Comparison of the Three Most Prevalent Pay Equity Analysis and Remediation Methods

Methodology White Paper 2020

by Zev J. Eigen, Allison Adams + Andrea Palmiter

Fair Pay

Workplace

Table of Contents

Introduction

Pay equity has taken center stage as laws in the United States, the UK, EU, and elsewhere ramp up, as the social pressure for gender and racial equity mounts, and, businesses everywhere need to hire and retain their best talent, especially in volatile economic times. There is a lot of misleading information out there, and blind reliance on experts is perilous. Getting pay equity analyses right requires knowledge and expertise in pay equity laws and their enforcement, statistics and an understanding of compensation and benefits best practices. Ensuring that pay equity is accomplished optimally is more important than ever when economic forecasts are uncertain. As budgets tighten, allocating scarce dollars effectively and efficiently is crucial for many businesses that want to remain in compliance to reduce risk of catastrophic legal, HR and PR risks.

This paper sheds a much needed light on the choices faced by organizations that want to get pay equity right to honor their commitment to this important social and legal issue. The paper compares the three most commonly applied methodologies for identifying and remediating pay equity concerns in an organization. We use a data set borrowed from a real employer that has been anonymized with permission to illustrate the similarities and differences across these methods. Specifically, we show when each method identifies concerns at the group and individual levels, to whom and how much each method recommends providing remuneration, and most importantly, whether each method succeeds in eliminating cause for concern from a statistical/analytic perspective for any groups in which there is initial evidence of a pay equity issue.

Pay Equity Statistical Review

The goal of pay equity reviews can be to comply with pay equity laws at the applicable state and federal levels in the United States, or globally, where applicable. It can also be more broadly defined as trying to do what's right or fair, (even if not compliant with laws), which can be elusively subjective. This paper adopts the assumption that most employers prefer a method that is likely legally compliant, and may be reasonably regarded as fair as the primary motivation.

At a minimum, employers seeking first to comply with laws may rely on appropriate methods that enable the accomplishment of that goal as a prerequisite for taking other actions that may make things more uniform or objective. The rationale is to first see if the playing field is level. If it is not, then level it. Then, once the playing field is leveled, then other means may be adopted to increase the uniformity with which policies and practices are doled out to prevent future inequities, or to ensure more uniformity and consistency generally.

Three components of Pay Equity Laws

1.

Substantially Similar Groups or "SSGs"

First, employers need to group their employees based on performance of work that is "substantially similar," "comparable," or "equal" (or similar wording) based on employees' skill, effort, accountability and working circumstances. This paper uses the acronym "SSG" to denote a "substantially similar group" of employees. Job titles are often too granular as a grouping schema because there are commonly job titles that could reasonably be coalesced into SSGs. For instance, "assistant" and "administrative aid." Job level or grade may be too broad. So, it is critical that significant attention be paid to how employers form SSGs for the purpose of reviewing for pay equity compliance. For simplicity's sake, the employer data used in this paper contains six SSGs of varying headcounts. Summary statistics for the groups is displayed in Table I.

2.

Statistical Evidence of Differences in Wages in SSGs

Second, employers need to statistically estimate whether employees in SSGs are paid less because of their protected category status. All statutes include gender as a protected category. Missouri and Montana only permit women to bring suits, not men. All other statutes permit all genders to bring suits. Some other jurisdictions include race as a protected category, and still other jurisdictions include other categories such as sexual orientation, age and military status (among others). Evaluating whether there is statistical evidence of a protected class being paid less is a straightforward matter of parametrically testing whether and to what extent the distribution of compensation for men varies statistically significantly from the distribution of compensation for women. For simplicity's sake, this paper will use the comparison of gender (male vs. female, or sometimes more broadly defined as "male" vs. "non-male"), however the same math and methods applies to the analysis of any other protected category (e.g. white vs. nonwhite).

3.

Defenses When There is Statistical Evidence of Differences in Wages in SSGs Because of a Protected Category Status

Third, pay equity laws at the federal and state levels enumerate defenses of which employers may avail themselves if and when there is statistical evidence of gaps that are concerning. These defenses are typically neutral, job-related factors that measure disparities in the quantity and quality of output of work. Some statutes include a "catch all" defense of gender neutral, job related criteria. Examples of these factors are tenure, years of relevant work experience, location and educational attainment. The universally accepted statistical test for this step is multivariate regression modeling. Including a coefficient in the model associated with individuals' gender is an accepted technique to evaluate whether, net of the control variables (defenses), gender may not be ruled out as a contributing explanatory factor. All three of the methods described below rely on regression modeling in varying ways.

Overview of Data and the Three Methods

Below this paper summarizes each of the methods under scrutiny. The data set used for this comparative analysis consists of 496 employees spanning six SSGs. The summary statistics for each are shown in Table I below.

