EUROPEAN COMMISSION
EUROSTAT

Directorate F: Social statistics

Doc. DSSB/2017/Dec/3

Item 3

Adjusted gender pay gap

MEETING OF THE BOARD OF THE EUROPEAN DIRECTORS OF SOCIAL STATISTICS

LUXEMBOURG, 4-5 DECEMBER 2017

BECH BUILDING, ROOM B2/404

The purpose of this document is to inform the DSS Board about Eurostat's proposal to adjust the gender pay gap using the Structure of Earnings Survey data. Firstly, the gender policy context and the unadjusted gender pay gap indicator are introduced. Then, the adjustment method is outlined and the results are shown, including first estimations for the adjusted gender pay gap. Finally, the outcome of the discussion in the LAMAS Working Group is presented.

1. Policy background

Reducing the gender pay gap (GPG) is one of the key priorities of gender policies at the EU and national levels. At the EU level, the European Commission prioritised "reducing the gender pay, earnings and pension gaps and thus fighting poverty among women" as one of the key areas in its "Strategic engagement for gender equality 2016-2019".

One of the European Commission's actions in this area is the European Equal Pay Day (in early November) with Europe-wide information activities to raise awareness of EU citizens on possible discriminations between men and women. The unadjusted gender pay gap (unadjusted GPG) gets high visibility in this context as the key indicator to monitor progress in the field of equal earnings for men and women. The importance of the unadjusted GPG indicator is now reinforced by its inclusion in the Social Scoreboard of the European Pillar of Social Rights.

2. The unadjusted gender pay gap

The unadjusted GPG is calculated as the relative difference between the average hourly earnings of women and men. It thus provides a simple indicator of wage inequalities which explains its wide use by policy makers.

However, the unadjusted GPG entangles in its measurement both possible discrimination between men and women, in terms of "unequal pay for equal work", as well as the impact of differences in the average characteristics of men and women in the labour market.

2. The adjusted gender pay gap

Against this backdrop, Eurostat has developed a methodology to adjust the GPG using the Structure of Earnings Survey (SES) microdata (detailed methodology and results provided in the document annexed). The methodology is based on the Blinder-Oaxaca decomposition. The SES microdata provide information on the earnings of individual employees as well as on some personal, job and enterprise characteristics.

Within the decomposition approach, it must be decided which earnings structure constitutes the non-discriminatory benchmark against which to decompose the difference between hourly earnings of men and women. It is assumed, in accordance with the definition of the unadjusted gender pay gap, that the male earnings structure constitutes this benchmark.

Eurostat's project provides a decomposition of the difference between male and female earnings into explained and unexplained parts. The explained part is the gap between male and female earnings which is due to the differences in the average characteristics (sector of activity, age, occupation, etc.) of male and female employees. The unexplained part measures the difference between the financial returns to men and women with the same characteristics.

The adjusted GPG can be estimated as the unexplained part of the unadjusted GPG.

3. Results

Across EU Member States[1], the overall explained part varies from -281% in Romania[2] to 72% in Belgium. The explained part is negative in ten Member States: Bulgaria, Hungary, Italy, Ireland, Latvia, Lithuania, Luxembourg, Poland, Romania and Slovenia, zero in Portugal and positive in fifteen Member States. The positive (respectively negative) explained part means that men had, on average, better (respectively poorer) characteristics compared to women in the labour market. In the latter case, the adjusted GPG is higher than the unadjusted figures.

The explained part is mostly driven by economic activity, occupation, education and enterprise size. Economic activity accounts for more than 10% (irrespective of the sign of the adjustment) of the difference between the log hourly earnings of men and women in twenty-two Member States, occupation in seventeen, education in fourteen and enterprise size in twelve.

Across EU Member States, the adjusted GPG varies from 1.8% in Belgium to 22.7% in Lithuania, whereas the unadjusted GPG varies from 4.5% in Romania to 28.1% in Estonia. When comparing the ranking of EU countries (without Croatia and arranged from smallest to largest GPG) before/after adjustment we observe the most significant impact for Lithuania, Romania, Poland, Slovenia, Bulgaria and Italy moving by at least 10 positions downwards in the ranking as well as for Germany, Austria, the Netherlands and Denmark moving by at least 10 positions upwards.

4. Outcome of the LAMAS consultations and next steps

In July 2017, Eurostat consulted LAMAS as well as the contact persons in the NSIs responsible for providing the annual unadjusted gender pay gap, on a proposal to adjust the GPG. This proposal, that included a methodology and test results, was also discussed at the LAMAS meeting of 10-11 October 2017. Eurostat proposed, in particular, that the adjusted GPG indicator become the new headline indicator with the unadjusted GPG as the background indicator.

The consultations showed that a majority of LAMAS delegates were not in favour of making the adjusted GPG the new headline indicator instead of the unadjusted one..

