ORCAA

Comment on LL 144 Proposed Rules

Submitted to DCWP | October 24, 2022

ORCAA is an independent algorithmic auditing consultancy. We have conducted algorithmic audits of hiring tools, and have had dozens of discussions related to LL 144 with vendors of AEDTs, employers that use AEDTs, and law firms. This has given us an understanding of various stakeholders' perspectives.

Depending on the rules adopted, this law could provide substantial protection and transparency for candidates, or it could permit inconclusive audits that merely continue the status quo. In any case, it will set an important precedent for similar laws elsewhere. We believe the key principles are to focus on outcomes that matter to candidates, and to consider each AEDT *as it is actually used*, which can vary across employers.

In addition to our comments below, we attach a mock Bias Audit Report we developed (see Appendix A). We believe this is a viable template and would be glad to help DCWP adopt or adapt it.

Per employer bias audits would be more effective

The example in §5-301(a) of the proposed rules suggests a vendor can conduct a general Bias Audit of an AEDT using "historical data it has collected from employers". We argue this is not feasible in general, and even when it is, it is often insufficient to address the critical question of whether different candidates in certain protected classes have a fair chance of being hired when the AEDT is involved. Bias Audits done on a per employer level, which address an employer's particular use of an AEDT and focus on downstream outcomes rather than scores, would provide a clearer view of whether disparate impact is occurring in the use of the AEDT.

<u>Vendors may not have access to employer usage data</u>

Many vendors do not have access to the data necessary to conduct a bias audit as outlined in the proposed rules. Information about specific candidates – who they are and the hiring decisions made about them – generally belongs to the prospective employer. Some vendors may provide AEDTs that their customers run locally on their own systems. In these cases, the vendor may never have the chance to collect data on those candidates. In other cases the candidates run through the vendor's system, but the vendor's contracts with customers limit what candidate data they can access or store, or how they can use it. These data limitations imply many vendors could not produce bias audits as proposed.

Most vendors do not track downstream outcomes

Employment decisions, not scores, are what matters in terms of fair hiring. Even if vendors have access to some candidate data, it is likely insufficient to address what actually happened to candidates. Many AEDTs produce a score (e.g., a "culture fit" score, as in the second example table in the proposed rules). This is the AEDT output the vendor is most likely to be able to report on, and many or most bias audits based on the proposed rules would focus on score analyses.

In practice AEDT scores are often binned according to numeric thresholds chosen by the employer, and candidates are then presented with simpler bin labels (e.g. Low, Medium, or High) instead of the raw score. These labels are then one of multiple inputs to an employment decision (e.g., whether a candidate is selected for an interview).

Focusing on scores ignores what happens downstream. For one, it misses the critical issue of thresholds. Consider the second example table in the proposed rules, which shows a bias audit based on scores. If the Medium/High threshold were 80, then only three of the 14 groups (White Males, Multiracial Males, and Multiracial Females) would have High average scores. But if the threshold were 70, then half of the groups would have High average scores.

These scenarios would likely lead to substantially different hiring outcomes in terms of average selection rates by group. The first scenario is far more likely to produce a disparate impact than the second, but the proposed score analysis offers no way to capture this important distinction. Since different employers can set different thresholds, a score analysis that pooled data from multiple employers would shed even less light on the employment outcomes that really matter.

Aggregating across customers can mislead

Even if vendors had access to candidate data and downstream employment decisions, the kind of pooled analysis shown in the first example table could be misleading. It could allow discriminating employers to fly under the radar, or make blameless vendors or employers look bad.

Biases by separate employers using a single AEDT could "cancel out" in a combined analysis. We show in Appendix B how this could happen with the first example table from the proposed rules. The example table exhibits gender parity: 48.0% of Males are selected versus 47.0% of Females, for an Impact Ratio of 0.979. Now we imagine the example table was based on data from two firms, A and B, whose corresponding tables are shown separately. Firm A favors Male candidates (51.7% of Males selected versus 40.9% of Females; Impact Ratio = 0.792) and Firm B favors Female candidates (53.1% of Females selected versus 44.4% of Males, Impact Ratio = 0.837). But these significant disparities disappear in the combined analysis.

