Testing for ethnic bias in HackerEarth

While skill-based assessments help in evaluating candidates accurately, it's also important to check whether these assessments are free from any bias and have a disparate impact. Hence, HackerEarth joined hands with the High Beam Global team, a global market research agency that specializes in data collection, reporting, and analytics solutions, catering to organizations across the globe.

This exercise was performed to determine if there was any bias in HackerEarth assessments against people from different ethnicities, races, and genders. The sample size was 210, with 30 people each from 7 different ethnicities. Participants were screened on the basis of the following using an online, pre-screening questionnaire:

Age:

21-45 years old

Education:

At least an undergraduate degree or diploma in Computer Science/Applications or Engineering/Mathematics

Job role and description:

Employed in the areas of IT Support Services/Core IT services, Software Development and/or Testing, Research and Development and Analytics.

Participants also needed to have knowledge of software programming/coding in computer languages especially Python.

Python was chosen because it's one of the most commonly used programming languages and is used in a wide variety of industries for different computational purposes.

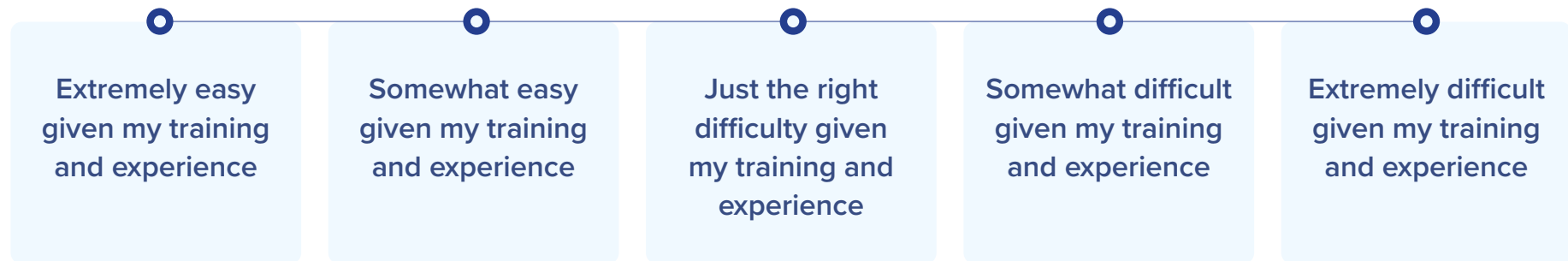The assessment results were analyzed based on the following parameters:

- Score distribution
- Attempted time distribution
- Perception parameters such as perceived relevance*, perceived bias*, and *perceived difficulty*.

Perceived relevance

Measures the relevance of the questions in the assessment with a participant's regular day-to-day work as a Python programmer. The responses were recorded on a 7-point scale, anchored at the end-points as 1=Not at all relevant' and 7=Extremely relevant.

Perceived difficulty

Measures a participant's evaluation of the difficulty level of the assessment based on their training and experience. The responses were recorded on a 5-point, Likert-type scale.

| Extremely easy given my training and experience | Somewhat easy given my training and experience | Just the right difficulty given my training and experience | Somewhat difficult given my training and experience | Extremely difficult given my training and experience |

Perceived bias

Measures a participant's evaluation of whether the structure and content of the assessment displays any bias for or against their particular ethnic/race group. The responses were recorded on a 5-point, Likert-type scale.

| Strongly biased against my race/group | Somewhat biased against my race/group | No ethnic/race bias | Somewhat biased in favour of my race/group | Extremely biased in favour of my race/group |

These were analyzed using the analysis of variance (ANOVA) test. This test is commonly used in the field of statistical inference to analyze the difference between the mean values obtained from different groups. It is used to determine if the differences observed for a certain metric among different groups (that are selected based on certain parameters) are unlikely to be due to a random chance.

Here are the findings from the exercise:

1.

There was no significant difference between the ethnic groups with respect to the time taken to complete the assessment or the number of questions attempted.

2.

There were no statistically significant differences between the ethnic groups on perceived relevance and perceived difficulty of assessment and perceived ethnic bias.

3.

There is no statistical evidence of any gender, age, or educational attainment bias (graduates and undergraduates only) in the three direct assessment metrics: total score, time spent on the assessment, and the no. of questions attempted.

4.

There was evidence of difference in perceptions, while assessing the three perceptual metrics. These were as follows:

- Females had a significantly lower score on perceived difficulty and perceived significantly lower bias against their ethnic group.

- The younger age group showed significantly higher perceptions of perceived difficulty statistically, although the difference was less than 0.5 scale points.

- Graduates had a statistically significantly higher perceived relevance of the test score than undergraduates.

- Graduates had a significantly lower score on bias against their own ethnic group. The difference was less than 0.5 scale points and not likely to be practically significant.

5.

There was evidence of ethnic differences when the total score of the assessment was considered. East Asians scored significantly higher than non-Hispanic White or Jewish and Other Single Race. This does not mean that the assessment methodology is necessarily biased in favor of East Asians, it simply means that East Asians are able to perform better on such assessments.