Table I: Summary Statistics of the 6 Groups

SSG	Headcount	Mean	STD	MIN	25%	50%	75%	MAX
Engineering	77	$ 126,179	$ 15,106	$ 58,200	$ 115,380	$ 129,760	$ 132,540	$ 152,040
IT	30	$ 104,847	$ 13,852	$ 83,043	$ 94,534	$ 102,430	$ 111,504	$ 132,000
Research and Development	88	$ 94,600	$ 52,698	$ 45,750	$ 58,954	$ 66,335	$ 109,410	$ 291,000
Sales	110	$ 53,669	$ 23,729	$ 20,000	$ 31,110	$ 51,340	$ 70,955	$ 101,460
Support	51	$ 91,880	$ 30,763	$ 40,000	$ 65,500	$ 90,000	$ 111,900	$ 162,000
Technicians	141	$ 145,262	$ 52,729	$ 40,000	$ 94,000	$ 140,000	$ 190,440	$ 252,000

The first three methods rely on confidence intervals as the primary way of flagging people for review. To replicate this approach, this paper uses z scores to determine the number of standard deviations from the mean a data point is for each individual. We calculate the z scores for each individual and estimate how many employees are more than the acceptable limits away. **In replicating the most common methods, this paper uses the following levels:**

- 70% Confidence Interval = ± 1.04 SDs
- 80% Confidence Interval = ± 1.28 SDs
- 90% Confidence Interval = ± 1.65 SDs
- 95% Confidence Interval = ± 1.96 SDs

In the descriptions below in Methods 1 through 3, this paper uses the term, "outlier" to refer to individuals flagged as beyond the defined parameters of the applied confidence intervals.

Method 1: One Overall Fixed Effect Model

Identify and adjust outliers using a model that spans the entire data set (whole organization), controlling for assignment to SSG (fixed effect model).

1(a): Remediate by adjusting flagged negative outliers regardless of gender. So, both male and female employees are eligible to be flagged to receive a remedy.

1(b): Adjust only the negative outliers identified in the disadvantaged protected category.

The convenience and ease of running one fixed effect regression model is the main attraction of Method 1. However, there are many reasons to suspect that the ease of running one model would not be outweighed by three problems: First, this model is almost certain to both under- and over-correct, making the costs saved in running one simple model almost universally likely outweighed by the costs of over-remediating unnecessarily. This method almost invariably generates false positives (flagging for remediation groups that show no statistical evidence of concerning gender differences), and over and under remediates at the individual level. Second, with one model, one may not tailor controls to groups. This is absolutely critical. For controls to be viable, they need to be neutral and job related. What about controls like educational attainment that may be job related for some SSGs like “Research and Development,” but are not applicable at all to SSGs like “Support,” or “Facility Services?” This Method doesn’t allow for tailorability and that lack of flexibility is a significant disadvantage here. Third, Method 1’s legal defensibility is highly questionable. For instance, if the employer of the dataset used herein were sued based on the 88 people in the “Research and Development” group, the appropriate regression model used by the Agency, Plaintiff or Defendant would be to regress compensation on gender and the applicable valid control variables for the 88 employees in that SSG. Method 1 suffers because the method applied here is not the same as the method most likely applicable in the event of a suit.

Method 2: **Group by Group Model; Standard Deviation/Outlier Approach**

Identify and adjust outliers using multiple regression models for each SSG large enough for robust statistical analysis.2(a): Remediate by adjusting flagged negative outliers regardless of gender, just like Method

> **1(a):** but the models are siloed by SSG. So, both male and female employees are eligible to be flagged to receive a remedy here as well.
>
> **2(b):** Adjust only the negative outliers identified in the disadvantaged protected category within the SSG.

Method 2 avoids the problems of trying to fit a single model to the data. It also avoids the problem of applying controls to groups to which they should not be applied. Assuming that it first evaluates whether there is a statistically significant gap net of controls applied in the multivariate regression model, the primary downside of this method is likely requiring greater average per person cost to remediate than needed. This Method likely produces a higher per person cost to remedy, and a smaller number of people to whom a remedy would be offered. That is, with Method 2, more money is given to a smaller number of people. For both Methods 1 and 2, there is a risk that outlier identification and compensation will only improve the pay for a small subset of statistically disadvantaged women (when evidence from steps 2 and 3 show women as the statistically disadvantaged protected individuals). This in turn, could mean that even after applying Methods 1 or 2, the SSG could still show a statistically significant gender coefficient when the regression is re-run with the proposed remediation figures. So, Methods 1 and 2 run the risk of failing to optimally remediate, and in a way that is consistent with legal requirements. They may be regarded as helping pay people more fairly, because compressing distributions means more uniformity, however, there is nothing in any laws that forbid employers from paying people 1, 2, 3, or even 4 standard deviations below the mean as long as both men are women are equally treated, and there is no statistical evidence that women are paid less because of their gender.