Similarly, consider two firms C and D using the same AEDT, where Firm C considers 10,000 candidates and Firm D considers just 100 candidates. Suppose Firm C strongly favors Male candidates and Firm D has perfect gender parity. The combined analysis will suggest that the AEDT favors men, simply because Firm C's data dominates. This result is unfair to Firm D and, arguably, to the AEDT vendor (after all, Firm D was able to use the AEDT without a gender disparity).

The proposed rules do not address applicant sourcing

The pool of applicants an employer reviews, as well as the method the employer uses to generate, or "source", the pool, can contribute to bias in the hiring process. For example, suppose an employer manages to (unfairly) target its recruiting so that it attracts an overwhelmingly White applicant pool. These applicants then pass through a resume-filtering AEDT, and are screened out at comparable rates across race groups. The result will be an overwhelmingly White set of hires. Given the proposed definition of a Bias Audit, this would not be seen as a problem.

Conversely, when the labor pool for a given job is unbalanced, Bias Audits that address sourcing would effectively incorporate this context. For instance, if 80% of security guards are male, then we should not be surprised if applicant pools for security guard jobs skew male.

Bias Audits could address the sourcing issue by including a demographic breakdown of the applicant pool considered by the AEDT, alongside a similar breakdown of a relevant comparison population. Our mock report (see Appendix A) includes an example. The Bureau of Labor and Statistics would be a natural source for comparison data at a national level, and more local statistics might be available.

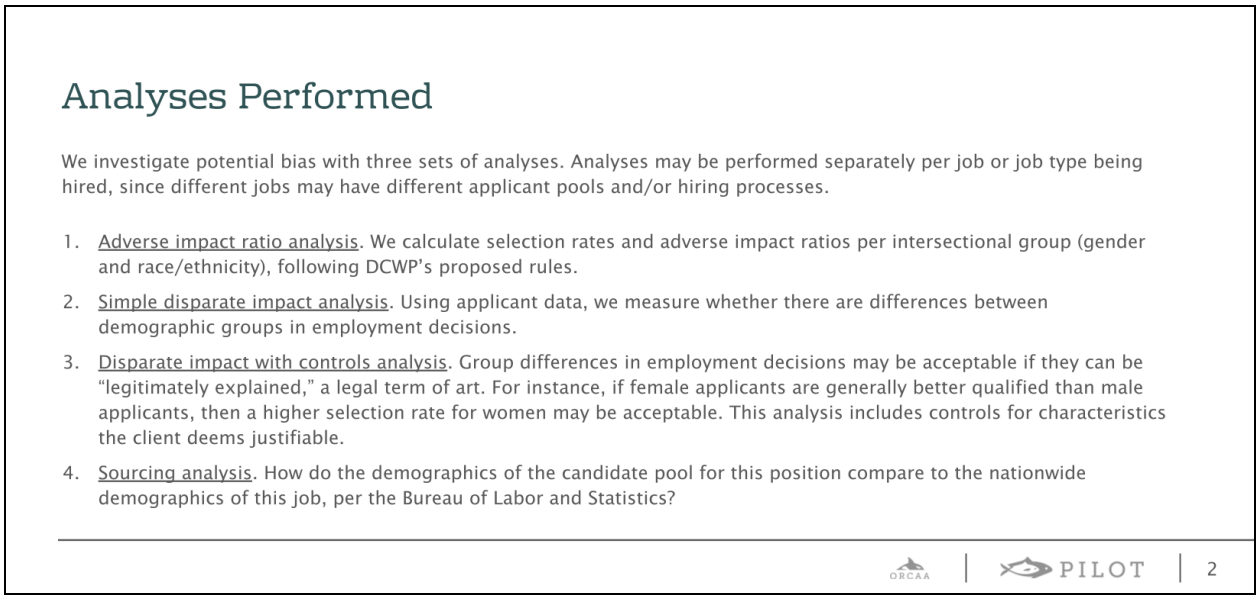
The proposed rules do not address sample size issues

Some samples are too small to do robust statistical analysis. The example tables in the proposed rules show intersectional analyses (each unique combination of gender and race/ethnicity is a separate group); this leads to 14 groups. The first example table shows 24 Native American / Alaska Native males, and 17 Native American / Alaska Native females. These samples are simply too small to trust the estimates of selection rates for these groups, or the corresponding Impact Ratios.[1] One alternative would be to allow (or prescribe) coarser analyses when there is little data. For instance, below a certain size threshold it might make sense to analyze gender and race/ethnicity separately rather than intersectionally, or to condense race/ethnicity categories. Offering guidance on how to deal with small samples would mitigate the risk of Bias Audits containing unreliable statistics.