LAMAS broadly agreed on the decomposition method proposed to calculate the adjusted GPG for countries and the EU as a whole. One the one hand, a group of Member States agreed with the Oaxaca decomposition method proposed by Eurostat (some of them already use the same method at the national level, e.g. Belgium, Finland and Germany). On the other hand, there was a number of Member States that were in favour of a more thorough discussion on the Oaxaca decomposition method pointing at some other possible options (e.g. using different earnings structures as non-discriminatory benchmarks) or some limitations in that method.

LAMAS agreed that Eurostat would publish the results as a methodological paper in "Statistics Explained" along with references to national adjusted GPG's, to make users aware that results depend on the data used and the assumptions made.

Eurostat will consult LAMAS in written on a layout to possibly publish the adjustment factors, alongside the adjusted GPG, as experimental statistics. Alternative labels for the "adjusted GPG" will also be proposed.

[1] No 2014 SES data for Croatia were available when the analysis was carried out.

[2] The results for Romania are provisional.

Eurostat hopes that the methodology proposed and the results obtained will stimulate further discussion within the European Statistical System on a common method to adjust the GPG indicator.

ANNEX

1. Introduction

Reducing the gender pay gap is one of the key priorities of gender policies at the EU and national levels. At the EU level, the European Commission prioritised "reducing the gender pay, earnings and pension gaps and thus fighting poverty among women" as one of the key areas in its "Strategic engagement for gender equality 2016-2019". One of the European Commission's actions in this area is to continue to mark the European Equal Pay Day (in early November) with Europe-wide information activities to raise awareness of EU citizens on possible discriminations between men and women. The unadjusted gender pay gap (unadjusted GPG) gets high visibility in this context as the key indicator to monitor progress in the field of equal earnings for men and women. The importance of the unadjusted GPG indicator is now reinforced by its inclusion in the Social Scoreboard of the European Pillar of Social Rights.

2. The unadjusted GPG

The unadjusted GPG is published annually by Eurostat in cooperation with the National Statistical Institutes (NSIs) of the EU Member States and the EFTA counties. It is based on the methodology of the Structure of Earnings Survey (SES) which is carried out with a four-yearly periodicity and on national sources for the years between the SES years (from reference year 2007 onwards). The scope and coverage of the unadjusted GPG are as follows: NACE sections B to S without O, enterprises with 10 employees or more, no restrictions for age and hours worked, and part-timers included. The definition of the unadjusted GPG, expressed in percentage, is as follows:

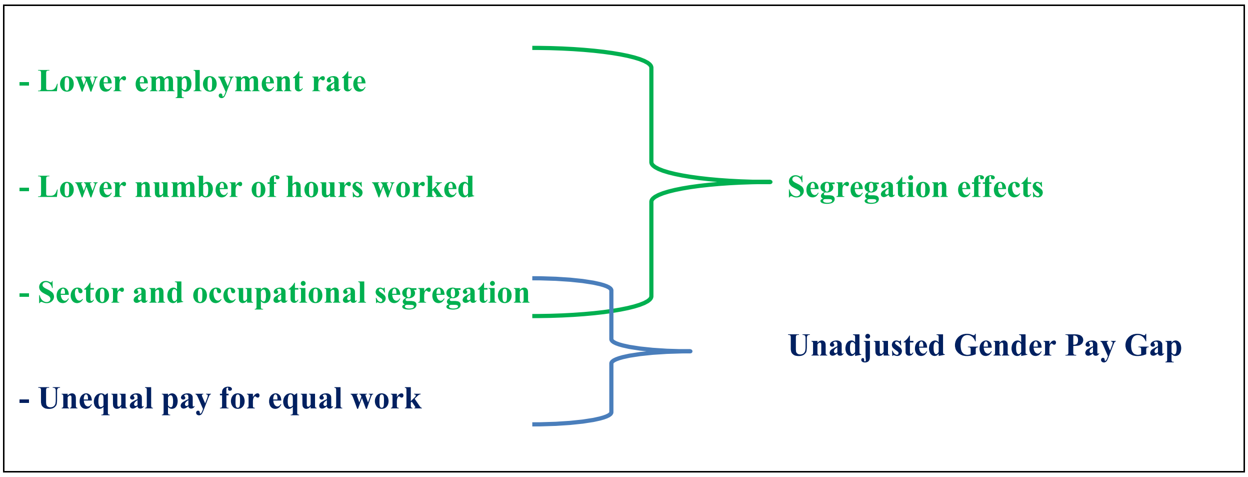
$$\frac{\text{Mean hourly earnings of men} - \text{Mean hourly earnings of women}}{\text{Mean hourly earnings of men}}$$

As an unadjusted indicator, the GPG gives an overall picture of the differences between men and women in terms of pay and measures a concept which is broader than the concept of equal pay for equal work. A part of the difference in earnings of men and women can be explained, among others, by sectoral and occupational gender segregations. Sectoral gender segregation means that women tend to be clustered in low paying sectors whereas men in high pay sectors. Occupational gender segregation is a result of men being more often promoted to supervisory and management positions compared to women. As a consequence, the unadjusted gender pay gap is rather a complex indicator which entangles in its measurement both the possible discrimination between men and women in the terms of "unequal pay for equal work" as well as the consequences of sectoral and occupational gender segregations in the labour market on the earnings of women compared with men.