The remediation approaches in Methods 1(a) and 2(a) omit the question of whether the distribution of wages for women is statistically different from the distribution of wages for men. Instead it asks whether there are people (male or female) who are different from the overall mean. Because this method skips a critical step, the remediation philosophy for 1a and 2a may not help employers comply with laws. Instead, some suggest that it is focused on fairness. However, here, even the well-intentioned focus on fairness could be regarded as fundamentally unfair. The argument in favor of this method goes that by compressing the distribution of pay for men and women alike, the employer would be treating all employees fairly/equally. However, this point is severely undermined by the fact that treating men and women equally only makes sense if there were a level playing field to start. It does little or nothing to address problems if women are underpaid because of their gender. Like the cartoon in Figure I, offering men and women an equal distance for a race, does little to rectify existing underlying problems. By paying both men and women more, in situations in which there is evidence of systematic gender bias in favor of one gender, this method is likely to exacerbate potential legal concerns.

Figure I: Cartoon Illustrating the Problem of Methods that Fail to First Level the Playing Field

Method 3: **Group by Group Model; Statistical Significance/ "But For" Gender Status**

Identify SSGs showing statistically significant gaps without any controls applied. Apply parametric tests (t-tests) to evaluate whether men and women are paid differently because of gender. For groups in which there is statistical evidence at p-value $< .05$, evaluate whether and to what extent the regression model including gender and the applicable control variables impacts the p-value for the gender coefficient to render it no longer statistically significant ($p \geq .05$). Adjust people in the disadvantaged protected class based on this model, net of the gender coefficient regardless of their position on the distribution (regardless of whether they are "outliers"). Use the net predicted values derived from the regression model as a guide for remediation to disadvantaged individuals. This takes into account the degree to which people are likely paid less because of gender whether they are at the bottom, middle or the top of the distribution if statistical evidence suggests that they likely would have received greater compensation but for their gender status.

Comparison of Methods' Identification of Concern at the Group Level

As a starting point, it is helpful to compare standard accepted parametric testing results without controls applied to the results controlling for tenure (measured in years working with the organization from the most recent date of hire). Table II below shows these results. The Support group shows the most robust evidence of a gender gap that may not be ruled out at the 95% confidence level both with and without controlling for tenure. The R&D group is barely outside of statistical significance without controlling for tenure (p =.06) and is statistically significant when controlling for tenure (p = .01). The Technicians group is not statistically significant after controlling for tenure (p = .063), but is statistically significant without any controls applied. The other three groups, Engineering, IT and Sales all show no statistical evidence that gender is a factor in the variation in compensation for employees in those groups.

Table II: Which Groups Show Statistically Significant Gaps by Gender?

SSG	T-test p values	Gender coefficient p value controlling for tenure
Engineering	0.295	0.387
IT	0.205	0.183
R&D	.06*	.01**
Sales	0.118	0.101
Support	.03**	.019**
Technicians	.015**	.063*

As noted above, this paper compares outcomes across the methods based on four criteria:

1. Which groups are flagged for review?
2. What is the recommended budget to remedy flagged groups?
3. To how many individuals is a remedy recommended?
4. After each method's recommended remedy was applied, was the method successful in no longer showing any statistical evidence that gender was explaining variation in compensation?

Comparison 1

Which Group(s) Are Flagged for Review?

The first point of comparison across the three methods is whether a group is flagged as showing evidence of concern using the different methods noted above.

Method 3 stands alone in its reliance on p-values at the group level as opposed to simple standard deviation measurements as explained above. To compare apples-to-apples, we included a version of Method 3 that uses a p-value of .10 to compare with the other methods where a 70% confidence interval is applied, and a .05 p-value threshold in the 90% confidence interval comparison set.

As demonstrated in Table III below, there is some variation in how the methods flag SSGs for review. Method 1 (a and b) over-identify groups showing concern. This is likely due to the unobserved interaction among the effect of being in a group with other included coefficients (tenure).

Otherwise, there is consensus among Methods 2 and 3 that R&D and Support are flagged for review. The exception to this is if Methods 1 or 2 are applied at the 70% confidence level. In that case, the Technicians group is not flagged for review. By comparison, Method 3 (applying a p value of .10 as the threshold), would flag the Technicians group for review.

70% Confidence Interval // p < .10 for Method 3			
SSG	**Method 1**	**Method 2**	**Method 3**
	1 overall model	*Group Basis/Std Dev*	*Group Basis/ p-value*
Engineering	Yes	No	No
IT	Yes	No	No
R&D	Yes	Yes	Yes
Sales	Yes	No	No
Support	Yes	Yes	Yes
Technicians	Yes	No	Yes

80% Confidence Interval		
SSG	**Method 1**	**Method 2**
	1 overall model	*Group Basis/Std Dev*
Engineering	No	No
IT	Yes	No
R&D	Yes	Yes
Sales	Yes	No
Support	Yes	Yes
Technicians	Yes	No

90% Confidence Interval // p < .05 for Method 3				
SSG	**Method 1a**	**Method 1b**	**Method 2**	**Method 3**
	1 overall model (M & F eligible for remedy)	*1 overall model (disadvantaged group only eligible for remedy)*	*Group Basis / Std Dev*	*Group Basis / p-value*
Engineering	No	No	No	No
IT	Yes	Yes	No	No
R&D	Yes	Yes	Yes	Yes
Sales	Yes	Yes	No	No
Support	Yes	No	Yes	Yes
Technicians	Yes	Yes	No	No

Comparison 2

What is the Recommended Budget to Remedy Flagged Groups?