[1] Consider the statistics "rule of thumb" based on the central limit theorem, that at least 30 data points should be used to calculate a population average.

Appendix A: Mock Bias Audit Report

Note: The below version is current as of October 20, 2022. Future updates will be reflected in the live version of the mock report, hosted on our website.



Audit Scope

Client: NewCo

Automated Employment Decision Tool: ToolX is used by NewCo to help determine which candidates are invited for an interview. After submitting an application, a candidate is given the ToolX survey. Their responses are assessed by ToolX, which gives a recommendation: "Interview" or "Screen". This recommendation is shown to a NewCo hiring manager, who can follow it or not, at their discretion. Hiring managers generally, but not always, follow the recommendations.

Employment decision: Whether the candidate was selected for an interview (Y/N)

Variables that may "legitimately explain" group differences: Years of experience

Protected classes addressed: Race/ethnicity. This is inferred for each candidate using the BIFSG method.*

Scope of data:

Job title	# of candidates	Starting from	Until
Retail salesperson	698	1/1/2021	12/31/2021
Engineer	460	5/1/2021	12/31/2021

Notes:

- Prior to 5/1/2021 NewCo hired Engineers using a traditional hiring process, without ToolX

* For documentation on inference methods, see [Pilot url]

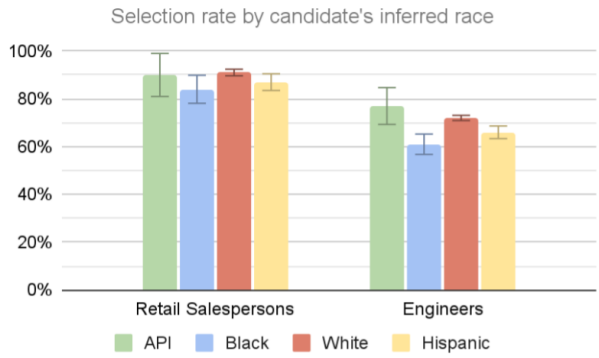
1. Adverse Impact Ratio Analysis

		# of applicants	# selected for interview	Selection rate	Adverse impact ratio
Male	Hispanic	83	69	83.1%	0.881
	White	223	202	90.6%	0.960
	Black	140	104	74.3%	0.787
	API	66	59	89.4%	0.947
	AI/AN	14	10	71.4%	0.757
	Two or more	51	43	84.3%	0.893
Female	Hispanic	95	84	88.4%	0.937
	White	212	192	90.6%	0.960
	Black	130	100	76.9%	0.815
	API	71	67	94.4%	1.000
	AI/AN	9	6	66.7%	0.706
	Two or more	64	48	75.0%	0.795

This table calculates selection rates and adverse impact ratios for each intersectional group (gender * race/ethnicity), following the sample bias audit table in DCWP's proposed rules for the law.

We note that the proposed rules do not include thresholds of acceptability for selection rates or the adverse impact ratios, or for the number of candidates that constitute a sufficient sample for statistical analysis.

2. Simple Disparate Impact Analysis



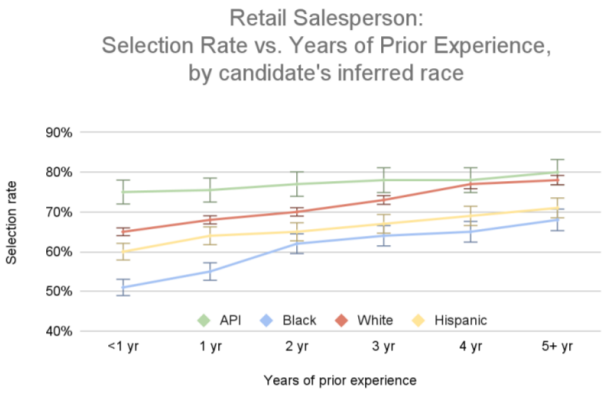
This bar chart shows the selection rates of candidates by their inferred race. For each position NewCo hired, there is a separate set of bars. The "whiskers" at the top of each bar show a 95% confidence interval of the true selection rate for that group.

