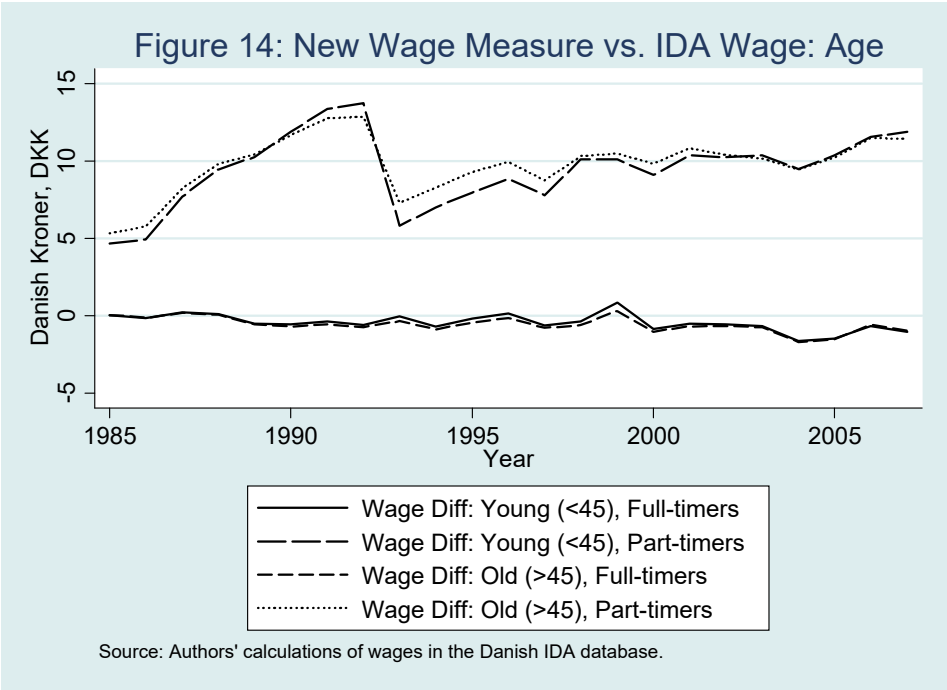


Footnote: The difference is TIMELON-NEW WAGE MEASURE. Thus a positive value indicates that TIMELON is higher than our new measure. There are almost no observations with "Bad Quality", "Full-timer" status, so these are omitted.

Finally, in Figure 13 and 14 we show the difference between our wage measure and the IDA wage measure across age and gender and across time. Again, it is clear that we primarily correct the wages of part-timers. However, this correction is somewhat more pronaunced for men and slightly more pronaunced older workers.



Footnote: The difference is TIMELON-NEW WAGE MEASURE. Thus a positive value indicates that TIMELON is higher than our new measure. There are almost no observations with "Bad Quality", "Full-timer" status, so these are omitted.



Note: The difference is TIMELON-NEW WAGE MEASURE. Thus a positive value indicates that TIMELON is higher than our new measure. There are almost no observations with "Bad Quality", "Full-timer" status, so these are omitted.

Discussion and Further Research

We now turn to discussing a number of areas where we think there is still room for improvement. Firstly, as mentioned earlier, the level of the part-timer wages still seems too high. One possible solution to this problem is to correct the IDA wage in a less conservative manner. As explained in Section 4, if the hours of part-timers vary a lot across pay cycles, we under-correct by assuming a minimum of variation. Instead, we could use an (ad hoc) average of the upper and lower bounds on the fraction of pay cycles in the top-bracket and in that way assume an intermediate amount of variation. Secondly, research should try to shed light on the extent to which the part-timer/full-timer classification used by Statistics Denmark distorts the compositions of samples and thus changes the outcomes of wage comparison exercises. For instance, some industries are more volatile than others in terms of the labour required. Agriculture is seasonal, construction is highly cyclical, and manufacturing sometimes depends strongly on market demand. In results reported in Chapter 3 of Lund and Vejlin (2015), we find that these industries have a higher proportion of workers just outside of the top-bracket, and a correspondingly

lower proportion in the top-bracket. We also find that the largest difference between the IDA wage and our Lønstatistik benchmark is in agriculture, and that the biggest reductions in the biases once we look at our new wage measure are in manufacturing and construction. These results indicate that the IDA wage is not so reliable when used to compare the wage levels across industries, since workers are pushed out of the top-bracket as in Section 4. Unfortunately, our new wage measure does not really reduce the bias in agriculture, and the bias is not particularly large in construction or manufacturing, which could indicate that there are other biases playing a role as well. This finding call for more research into whether the IDA wage can be used for cross-sectional wage comparisons at all, and into whether there are ways of eliminating the biases.

