Bias audit: summary of results

AI Shortlist, Spotted Zebra

Independent bias audit conducted by VerifyWise under New York City Local Law 144 of 2021 and the implementing rules of the Department of Consumer and Worker Protection.

Date of bias audit	25 September 2026
Valid until	25 September 2027
Tool audited	Spotted Zebra AI Shortlist
Independent auditor	VerifyWise

What the tool does and what was measured

AI Shortlist scores a candidate's CV against five role criteria, classifying the evidence for each as full, potential or no evidence. The overall score is the proportion of the five criteria classified as full evidence.

One or more criteria may operate as a knockout. As Spotted Zebra describes it, a candidate without full or potential evidence in this criterion receives 0% overall, with the remaining classifications still returned. In the configuration audited, the knockout was set to require full evidence: every candidate without full evidence for that criterion scored 0% overall.

This audit reports a **scoring rate**, being the proportion of candidates in each category who scored **above** the median for the sample, as provided for in section 5-301(c) of the rules.

Median score for the full sample: 60.0%

Number of candidates assessed: 5,000.

Source and explanation of the data

This audit used **test data**, as permitted by section 5-302(b) of the rules where insufficient historical data is available to conduct a statistically significant bias audit.

Why historical data was not used. Demographic capture has been enabled for one customer of AI Shortlist, through which approximately 1,500 to 2,000 candidates have recently been processed. The only characteristic captured for those candidates is sex. Race and ethnicity were not collected. A small number of candidates across other customers may have both recorded, in very small numbers.

Local Law 144 requires impact ratios for sex, for race and ethnicity, and for the intersections of the two, which is 14 categories across the seven race and ethnicity groups. Where race and ethnicity is not collected, neither those categories nor any intersectional category can be calculated from historical data, whatever the number of candidates processed. The audit therefore used test data, as the rules permit where historical data is insufficient for a statistically significant bias audit. Should race and ethnicity capture begin at reportable volumes, future audits of this tool are expected to use historical data.

How the test data was generated and obtained. VerifyWise generated 5,000 synthetic candidate CVs for a single Account Executive role. The role definition was written independently by VerifyWise and supplied to Spotted Zebra, who configured the five scoring criteria from it using their standard process. Each synthetic candidate was assigned a sex, race or ethnicity, age band and disability status drawn from a balanced allocation, together with a qualification profile ranging from weak to strong. CVs were generated from those profiles by a large language model against a fixed configuration and validated before delivery.

The corpus was delivered in two halves, scored together in a single run on a single configuration. 2,493 CVs carry no demographic signal of any kind: no name, no email, no membership or activity that would indicate a protected characteristic. The remaining 2,507 carry the markers a real CV would carry, including a name, an email address and club, community and activity memberships, matched to each candidate's assigned demographics. The two halves allow the tool to be measured both on documents carrying no demographic signal and on documents resembling real applications.

Demographic attributes were held by VerifyWise and never disclosed to Spotted Zebra, who scored the CVs without knowing which candidate belonged to which group. Scores were returned keyed by candidate identifier and joined back to the demographic data by VerifyWise.

Number of individuals assessed who fall within an unknown category: 0. Every synthetic candidate carries an assigned sex and race or ethnicity by construction.

Number of individuals assessed but excluded from the calculations: 0. All 5,000 candidates were scored and all 5,000 are included.

Results

Sex categories

Category	Candidates	Scored above median	Scoring rate	Impact ratio
Male	2,500	997	39.88%	1.000
Female	2,500	995	39.80%	0.998

Race and ethnicity categories

Category	Candidates	Scored above median	Scoring rate	Impact ratio
American Indian or Alaska Native	715	288	40.28%	1.000
White	714	287	40.20%	0.998
Hispanic or Latino	714	286	40.06%	0.994
Asian	715	286	40.00%	0.993
Black or African American	714	284	39.78%	0.987
Two or More Races	714	281	39.36%	0.977
Native Hawaiian or Other Pacific Islander	714	280	39.22%	0.974

Intersectional categories of sex and race or ethnicity

Category	Candidates	Scored above median	Scoring rate	Impact ratio
Male, White	357	146	40.90%	1.000
Female, American Indian or Alaska Native	358	145	40.50%	0.990
Female, Hispanic or Latino	357	144	40.34%	0.986
Male, Asian	358	144	40.22%	0.984
Male, American Indian or Alaska Native	357	143	40.06%	0.979
Male, Black or African American	357	143	40.06%	0.979
Female, Asian	357	142	39.78%	0.973
Male, Hispanic or Latino	357	142	39.78%	0.973
Female, Black or African American	357	141	39.50%	0.966
Female, Native Hawaiian or Other Pacific Islander	357	141	39.50%	0.966
Female, Two or More Races	357	141	39.50%	0.966
Female, White	357	141	39.50%	0.966
Male, Two or More Races	357	140	39.22%	0.959
Male, Native Hawaiian or Other Pacific Islander	357	139	38.94%	0.952

No category was excluded under the 2% threshold in section 5-301(d). The smallest intersectional category represents 7.14% of the scored sample.

Age band

Not required by Local Law 144, which names sex and race or ethnicity. Included because the audit population carries it.

Category	Candidates	Scored above median	Scoring rate	Impact ratio
Under 40	2,500	997	39.88%	1.000
40 or older	2,500	995	39.80%	0.998

Disability status

Category	Candidates	Scored above median	Scoring rate	Impact ratio
No disability	2,500	998	39.92%	1.000
Disability	2,500	994	39.76%	0.996

Interpretation

No impact ratio falls below 0.80. The lowest is 0.952, for Male candidates recorded as Native Hawaiian or Other Pacific Islander in the intersectional analysis.

A ratio of 0.80 is the conventional threshold, drawn from the federal four-fifths rule, at which a difference in outcomes is treated as warranting further examination. Local Law 144 does not set a permitted floor and does not itself reference that rule. It requires that these figures be calculated and published, which is what this summary does.

Three points a reader should weigh alongside the figures.

The candidate population was balanced by design. Each protected category contains approximately equal numbers of candidates and an equal spread of qualification profiles. Where a scoring rate is measured against the sample median, a balanced population tends to produce impact ratios near 1.00. These results are therefore evidence that the tool's scoring does not respond to the demographic signal carried in the CVs. They are not a measurement of how the tool performs on the live applicant population of any employer deploying it, whose composition differs.

The tool discriminates strongly on qualification. Synthetic candidates were built to five qualification levels and mean scores across those levels were 0.0%, 0.1%, 59.6%, 79.5% and 99.9%. A tool with no discriminating power would also produce impact ratios near 1.00, so this gradient establishes that the tool was operating as intended when it returned no demographic difference.

A second analysis compared the two halves of the corpus. Within each qualification level, candidates whose CVs carried demographic markers scored 0.1 percentage points differently from candidates whose CVs carried

none, a difference indistinguishable from chance ($p = 0.20$). This measures outcomes only. It does not establish whether the markers reached the scoring stage, and it is not a measurement of Spotted Zebra's redaction step.

The full audit report contains the method, the further analyses conducted and the limitations of this audit.

Revision, 28 September 2026. The description of the knockout now carries Spotted Zebra's own account of it, attributed to them, and states separately the configuration that was audited, in which the knockout required full evidence.

No figure, scoring rate or impact ratio changed. The audit, its date and its results are as issued on 25 September 2026.