In this case "selection" means being invited for an in-person interview, since the automated employment decision tool being audited is a pre-interview survey.

For retail salespersons, White and API candidates have higher selection rates than Black or Hispanic candidates, but the overlapping confidence intervals suggest the differences are not significant.

For engineers, Black and Hispanic candidates have lower selection rates and the confidence intervals do not overlap, suggesting the difference is statistically significant.

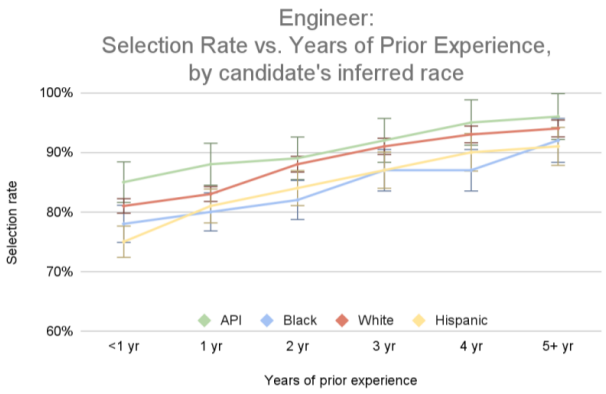
3. Disparate Impact with Controls: Retail Salesperson



This graph shows the selection rate of retail salesperson candidates according to their years of prior experience. Each inferred race group of candidates is shown as a separate line. The "whiskers" around each point of each line are 95% confidence intervals.

For each inferred race group, selection rate increases with years of prior experience, which makes sense. However there are persistent differences: at every level of prior experience, API candidates are most likely to be selected, followed by White, then Hispanic, and finally Black candidates. Moreover, the confidence intervals for Hispanic and Black candidates are strictly below those of API and White candidates, so the differences here are likely to be statistically significant.

3. Disparate Impact with Controls: Engineer

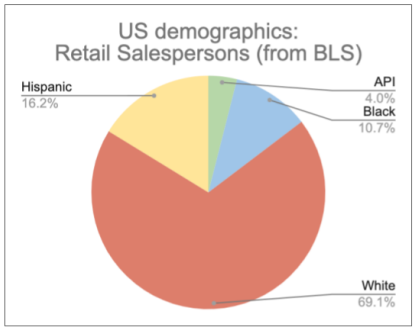


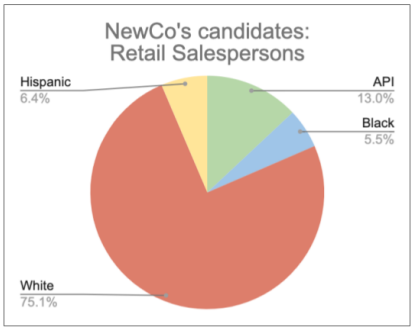
This graph shows the selection rate of engineer candidates according to their years of prior experience. Each inferred race group of candidates is shown as a separate line. The "whiskers" around each point of each line are 95% confidence intervals.

For each inferred race group, selection rate increases with years of prior experience, which makes sense. However there are persistent differences: for every level of prior experience, API candidates are most likely to be selected, followed by White candidates, with Black and Hispanic candidates less likely. There is considerable overlap between the confidence intervals (no line is entirely "on its own" above or below the others), so the differences are unlikely to be highly statistically significant.

ORCAA | PILOT | 7

4. Sourcing Analysis: Retail salesperson





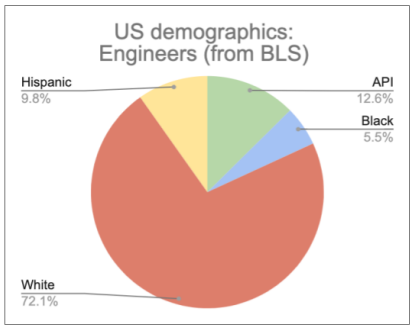
The pie chart on the left shows the demographic breakdown of all US retail salespersons, per BLS' Labor force characteristics by race and ethnicity, 2020 (most recent available; "Retail salespersons" category).