It must be noted that the unadjusted GPG does not capture all segregation effects between men and women in the labour market. In particular, a higher proportion of women is 'inactive' (i.e. not seeking or not available to work) as illustrated by a lower employment rate. Moreover, the average number of hours worked by (active) women is lower than for men, which is not captured by the GPG which is calculated on an hourly basis.

Diagram 1 summarizes the main effects that play a role on the overall gap between the average earnings of all women of working age compared with men:

Diagram 1: decomposition of the 'overall' gender earnings gap



- Lower employment rate
- Lower number of hours worked
- Sector and occupational segregation
- Unequal pay for equal work
Segregation effects
Unadjusted Gender Pay Gap

In view of this, the European Commission's DG for Justice and Consumers expressed a strong interest in the decomposition of the unadjusted GPG into the share which is due to composition effects (different jobs held by men and women) and the remaining part (adjusted GPG) that would provide a better approximation of possible discrimination.

Against this backdrop, Eurostat developed a methodology to adjust the gender pay gap which was discussed at the LAMAS.

3. The adjusted GPG and adjustment factors

Eurostat used the Oaxaca decomposition (also called Blinder-Oaxaca decomposition) to estimate the adjusted GPG with the Structure of Earrings Survey (SES) as the data source. The main part of the document will only outline this method. More detailed information can be found in the annex.

The Oaxaca decomposition is carried out in two stages which are a regression analysis and a decomposition analysis. The regression analysis is carried out separately for the earnings structure of men and the earnings structure of women. The regression equations provide insights into the male and female earnings structures by showing the relationship between log hourly earnings and characteristics (observed personal, job and enterprise characteristics) for men and women, separately.

In a second step, the difference between the means of log hourly earnings of men and women is decomposed into two parts: an unexplained part and an explained part. Within the decomposition approach, it must be decided which earnings structure constitutes the non-discriminatory benchmark against which to decompose the difference between the means of hourly earnings of men and women. We assume, in accordance with the definition of the unadjusted gender pay gap, that the male earnings structure constitutes this benchmark.

The final decomposition equation for the difference between the means of log hourly earnings of men (M) and women (W) is as follows:

$$\Delta = \overline{\ln y}^{M} - \overline{\ln y}^{W} = \underbrace{\left(\hat{\beta}_0^M - \hat{\beta}_0^W\right) + \sum_{k=1}^{K} \bar{x}_k^W \left(\hat{\beta}_k^M - \hat{\beta}_k^W\right)}_{\textbf{\textit{Unexplained}}\text{ part of }\Delta} + \underbrace{\sum_{k=1}^{K} \hat{\beta}_k^M \left(\bar{x}_k^M - \bar{x}_k^W\right)}_{\textbf{\textit{Explained}}\text{ part of }\Delta}$$

where:

- $\overline{\ln y}$ represents the natural log of mean of hourly earnings of men (M) and women (W),
- $\bar{x}_k$, from k=1 to k=K, are the means of the variables covering the observed personal, job and enterprise characteristics for men (M) and women (W),
- β_0 is a constant and β_k, from k=1 to k=K, are the parameters for the corresponding variables covering the observed characteristics for men (M) and women (W).

The first part of the equation, which is marked as ***Unexplained***, measures the difference between financial returns to men and women. In other words, we calculate what a female worker with the average characteristics would have earned if she had been treated in the same way as a typical male worker, and compare these earnings with what she actually earns. This unexplained part could be considered as a better approximation of potential discrimination, i.e., how the labour market rewards men and women in an unequal way for the same work. However, the interpretation of this part as discrimination requires some caution as other explanatory variables not included in the regression (e.g. because they are unobserved), such as career breaks, may change the results of the decomposition. This is why the term "better approximation" is used.

The second part of the equation, which is marked as ***Explained***, measures the part of the difference between the means of log hourly earnings of men and women which is due to the differences in average characteristics for men and women

Each of the components, *Explained* and *Unexplained*, can be expressed as a proportion of the overall difference *Δ*. We can also express each subcomponent of *Explained* as a proportion of the overall difference Δ. This allows to estimate a magnitude of the effects of the specific characteristics explaining the overall difference Δ. Those subcomponents can be called adjustment factors.

Finally, the adjusted GPG can be calculated as follows:

$$GPG_{adjusted} = GPG_{unadjusted} * \left(1 - \frac{Explained}{\Delta}\right)$$

$$\text{or } GPG_{adjusted} = GPG_{unadjusted} * \frac{Unexplained}{\Delta}$$

This means that the adjusted GPG is the unexplained part of the unadjusted GPG.