Each method may be used to generate a budget to remediate the identified statistically concerning group or groups. Table IV below represents the array of remediation options for each of the methods.

There are two observations worth noting:

1. Across all methods, narrowing the bandwidth or tolerance decreases the recommended budgets. That is, going from the 70% level to 80% to 90% for Methods 1 and 2, and from a p-value threshold of .10 to .05 for Method 3, consistently decreases the recommended budgets across the board as expected. So, great care should be taken in selecting a confidence interval or p-value threshold if different from 90% (for Methods 1 and 2) or a p-value greater than .05 for Method 3.

2. Looking at the last chart in Table IV, Method 2(b) is the smallest at $53k, followed by Method 2(a) at $59.6k. Methods 1(a) and 1(b) are in the middle at $171k and 152k respectively. Applying Method 3 yields the largest remediation amount at $348k.

Table IV: Recommended Budgets by Method

70% Confidence Interval (~1 STD) // p < .10 for Method 3					
SSG	**Method 1 (a)**	**Method 1 (b)**	**Method 2 (a)**	**Method 2 (b)**	**Method 3**
	1 overall model (M & F eligible for remedy)	*1 overall model (disadvantaged group only eligible for remedy)*	*Group Basis / Std Dev (M & F eligible for remedy)*	*Group Basis / Std Dev (disadvantaged group only eligible for remedy)*	*Group Basis / p-value (disadvantaged group only eligible for remedy*
Engineering	$ 11,026	$ 8,217	$ -	$ -	$ -
IT	$ 43,736	$ 39,071	$ -	$ -	$ -
R&D	$ 76,192	$ 62,333	$ 110,483	$ 100,097	$ 352,015
Sales	$ 151,358	$ 144,194	$ -	$ -	$ -
Support	$ 65,437	$ 55,148	$ 115,085	$ 93,069	$ 176,598
Technicians	$ 359,616	$ 330,327	$ -	$ -	$ 75,169
Total	**$ 707,364**	**$ 639,290**	**$ 225,568**	**$ 193,166**	**$ 603,781**

80% Confidence Interval (~1.3 STD)					
SSG	**Method 1 (a)**	**Method 1 (b)**	**Method 2 (a)**	**Method 2 (b)**	**Method 3**
	1 overall model (M & F eligible for remedy)	*1 overall model (disadvantaged group only eligible for remedy)*	*Group Basis / Std Dev (M & F eligible for remedy)*	*Group Basis / Std Dev (disadvantaged group only eligible for remedy)*	*Group Basis / p-value (disadvantaged group only eligible for remedy*
Engineering	$ -	$ -	$ -	$ -	
IT	$ 27,067	$ 24,759	$ -	$ -	
R&D	$ 29,976	$ 24,787	$ 47,637	$ 47,637	
Sales	$ 109,686	$ 102,975	$ -	$ -	
Support	$ 40,434	$ 30,417	$ 72,227	$ 56,388	
Technicians	$ 220,583	$ 196,283	$ -	$ -	
Total	**$ 427,747**	**$ 379,221**	**$ 119,865**	**$ 104,025**	

90% Confidence Interval (~1.65 STD) // p < .05 for Method 3					
SSG	**Method 1 (a)**	**Method 1 (b)**	**Method 2 (a)**	**Method 2 (b)**	**Method 3**
	1 overall model (M & F eligible for remedy)	*1 overall model (disadvantaged group only eligible for remedy)*	*Group Basis / Std Dev (M & F eligible for remedy)*	*Group Basis / Std Dev (disadvantaged group only eligible for remedy)*	*Group Basis / p-value (disadvantaged group only eligible for remedy*
Engineering	$ -	$ -	$ -	$ -	$ -
IT	$ 14,152	$ 12,050	$ -	$ -	$ -
R&D	$ 11,969	$ 10,442	$ 25,976	$ 25,976	$ 239,104
Sales	$ 57,871	$ 54,387	$ -	$ -	$ -
Support	$ 1,888	$ -	$ 33,653	$ 27,336	$ 108,675
Technicians	$ 85,550	$ 74,975	$ -	$ -	$ -
Total	**$ 171,430**	**$ 151,854**	**$ 59,629**	**$ 53,312**	**$ 347,780**

Comparison 3

To How Many Individuals is a Remedy Recommended?

This measure shows very stark differences among the Methods, but especially between Method 3 on the one hand, and Methods 1 and 2 on the other. First, as demonstrated in Table V, Method 3 offers a remedy to a vastly larger percentage of impacted employees.