References

Journal Articles

Bagger, J., Fontaine, F., Postel-Vinay, F., and Robin, J. M. (2014): "Tenure, Experience, Human Capital and Wages: A Tractable Equilibrium Search Model of Wage Dynamics", American Economic Review, 104(6): 1551-96.

Buhai, I. S., Portela, M. A., Teulings, C. N., Van Vuuren, A (2014): "Returns to Tenure or Seniority?", Econometrica, 82(2), p. 705-730.

Lund, C. G., Vejlin, R.: "*Documenting and Improving the Hourly Wage Measure in the Danish IDA Database*", Department of Economics and Business, Economics Working Papers, 2015-06

Taber, C., Vejlin, R. (2016): "Estimation of a Roy/Search/Compensating Differential Model of the Labour Market", NBER Working Paper 22439

Online Resources

Online (1): Statistics Denmark's online documentation of the hourly wage in the IDA database. http://www.dst.dk/da/Statistik/dokumentation/Times/ida-databasen/ida-ansaettelser/timelon.aspx

Sources from Statistics Denmark

Statistics Denmark: "Hourly Wages and Working Hours in each Job", Arbejdsnotat 31, 1991.

Statistics Denmark: "Udviklingen i IDA's timeløn mellem 1992 og 1993", un-dated and anonymous memo.

Statistics Denmark: A sample of old program codes, un-dated and confidential.

Statistics Denmark: An Excel working file, un-dated and confidential.

Appendix A – Details of our Performance Analyses

The previous section documented the parametric form and historical parameter values of the estimation method of Statistics Denmark, and the Introduction highlighted a poor performance of the wage measure on four accounts – one time series issue in the period 1986-1993, another in the period 2003-2004, one level issue

among part-timers, and a falling wage from the first year of employment to the second year. This section provides the details of our performance analysis.

In terms of the sample, we use the longest possible time frame, 1985-2007. We also pick the most comprehensive population possible from the universe of all Danish establishments and workers with a job in the last week of November. In particular, we include all workers in the public sector and all other sectors, both men and women, full-timers and part-timers of all hours, and workers of all ages between 15 and 74, which are the age limits in the IDA database. However, due to data access issues and technical complications we discard the second jobs (type B-jobs). By only focusing on the main jobs, we lose 6.1 million observations or 9.8%. We also exclude workers with zero pensions or with periods of employment that are reported as zero or have been zero due to sickness or unemployment. We lose 6.8 million observations with these exclusions, or 12.0%. In total, we have 4.1 million unique workers, 0.51 million unique establishments, and 49.7 million pooled observations.

With the sample, we replicate the IDA wage measure and test it. The first case of a poor performance is the time series of the wages of part-timers from 1987 to 1993, documented in Figure 1 in the Introduction. Table 3 shows the magnitudes.

Table 3. Average Hourly Wages of Part-Timers and Full-Timers, 1985-1993

Year	Full-Timers, in DKK	Part-Timers, in DKK	Relative Difference, in %
1985	94.3	90.1	-4.5
1986	98.0	93.7	-4.4
1987	107.9	104.5	-3.1
1988	115.5	112.8	-2.4
1989	122.8	119.7	-2.5
1990	130.7	126.7	-3.1
1991	136.5	133.0	-2.5
1992	138.6	135.4	-2.3
1993	137.2	128.3	-6.5

Note: The average Relative Difference from 1994-2007 was -8.0%.

The next two poor performances of the IDA wage are the negative growth rate of the full-timer wage from 2003 to 2004 and the very high part-timer wage level (Figures 1 and 2 in the Introduction). In our assessments of these performances, we use a register of hourly wages from 1997 by Statistics Denmark called "Lønstatistik". The data contains a number of pay and time components that can be assembled to different kinds of hourly wage concepts. With our documentation of the IDA wage

in mind, we should be in a very good position to pick the right components and assemble them so that the concept is as close as possible to that of the IDA wage. We define what we call the "Wage and Salary Statistics wage" as

$$wage_{WS} = \frac{sftj - pens}{timprae + timfra - timover},$$

where $sftj$ is "samlet fortjeneste" i.e. a gross earnings measure, "$pens$" is pensions, "$timprae$" is hours worked, "$timfra$" is hours of certain kinds of absence, and "$timover$" is over-time.