4. Results and analysis

The 2014 data of the Structure of Earnings Survey (SES) are used in the decomposition and estimations of the adjusted GPG. Table 1 shows the overall explained part (*Explained*) and its subcomponents (adjustment factors) expressed as a percentage of the difference between log hourly earnings of men and women (*Δ*).

Table 1: The overall explained part and its components (adjustment factors) expressed as a percentage of the difference between log hourly earnings of men and women, 2014

Country	Overall explained	Age*	Education	Occupation	Job experience**	Employment contract	Working time	Economic activity	Enterprise size	Enterprise control
BE	**72**	2	18	38	2	-1	23	-8	-6	4
BG	**-39**	-3	-32	-39	-17	1	1	29	12	9
CZ	**18**	-3	0	2	0	1	1	17	0	0
DK	**46**	0	-3	8	-1	-3	-1	32	-4	18
DE	**70**	-1	4	10	2	0	32	25	-3	1
EE	**27**	3	-6	11	-2	0	3	22	-2	-1
IE	**-10**	2	-5	8	4	0	2	-4	-14	-4
EL	**35**	10	-16	-1	-1	0	11	32	-1	1
ES	**24**	2	-9	6	8	1	4	22	-8	-2
FR	**37**	1	-6	-3	1	0	-3	52	-2	-3
HR	**:**	:	:	:	:	:	:	:	:	:
IT	**-94**	-15	-57	-94	5	-3	64	52	-15	-32
CY	**6**	-5	-6	28	-8	0	1	7	-12	0
LV	**-8**	6	-31	-15	-16	1	0	74	-25	-2
LT	**-71**	-1	-31	-23	-32	0	5	32	-24	3
LU	**-50**	33	-16	-2	8	-3	8	-7	-31	-41
HU	**-17**	-5	-27	-47	-8	1	-21	90	-21	22
MT	**6**	12	-19	14	8	0	4	5	-17	-2
NL	**47**	11	-2	12	3	0	21	-32	-5	39
AT	**57**	0	1	18	8	-5	18	19	-1	0
PL	**-134**	-15	-98	-56	-12	-3	3	42	21	-14
PT	**0**	1	-34	9	1	-1	-1	37	-10	-2
RO***	**-281**	0	-147	-193	-44	0	10	222	-118	-10
SI	**-121**	-9	-90	-52	-24	-8	3	88	-6	-21
SK	**11**	-3	-6	-6	-2	0	4	18	1	5
FI	**38**	-2	-4	24	0	1	-1	16	-5	8
SE	**55**	-1	-10	-1	-1		3	57	-8	17
UK	**36**	-1	-1	28	1	-1	4	15	-4	-7
IS	**:**	:	:	:	:	:	:	:	:	:
NO	**42**	0	-9	-9	1	0	11	50	-5	3
CH	**43**	4	11	15	3	0	-3	13	-1	0

* Including the results for the "age" and "age squared" variables
** Including the results for the "job experience" and "job experience squared" variables
*** Provisional results excluding employees having weights in 5 percentiles of the largest sample weights.

Across EU Member States[3], the overall explained part varies from -281% in Romania[4] to 72% in Belgium. The explained part is negative in ten Member States: Bulgaria, Hungary, Italy, Ireland, Latvia, Lithuania, Luxembourg, Poland, Romania and Slovenia, zero in Portugal and positive in fifteen Member States (no data for Croatia was available when the analysis was carried out). The positive explained part means that men had, on average, better characteristics compared to women in the labour market, whereas the negative explained part means that women had, on average, better characteristics compared to men in the labour market. In the latter case, the adjusted GPG is higher than the unadjusted figures.

The explained part is mostly driven by economic activity, occupation, education and enterprise size. Economic activity accounts for more than 10% (irrespective of the sign of the adjustment) of the difference between log hourly earnings of men and women in twenty-two Member States, occupation in seventeen, education in fourteen and enterprise size in twelve.

As economic activity, occupation and enterprise size are categorical variables, the difference in average characteristics for these variables can be interpreted as gender segregation. For example, a positive adjustment factor for economic activity means that men worked, on average, in better paid NACE sections compared to women, leading to a lower adjusted GPG

[3] No 2014 SES data for Croatia were available when the analysis was carried out.
[4] The results for Romania are provisional.

compared with the unadjusted value. Conversely, a negative component means that women worked, on average, in better paid NACE sections compared to men, leading to an upward correction of the unadjusted GPG.

Among Member States, a negative adjustment factor for economic activity is observed only in Belgium, Ireland, Luxembourg and the Netherlands which is the only country with the proportion below -10%. On the other hand, in that country, a positive adjustment factor was recorded for the enterprise control meaning that the higher share of men compared to that of women work in the private sector where earnings are higher on average than in the public sector.