On average, nine times as many women receive a remedy when Method 3's recommended monetary changes are implemented as compared to the other Methods. This is because Method 3 reflects a philosophy that employees are counted as being "due" for a remedy if, net of the gender gap, there is statistical evidence that they are likely getting paid less than expected because of their gender (or at least there is probabilistic statistical evidence that suggests that the variation in compensation due to gender may not be ruled out applying the standard universally accepted statistical threshold of "statistical significance" at a p-value of less than .05). Put another way, with Method 3, it is completely irrelevant where people in the disadvantaged protected category fall on the distribution.

They could be at the top of the range, the middle or the bottom. If the model shows that the individual would have been paid more but for their protected category status, they are counted in the group to whom a remedy may be due. The other Methods, by contrast, do care where people are on the distribution; people in the middle or the top are excluded. This difference in approach will almost invariably lead to the results shown in this paper-- budgets to correct being allocated to fewer people.

Table V: Percentage of Female Employees Receiving a Proposed Remedy by Method

SSG	Method 1(a)	Method 1(b)	Method 2(a)	Method 2(b)	Method 3
	1 overall model (M & F eligible for remedy)	*1 overall model (disadvantaged group only eligible for remedy)*	*Group Basis / Std Dev (M & F eligible for remedy)*	*Group Basis / Std Dev (disadvantaged group only eligible for remedy)*	*Group Basis / p-value (disadvantaged group only eligble for remedy*
Engineering					
IT	8%	8%			
R&D	3%	3%	3%	3%	79%
Sales	8%	8%			
Support	12%		15%	12%	65%
Technicians	13%	13%			
Overall	**9%**	**9%**	**8%**	**7%**	**73%**

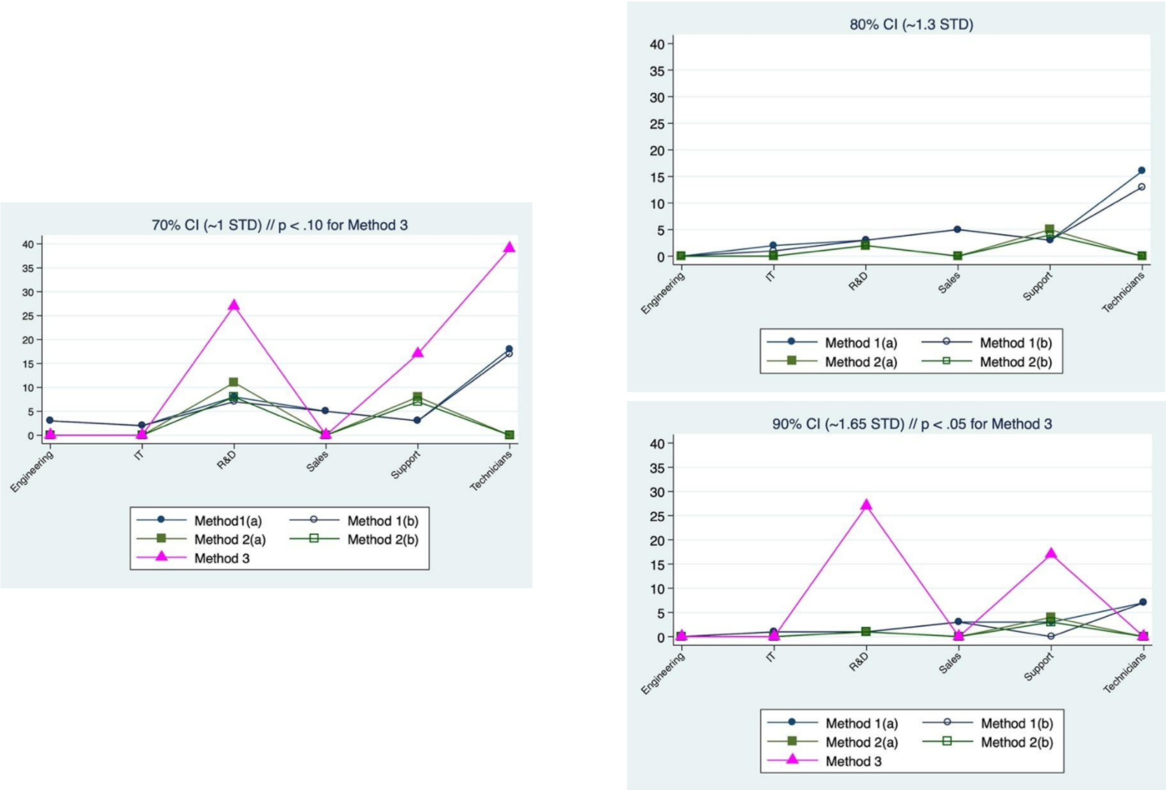
Second, as shown in Table VI, the average remedy per employee for Method 3 is about 36% less than what it is using the alternative methods. The overall average remedy for Method 3 is $7,904 per employee, as compared with $11,429, $12,654, $11,926, and $13,328 for Methods 1(a), 1(b), 2(a), 2(b) respectively.