Now, there are a number of sample issues that we must deal with since the samples in the IDA database and the Lønstatistik are far from identical. Firstly, the Lønstatistik is not a representative sample of the labour market, at least not in the form that we have access to. Indeed, firms run by the government and large firms are over-represented, and only 50% of all jobs are included. We use the skewed sample anyway. Secondly, since we only work with November jobs in the IDA database, we select the jobs in the Wage and Salary Statistics that seem to exist in November. However, it often happens that there are multiple such jobs for each worker in the Wage and Salary Statistics despite the fact that the comprehensive IDA database contains only one job. In that case we simply choose the job from the Wage and Salary Statistics with the highest earnings since we deem that job the most likely to exist in reality. Thirdly, there are large outliers in the Wage and Salary Statistics that we trim away by discarding 0.5% of both tails. Fourthly, the firm identifier cvrnr is sometimes missing or incorrect, compromising the sample match with the IDA database. The lack of correct firm identifiers is especially a problem in the initial years and accordingly our matched sample will improve over the years. We discard the year 1997 for this reason. Finally we match the two samples. Since the idea is to strictly compare the IDA and Wage and Salary Statistics wages for each worker and not compare the different samples, we discard all observations for which there is no counterpart in the other dataset.

Our constructed benchmark appears in Figures 2 and 3 for the years 1998-2007. Notice how the wage levels have risen compared to Figure 1 due to the use of the matched sample. There are three key points with the figures. Firstly, the full-timer IDA and Wage and Salary Statistics wages are very close to each other. The fact that the difference is so small for the group of workers that supposedly has the most precisely estimated wages in IDA (the full-timers) gives credence to the Wage and Salary Statistics wage as a useful and reliable benchmark despite its problems. Secondly, from 2003 to 2004 the Wage and Salary Statistics wages stay roughly constant, indicating a problem with the estimated IDA wage in this period. Thirdly, there is a very big difference between the Wage and Salary Statistics and estimated wages of the part-timers, of magnitude 20 DKK. Thus, it appears that the part-timer

wages are very poorly estimated in the IDA database. Indeed, different wage concepts or sample distortions are highly unlikely to explain the difference since the correspondence is so good for the full-timers.

We now turn to the fourth example of discomforting behaviour of the IDA wage, that of the falling wages from the first year of employment to the second. We restrict attention to the period 1994-2007 and first look at all workers together and run the following regression.

$$\log w_{it} = b_0 \cdot e_{it} + b_1 \cdot \tau_{it} + y_t + \alpha_i + \epsilon_{it}.$$

Here, w_{it} is the hourly wage of individual i, e_{it} is a vector of dummies capturing the overall labour market experience of worker i in year t, τ_{it} is a vector of dummies capturing the tenure of individual i in her job in year t, y_t is a year fixed effect, α_i is a worker fixed effect, and ϵ_{it} is an error term. The overall labour market experience is essentially measured as the total pensions accumulated since 1964 converted into full-time years and rounded down. Tenure is measured as the number of calendar years since hiring, so 0 in the first year and rising by one every year after that.

The object of interest from the regression is the vector of coefficients b_1 quantifying the relative increases in the wages since the year of hiring. The vector is the basis for the tenure profiles in Figure 4 in the Introduction[16] showing a highly unlikely fall in the IDA wage from the first year of employment to the second. Here we elaborate on that finding by showing that a similar pattern does not materialise among either part-timers or full-timers in isolation. As a result, the fall must be driven by a compositional change of the workforce from the year of hiring to the second year of employment.

Appendix B

To operationalize this idea, Statistics Denmark computes the following for main part-time jobs:

16. Note that the samples used for the two tenure profiles in the figure are different. We prefer using our standard sample for the IDA tenure profile (except that we use only the years 1994-2007) and not the matched sample used in Figures 1-3. For the Wage and Salary Statistics wage we also use the un-matched sample. We still pick the highest income job, but since there is no matching we now keep also the jobs without a cvrnr.Thus, the Wage and Salary Statistics wage tenure profile should not be seen as the exact true profile, but nonetheless it does indicate the size of the error.

T': The yearly full-time hours that apply for an employee with a period of illness or full-time unemployment. T' is proportional to T, with a factor of proportionality defined as u_{max} divided by $52 - V$.

t'_{min}: The lower bound on hours in the second job. Equal to zero if there is no second job.

T_{red}: Reduced number of working hours. $T_{red} = T' - t'_{min}$.

If the reduced number of working hours due to employment elsewhere is greater than the lower bound, $T_{red} > t_{min}$, then the second job is not sufficiently constraining, and the midpoint $0.5(t_{min} + t_{max})$ of the interval (t_{min}, t_{max}) is used as the number of hours in the main part-time job. On the other hand, if the reduced number of working hours is smaller than the lower bound on hours, i.e. if $T_{red} < t_{min}$, then it is very likely that the second job is a constraining factor, and the lower bound t_{min} on hours in the main part-time job is used as the estimate of hours.