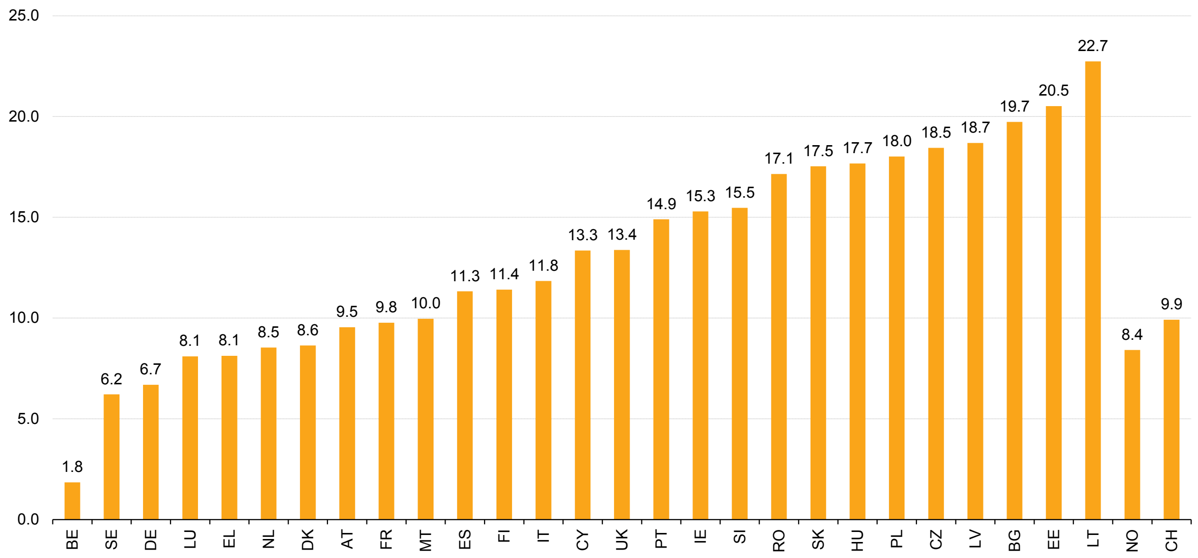
A negative adjustment factor for the enterprise size can be observed in all Member States, except Bulgaria, the Czech Republic (0%), Poland and Slovakia (1%). These factors are -10% or even below in 10 Member States and at 10% or above in Bulgaria and Poland.

A more mixed picture can be observed for occupation as the countries are split almost equally between those recording positive or negative adjustment factors.

As regards education, female workers had, on average, higher education than male workers in all Member States except Belgium, the Czech Republic, Germany and Austria. Education accounted for up to -10% of the difference between log hourly earnings of men and women in 13 Member States. Among all variables, it is the highest number of EU Member States recording a negative proportion of -10% or below.

Figure 1 shows the estimates of the adjusted GPG which are the results of applying a percentage of the overall explained part to unadjusted GPG presented in Figure 2.

Figure 1: Adjusted GPG (%), 2014



Across EU Member States, the adjusted GPG varies from 1.8% in Belgium to 22.7% in Lithuania, whereas the unadjusted GPG varies from 4.5% in Romania to 28.1% in Estonia.