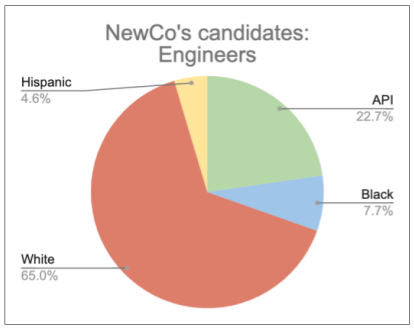
The pie chart on the right shows the demographic breakdown of NewCo's candidates for retail salesperson jobs, modeled using inferred race.

Compared to all US retail salespersons, the main differences are that NewCo's candidate pool is more white (75% vs 69%) and API (13% vs 4%), and less Hispanic (6% vs 16%) and Black (6% vs 11%).

ORCAA | PILOT | 8

4. Sourcing Analysis: Engineers





The pie chart on the left shows the demographic breakdown of all US engineers, per BLS' Labor force characteristics by race and ethnicity, 2020 (most recent available; "Architecture and engineering occupations" category).

The pie chart on the right shows the demographic breakdown of NewCo's candidates for engineer jobs, modeled using inferred race.

Compared to all US engineers, the main differences are that NewCo's candidate pool is more API (23% vs 13%), and less White (65% vs 72%).

Appendix B: Combining employer data could hide disparities

Combining the applicant data from (hypothetical) Firm A's and Firm B's tables below would produce the example Bias Audit table shown in §5-301(a).

Example Bias Audit table shown in §5-301(a)

			# of applicants	# selected	Selection Rate	Impact Ratio		Male selection rate
Hispanic or Latino		Male	205	90	43.9%	0.841		48.0%
		Female	190	82	43.2%	0.827		
Non/Hispanic or Latino	Male	White	412	215	52.2%	1.000		Female selection rate
		Black or African American	226	95	42.0%	0.806		47.0%
		Native Hawaiian or Pacific Islander	87	37	42.5%	0.815		
		Asian	321	167	52.0%	0.997		Impact ratio
		Native American or Alaska Native	24	11	45.8%	0.878		0.979
		Two or More Races	115	52	45.2%	0.866		
	Female	White	385	197	51.2%	0.981		
		Black or African American	164	75	45.7%	0.876		
		Native Hawaiian or Pacific Islander	32	15	46.9%	0.898		
		Asian	295	135	45.8%	0.877		
		Native American or Alaska Native	17	7	41.2%	0.789		
		Two or More Races	98	44	44.9%	0.860		

Firm A's table

FIRM A			# of applicants	# selected	Selection Rate	Impact Ratio		Male selection rate
Hispanic or Latino		Male	100	50	50.0%	0.909		51.7%
		Female	95	42	44.2%	0.804		
Non/Hispanic or Latino	Male	White	200	110	55.0%	1.000		Female selection rate
		Black or African American	113	55	48.7%	0.885		40.9%

		Native Hawaiian or Pacific Islander	44	21	47.7%	0.868		
		Asian	160	85	53.1%	0.966		Impact ratio
		Native American or Alaska Native	12	6	50.0%	0.909		0.792
		Two or More Races	58	28	48.3%	0.878		
	Female	White	191	80	41.9%	0.762		
		Black or African American	81	32	39.5%	0.718		
		Native Hawaiian or Pacific Islander	15	6	40.0%	0.727		
		Asian	150	59	39.3%	0.715		
		Native American or Alaska Native	9	4	44.4%	0.808		
		Two or More Races	50	19	38.0%	0.691		

Firm B's table

FIRM B			# of applicants	# selected	Selection Rate	Impact Ratio		Male selection rate
Hispanic or Latino		Male	105	40	38.1%	0.632		44.4%
		Female	95	40	42.1%	0.698		
Non/Hispanic or Latino	Male	White	212	105	49.5%	0.821		Female selection rate
		Black or African American	113	40	35.4%	0.587		53.1%
		Native Hawaiian or Pacific Islander	43	16	37.2%	0.617		
		Asian	161	82	50.9%	0.845		Impact ratio
		Native American or Alaska Native	12	5	41.7%	0.691		0.837
		Two or More Races	57	24	42.1%	0.698		
	Female	White	194	117	60.3%	1.000		
		Black or African American	83	43	51.8%	0.859		
		Native Hawaiian or Pacific Islander	17	9	52.9%	0.878		
		Asian	145	76	52.4%	0.869		
		Native American or Alaska Native	8	3	37.5%	0.622		

		Two or More Races	48	25	52.1%	0.864		