Table VI: Average Remediation per Employee by Method (90% / p = .05)

SSG	Method 1(a)	Method 1(b)	Method 2(a)	Method 2(b)	Method 3
	1 overall model (M & F eligible for remedy)	*1 overall model (disadvantaged group only eligible for remedy)*	*Group Basis / Std Dev (M & F eligible for remedy)*	*Group Basis / Std Dev (disadvantaged group only eligible for remedy)*	*Group Basis / p-value (disadvantaged group only eligble for remedy*
Engineering					
IT	$ 14,152	$ 12,050			
R&D	$ 11,969	$ 10,442	$ 25,976	$ 25,976	$ 8,856
Sales	$ 19,290	$ 18,129			
Support	$ 629		$ 8,413	$ 9,112	$ 6,393
Technicians	$ 12,221	$ 10,711			
Overall	**$ 11,429**	**$ 12,654**	**$ 11,926**	**$ 13,328**	**$ 7,904**

Lastly, as illustrated in Figure II, the headcounts of employees to whom each Method recommends providing remedial compensation are starkly different. Method 1 and 2 are particularly spartan. At the 70% level, Methods 1(a) and 1(b) recommend providing a remedy to 19 and 15 people. Methods 2(a) and 2(b) recommend 39 and 37 respectively. Method 3 recommends providing a remedy to 83 people at the p < .10 level.The 90% / p < .05 level presents even more stark differences in the numbers of people to whom remedies are recommended. Methods 1(a) and 1(b) suggest only 5 and 4 people get anything, and Methods 2(a) and 2(b) are at only 15 and 12, respectively. Method 3 recommends remedies for 44 individuals. Note too, that Methods 2 and 3 only recommend remedies in 2 of the 6 SSG—R&D and Support. Method 1(a) recommends a remedy in all groups except Engineering, and Method 1(b) recommends providing a remedy in all groups except Engineering and Support.

Figure II: Employees for Whom Remediations are Recommended by Method



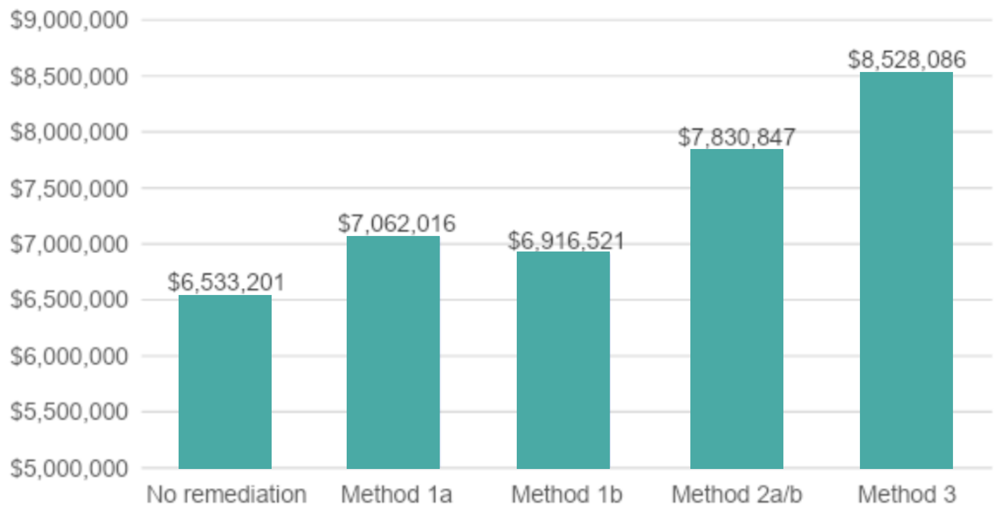
Measuring the Methods' Remediation Differences in Terms of Employee Earnings

To illustrate the likely effects of the differences that the Methods in terms of who is flagged to receive a remedy, and how much they are suggested to be given, this section offers a comparison of the cumulative of the Methods' compensation changes on employees' lifetime earnings. Three women from the example dataset demonstrate the variance in lifetime earnings.

For a woman receiving a pay equity adjustment to her compensation early in her career, remediation can have a substantial impact over time. Assuming a 3% cost of living increase every year for the remainder of a 45-year career and no other pay changes, an employee in the dataset in her fourth year of her career identified as needing remediation by all methods would see substantial changes in her earnings over time.