Figure 2: Unadjusted GPG (%), 2014

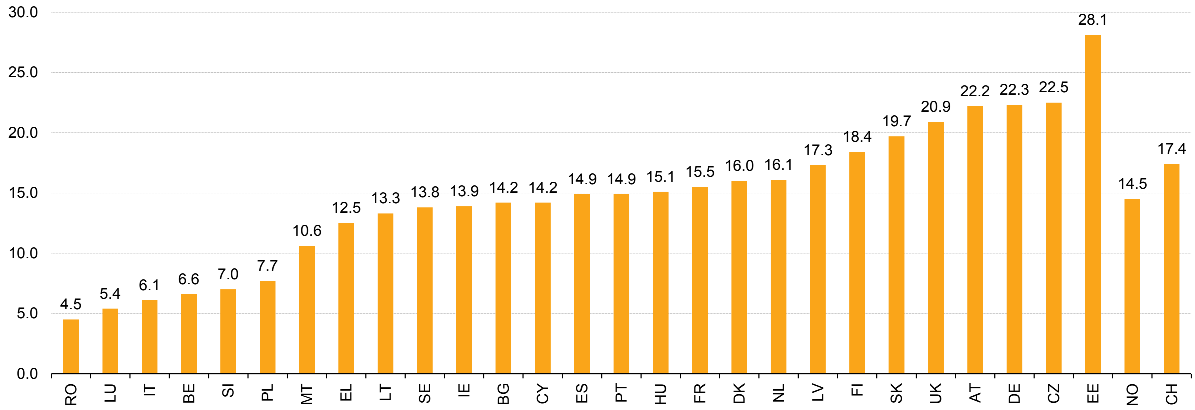
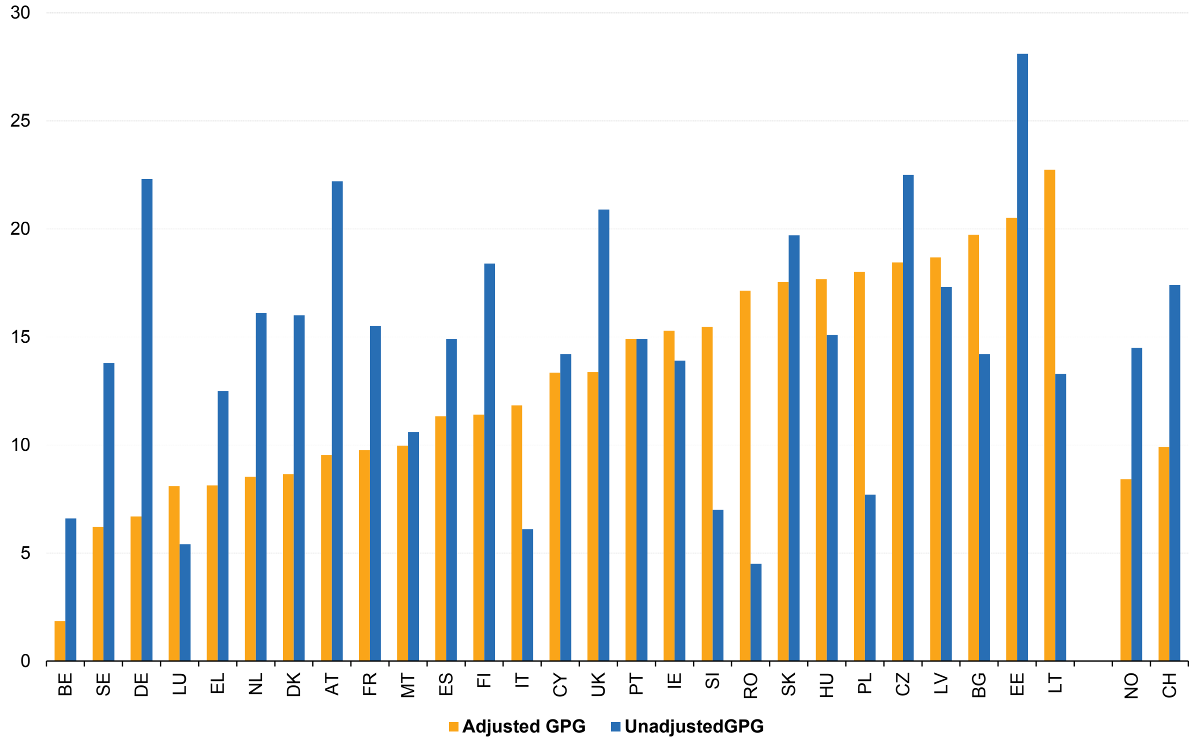


Figure 3 compares the adjusted and unadjusted GPGs. When comparing the ranking of EU countries (without Croatia and arranged from smallest to largest GPG) before/after adjustment we observe the most significant impact for Lithuania, Romania, Poland, Slovenia, Bulgaria and Italy moving by at least 10 positions downwards in the ranking as well as for Germany, Austria, the Netherlands and Denmark moving by at least 10 positions upwards.

Figure 3: Adjusted and unadjusted GPGs (%), 2014



The adjusted GPG for the EU has been calculated using the same method as for the EU unadjusted GPG, as the average of the adjusted GPGs in EU Member States weighted by the corresponding numbers of employees. The EU adjusted GPG, calculated for the whole economy except administration, is 11.2% against 16.7% for the unadjusted GPG in 2014.

5. Outcome of the LAMAS consultations and next steps

In July 2017, Eurostat consulted the LAMAS delegates and the contact persons in the NSIs responsible for providing the annual unadjusted gender pay gap, on a proposal to adjust the GPG. This proposal, that included a methodology and test results, was also discussed at the LAMAS meeting of 10-11 October 2017. Eurostat proposed, in particular, that the adjusted GPG indicator could become the new headline indicator with the unadjusted GPG as the background indicator.

The consultations showed that a majority of LAMAS delegates were not in favour of making the adjusted GPG the new headline indicator instead of the unadjusted one.

LAMAS broadly agreed on the decomposition method proposed to calculate the adjusted GPG for countries and the EU as a whole. One the hand, a group of Member States agreed with the Oaxaca decomposition method proposed by Eurostat (some of them already use the same method at the national level, e.g. Belgium, Finland and Germany). On the other hand, there was a number of Member States that were in favour of a more thorough discussion on the Oaxaca decomposition method pointing at some other possible options of the decomposition (e.g. using different earnings structures as non-discriminatory benchmarks) or some limitations in that method.

LAMAS agreed that Eurostat would publish the results as a methodological paper in "Statistical working paper" along with references to national adjusted GPG’s, to make users aware that results depend on the data used and the assumptions made.

Eurostat will consult LAMAS in written on a layout to possibly publish the adjustment factors, alongside the adjusted GPG, as *experimental statistics* in Eurostat's online database. Alternative labels for the 'adjusted GPG' will also be proposed.

6. Conclusions

Whereas the unadjusted GPG provides an overall measure of the gap between the hourly earnings of men and women, it cannot be interpreted as either an approximation or even an upper bound of possible discrimination in the sense of unequal pay for equal work which is the main concern of EU gender policy. Firstly, the part of the unadjusted GPG can be explained by the differences in the average characteristics of male and female workers. This explained part can vary significantly across the countries. Secondly, employed women may have, in some countries, better average characteristics than men. This is notably the case in some countries where the employment rate of women is lower than men which may create a selection bias in the population of the employed women. In such cases, the unadjusted GPG is lower than the adjusted figure.

For these reasons, the adjusted GPG indicator gives a better approximation of possible discrimination than the unadjusted figures. However, even the adjusted GPG should be interpreted with caution as other explanatory variables not included in the regression (e.g. because they are unobserved) may change the results of the decomposition.

Furthermore, the decomposition method allows measuring and comparing gender segregation effects in the European labour markets. These segregation effects can be analysed on the basis of the overall differences in average characteristics of men and women, e.g. in economic activities or occupations. Taken together, the adjusted GPG and the unadjusted GPG provide useful indicators to analyse the underlying causes of the differences recorded between the average earnings of men and women.

It is therefore Eurostat's intention to publish the decomposition methodology and results as "Statistical working paper".

The majority of Member States were in favour of keeping the unadjusted GPG as the headline GPG. Eurostat will conduct further consultations with the National Statistical Institutes aiming at publishing the adjusted GPG and the adjustment factors as 'experimental statistics'.

Eurostat hopes that the methodology proposed and the results obtained will stimulate further discussion within the European Statistical System on a common method to adjust the GPG indicator.

Annex: Oaxaca decomposition

Oaxaca decomposition, sometimes called Blinder-Oaxaca decomposition, is carried out in two stages which are a regression analysis and a decomposition analysis of the structure of earnings. In the first stage, a regression analysis is used to estimate the earnings equations separately for men (M) and women (W) as detailed in the following equations:

$$\ln y_i^M = \beta_0^M + \sum_{k=1}^{K} x_{ki}^M \beta_k^M + u_i^M$$

$$\ln y_i^W = \beta_0^W + \sum_{k=1}^{K} x_{ki}^W \beta_k^W + u_i^W$$

where:

- $\ln y_i$ represents the natural log of hourly earnings of individual i,
- x_{ki}, from k=1 to k=K, are variables covering the observed personal, job and enterprise characteristics that may impact on the log hourly earnings of individual i,
- β_0 is a constant and β_k, from k=1 to k=K, are the parameters for the corresponding variables covering the observed characteristics,
- u_i is a disturbance term for observation i.

Eurostat uses the observed characteristics which are available in the SES microdata provided by the countries (see Table 1).

Table 1: Observed characteristics covered by the variables x_k

Observed characteristics	Values	Codes of the corresponding SES variables
Personal and job characteristics		
Age	Age, age squared	Variable 2.2
Education	ISCED levels 0+1+2 , 3+4, 5+6, 7+8 *	Variable 2.5
Occupation	1-digit ISCO-08 occupation levels *[5]	Variable 2.3
Job experience (in the current enterprise)	Job experience in years, job experience in years squared	Variable 2.6
Employment contract	Indefinite duration, temporary duration, apprentice *	Variable 2.8
Working time	Full time, part time *	Variable 2.7
Enterprise characteristics		
Principal economic activity	All NACE rev. 2 sections except section O *	Variable 1.3
Enterprise size	Enterprises with 10-49, 50-249, 250-499, 500-999, 1000+ employees *	Variable 1.2
Enterprise control	Public, private *	Variable 1.4

* For the categorical characteristics, all categories without one (omitted category) are expressed as dummy variables.

[5] As suggested by some Member States, Eurostat will test using ISCO-08 at 2 digits instead of 1-digit level for the "occupation" variable.

After fitting separate regressions for men and women the following equalities hold for either male or female individual i:

$$\widehat{\ln y_i^M} = \hat{\beta}_0^M + \sum_{k=1}^{K} x_{ki}^M \hat{\beta}_k^M$$

$$\widehat{\ln y_i^W} = \hat{\beta}_0^W + \sum_{k=1}^{K} x_{ki}^W \hat{\beta}_k^W$$

These regression equations provide insights into the male and female earnings structures by showing the relationship between log hourly earnings and characteristics (observed personal, job and enterprise characteristics) for men and women. In this relationship, the estimated constants and coefficients measure the financial returns to male or female characteristics. Figure 1 shows this relationship for the bivariate regression with one explanatory variable.