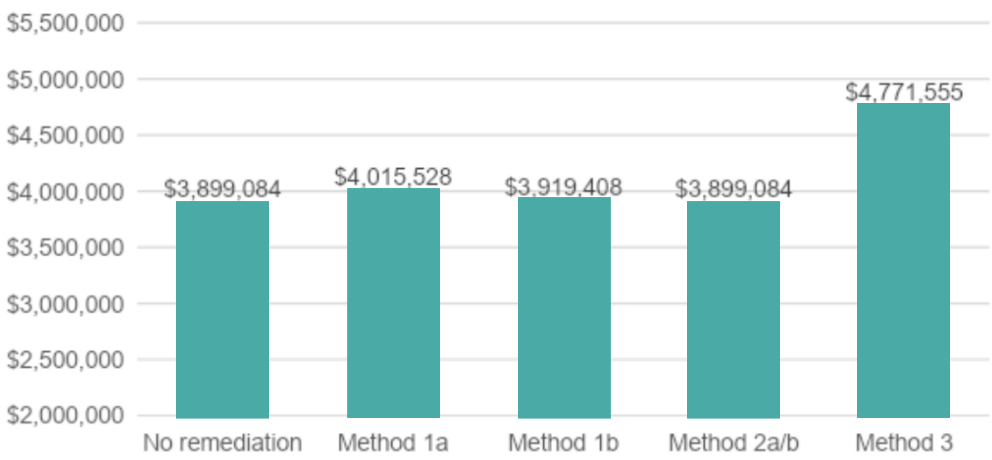
Changes in lifetime earnings range from an increase of $383,319 with Method 1(b) to nearly $2 million more over the course of her career (a 31% increase) with Method 3 than if there were no remediation, as shown in Figure III.

Figure III: Estimated Lifetime Earnings by Method for an Early Career Worker



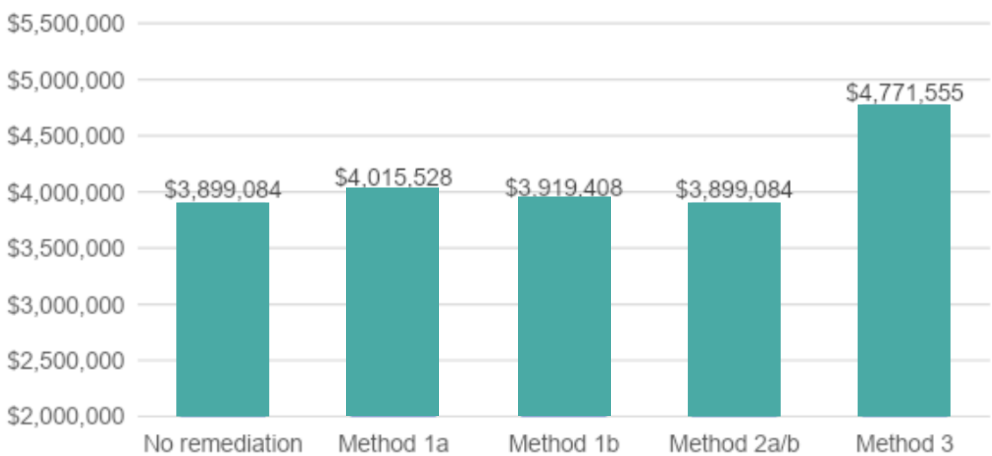
In some cases, not every Method identifies the need for remediation for an employee. For one employee in the data set, Methods 2(a) and 2(b) suggest no remediation. Method 1(a) suggests a remediation that results in a $116,444 increase in lifetime earnings and Method 1(b) results in $20,325 more, while Method 3 results in $872,472 more over the course of her career (Figure IV).

Figure IV: Estimated Lifetime Earnings by Method with No Remediation Identified by Methods 2 a/b



Similarly, sometimes Methods 2 and 3 identify the need for remediation, but Method 1 does not. In this case, remediation using Method 2(a) or 2(b) results in an additional $124,588 over the course of a career, while Method 3 increases lifetime earnings by $1,123,497, as shown in Figure V.

Figure V: Estimated Lifetime Earnings by Method with no Remediation Identified by Methods 1 a/b



In short, there are significant lifetime earnings implications for employees that come from the choices made by the employer. While this is on the one hand an exercise showing the impact of a point in time solution and it's echoing longitudinal repercussions, employers more interested in reducing legal risk should take note that statutory damage periods can be four to six years depending on the jurisdiction, so even employers with no altruistic motivation should regard these comparisons as presenting a sobering choice of methods when thinking of potential class-wide pay equity damages calculations.

This should be regarded in tandem with the next section, because applying a remedy that does not actually reduce the statistically significant gap to acceptable levels makes the damages calculations for Methods 1 and 2 even more concerning.

Does Each Method Succeed or Fail to Address the Statistically Concerning Pay Gaps?

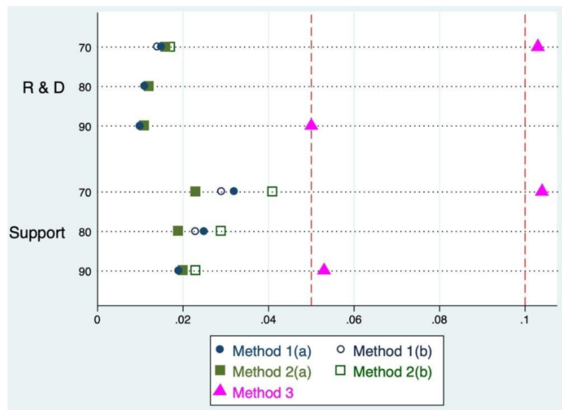
Recall from Table II, that when controlling for tenure, the gender coefficient remains statistically significant in the R&D group (p = .01), and the Support group (p =.02). After applying each Method's remediation recommendations to flagged identified employees, only Method 3 shows that both R&D and Support groups are no longer statistically significant at the applicable levels (p values less than .05).