In the second stage, we carry out a decomposition analysis of the difference between the means of log hourly earnings of men and women:

$$\Delta = \overline{\ln y}^M - \overline{\ln y}^W$$

The Oaxaca decomposition uses the following regression property for the means of log hourly earnings of men and women:

$$\overline{\ln y}^M = \hat{\beta}_0^M + \sum_{k=1}^{K} \bar{x}_k^M \hat{\beta}_k^M$$

$$\overline{\ln y}^W = \hat{\beta}_0^W + \sum_{k=1}^{K} \bar{x}_k^W \hat{\beta}_k^W$$

Within the decomposition approach, it must be decided which earnings structure constitutes the non-discriminatory benchmark against which to decompose the difference, Δ, between the means of log hourly earnings of men and women. We assume, in accordance with the definition of the unadjusted gender pay gap, that the male earnings structure constitutes this benchmark[6]. The constant and the coefficients in the men's equation are treated as the no-discriminatory benchmarks for the financial returns to characteristics of employees. In view of this, a *counterfactual* equation is constructed where the constant and coefficients in the women's equation are replaced by those of the men's equation:

$$\overline{\ln y}^{W*} = \hat{\beta}_0^M + \sum_{k=1}^{K} \bar{x}_k^W \hat{\beta}_k^M$$

This equation can be interpreted as what the average female worker would have earned if she had been paid on the same basis as an equivalent male worker (see also Figure 1 showing the Oaxaca decomposition for a single explanatory variable). The difference between the means of log hourly earnings of men and women can, then, be decomposed as follows:

$$\Delta = U + E$$

[6] The following other options are possible:
- the female earnings structure constitutes a non-discriminatory benchmark,
- both the female and male earnings structures constitute non-discriminatory benchmarks with some weighted average applied ,
- the whole population earnings structure constitutes a non-discriminatory benchmark.

where $U = \overline{\ln y}^{W*} - \overline{\ln y}^{W}$ and $E = \overline{\ln y}^{M} - \overline{\ln y}^{W*}$

The first part of the equation, which is marked as *U*, is the difference between the *counterfactual* and actual means of log hourly earnings of women. We calculate what a female worker with the average characteristics would have earned if she had been treated in the same way as a typical male worker, and compare these earnings with what she actually earns. After manipulation, this comparison, $\overline{\ln y}^{W*} - \overline{\ln y}^{W}$, can be expressed as:

$$U = \left(\hat{\beta}_0^M - \hat{\beta}_0^W\right) + \sum_{k=1}^{K} \bar{x}_k^W \left(\hat{\beta}_k^M - \hat{\beta}_k^W\right)$$

It measures the part of Δ which is due to the differences in the estimated constants plus the difference in the coefficients for men and women weighted by the averages of female characteristics. The difference between the constants and the difference between the coefficients measure the difference between financial returns to men and women. Figure 1 shows this difference for the bivariate regression with one variable. In the decomposition, the part *U* can be interpreted as the "unexplained" part of the difference in earnings between men and women (Δ). This unexplained" part could be considered as a better approximation of potential discrimination, i.e., how the labour market rewards men and women in an unequal way for the same work. Nevertheless, the interpretation of the *U* part as discrimination requires some caution as other explanatory variables not included in the regression (e.g. because they are unobserved) may change the results of the decomposition. This is why the term "better approximation" is used.

The second part of the equation, which is marked as *E*, is the difference between the actual mean of the log hourly earnings of men and the *counterfactual* mean of the log hourly earnings of women. This can be expressed as:

$$E = \sum_{k=1}^{K} \hat{\beta}_k^M \left(\bar{x}_k^M - \bar{x}_k^W\right)$$

It measures the part of Δ which is due to the differences in average characteristics of men and women weighted by the male coefficients. This represents the "explained" part (*E*) of the difference in earnings between men and women (Δ). Figure 1 shows this "explained" part for the bivariate regression with one explanatory variable.

The final decomposition equation for the difference between the means of log hourly earnings of men and women is as follows:

$$\overline{\ln y}^{M} - \overline{\ln y}^{W} = \underbrace{\left(\hat{\beta}_0^M - \hat{\beta}_0^W\right) + \sum_{k=1}^{K} \bar{x}_k^W \left(\hat{\beta}_k^M - \hat{\beta}_k^W\right)}_{\text{Unexplained}} + \underbrace{\sum_{k=1}^{K} \hat{\beta}_k^M \left(\bar{x}_k^M - \bar{x}_k^W\right)}_{\text{Explained}}$$

Each of the components, *E* and *U*, can be expressed as a proportion of the overall difference *Δ*. We can also express each subcomponent of *E,* i.e $E_k = \hat{\beta}_k^M(\bar{x}_k^M - \bar{x}_k^W)$, as a proportion of the overall difference Δ. This allows to estimate a magnitude of the effects of the specific characteristics explaining the overall difference Δ. However, the interpretation of the estimated coefficients can be misleading, when the explanatory variables are categorical with more than two categories. This is due to the coefficient estimates for the categories of a categorical variable being sensitive to a choice of the reference category. In the explained part

E, nevertheless, the overall effect of the categorical variable (e.g., the overall effect of economic activity) in the decomposition does not depend on the omitted category. It is, however, not the case for the unexplained part *U* where the overall effect of a categorical variable depends on the omitted category. For this reasons only the overall results for the unexplained part *U* can be taken into account and the results of the detailed decomposition for the unexplained component will not be interpreted.