In all other cases, after applying the recommended remediation amounts, the p-values for all both R&D and Support remain less than .05. This means that unless Method 3's remediation strategy is applied, for the two groups that showed statistical evidence that gender could not be ruled out as likely contributing to explaining the variation in pay, it remains the case that gender could not be ruled out as explaining pay. In other words, only Method 3 succeeds.

Figure VI depicts these results graphically for the Research & Development group and the Support Group. In both instances, it is clear that Method 3's results are the only ones that are above the applicable thresholds: p-value of .05 for the 95% level, and a p-value of .10 (the magenta triangle on the graphs). The graph contains vertical lines showing the applicable p-value cut offs of .05 and .10 for ease of reference.

Figure VI depicts these results graphically for the Research & Development group and the Support Group. In both instances, it is clear that Method 3's results are the only ones that are above the applicable thresholds: p-value of .05 for the 95% level, and a p-value of .10 (the magenta triangle on the graphs). The graph contains vertical lines showing the applicable p-value cut offs of .05 and .10 for ease of reference.

Figure III: Estimated Lifetime Earnings by Method for an Early Career Worker



Conclusion

This paper shows that the devil is in the details--not all pay equity review methodologies are created equally. Most importantly, only Method 3's remediation functioned as intended. For the two groups that started out showing statistical evidence of a pay imbalance in favor of men, Method 3's post-remediation tests showed no statistical significance at the appropriate threshold.

That is, Method 3 properly identified groups in which there is statistical evidence of a non-spurious gender disparity, and effectively minimally remediated those concerns. For Methods 1 and 2, even after the recommended remediation amounts were applied, those two groups continued to show statistical evidence of a pay disparity in favor of men. Methods 1(a) and 1(b) are certainly the easiest and most convenient to run for labor economists or statisticians, however, they appear to be the most precarious for employers. In fact, applying the 70% confidence interval using either Method 1(a) or 1(b) actually exacerbates the problem in one of the six groups (Engineering).

The lower remediation amounts suggested by Methods 1 and 2 may be attractive to employers seeking to get ahead of pay equity risk, but the unfortunate truth is that those lower numbers may be too good to be true. Applying them does not ameliorate the statistical evidence of concerns. While the appeal of cheaper, faster, or easier methods may be appealing, the old adage of "caviat emptor" applies with full force for employers considering this methodological approach. Additionally, the number of people receiving a proposed remedy should be taken into consideration. As of the writing of this paper, no pay equity law states that employers must compress distributions of pay.

No pay equity law states that it is unfair, illegal, immoral or unethical to pay people less than 1, 2, 3, or 4 standard deviations below the mean, so long as the employer does not pay women less than men because of their gender (or vice versa). So, using methods that compress distributions has little to no direct bearing on statutory compliance. It may indirectly or inadvertently help ameliorate problems but the brute force and lack of precision leading to over-corrections and under compliance should give employers serious concerns before implementing.

Citations

1. Zev J. Eigen is the Founder and Chief Data Scientist of Syndio. He is a former Professor (Northwestern; Yale; NYU), and former in-house counsel with Twentieth Century Fox Film Corporation. (PhD, MIT ’09; JD, Cornell Law School ’99; BS, Cornell University ’96)

2. Sometimes, those deploying Method 1 apply a Bonferroni correction (or equivalent), making matters even less like the model(s) used in the event of litigation.

3. The employee in this example has 4 years of work experience.

4. Except in Missouri and Montana where only women have the right to sue for pay equity violations under state laws.

Domain Expert Bios

Zev Eigen | Co-founder and Chief Data Scientist
Dr. Eigen is the co-founder and chief data scientist at Syndio and a nationally recognized expert on pay equity. Dr. Eigen earned his Ph.D. from MIT's Sloan School of Management. As a former professor, his research has been supported by grants from the National Science Foundation, Harvard Law School, MIT, and private sources such as Twentieth Century Fox Film Corporation. He holds a J.D. from Cornell Law School and a B.S. from Cornell University's School of Industrial and Labor Relations (with honors). He's been featured numerous times in business and financial press and has also acted as a testifying expert for pay equity.

Allison Adams | Data Scientist
Allison is a data scientist at Syndio who firmly believes in pizza rolls - not gender roles. She is skilled at developing advanced descriptive and inferential people analytics statistical models that empower organizations to make key business decisions. She holds her M.S in Information Systems Management with an emphasis in Data Analytics from Seattle Pacific University.

Andrea Palmiter | Advice and Analytics Manager
Andrea has over a decade of experience helping customers leverage data analytics to make organizational change. She has also worked as a management consultant with a focus on developing and implementing diversity and inclusion strategies for Fortune 500 tech companies and higher education institutions. Andrea holds a B.A. from Emory University and a master's from Harvard University in International Education Policy with a concentration in quantitative evaluation.