Harnessing the Wisdom of Crowds to Improve Hiring

December 30, 2016

At least as far back as the 19th century, statisticians found that groups of people are capable of making a more accurate decision than any single individual. Yet organizations rarely take advantage of this “wisdom of the crowd” to improve operations.

In a new program, behavioral scientists at the Behavioural Insights Team (BIT) in the United Kingdom are harnessing the decision-making power of groups to improve the way that organizations conduct hiring. Originally commissioned under Prime Minister David Cameron in 2010, the BIT is a leading example of government testing of public and organizational policy interventions through evidence-backed collaborations with behavioral scientists. Across a series of experiments, the BIT has been investigating how findings from behavioral science can be used to help organizations, including the BIT itself, improve hiring practices.

“Organisations spend eye-watering sums trying to attract the best talent because in many industries, the difference between the best and the good has real implications for the bottom line,” BIT behavioral scientists Kate Glazebrook, Theo Fellgett, and Janna Ter Meer write in a BIT blog post.

The team’s first set of experiments on innovating hiring practices was inspired by a nearly 200-year-old...
statistics experiment. The English statistician Francis Galton famously asked 800 townsfolk at a county fair in Plymouth, England, to guess the exact weight of an ox down to the pound. Guessers wrote their estimates on bits of paper, which Galton then analyzed. Galton had hypothesized that “oxen experts” such as farmers and butchers would have the best estimates, but the crowd proved this assumption wrong: The group’s estimate was more accurate than the individual guesses from experts.

Galton was shocked by the results: The median of all 800 guesses was extremely close to the exact weight of the ox. The group estimate was 1,197 pounds, and the actual weight of the ox was 1,198 pounds — a difference of just 0.08%.

This “wisdom of the crowd” demonstrated that, under the right conditions, groups of people can make more insightful decisions than individuals can, sometimes even besting the experts. Glazebrook and colleagues suspected that integrating more people into the hiring process could similarly improve the chances of picking out the best possible candidates from the pool of applications.

Many hiring decisions come down to superficial criteria, such as choosing to interview only graduates from certain universities or unconsciously favoring candidates based on traits like gender, ethnicity, or sexual orientation. By focusing on these perfunctory traits, organizations may miss out on highly qualified candidates. Additionally, these kinds of homogenous hiring practices can lead to situations in which employers miss out on the benefits that come from a diversity of perspectives. When everyone addresses problems in the same way (i.e., “groupthink”), teams can end up missing major concerns altogether.

Although companies are increasingly aware of the benefits of a diverse workforce, actually translating these goals into hiring practices has been a challenge. The BIT wanted to find out whether they could build better, more diverse teams by adopting a hiring strategy that could take advantage of the wisdom of crowds.

“One area of crucial importance to almost all organisations is recruitment, but research shows that a whole host of implicit biases result in suboptimal hiring decisions,” the BIT explains in their 2016 report. “Studies have shown that organisations are more likely to offer job interviews to candidates with ‘white-sounding’ names. Recruiters make snap judgements about individuals in interviews, and structure recruitment processes (e.g. sending a cover letter and CV) in ways that give too much weight to factors (gender, race, social class) that should be irrelevant to an individual’s ability to do a role.”

Research from APS Fellow Philip E. Tetlock (University of Pennsylvania) has demonstrated that people are better at forecasting outcomes when they work together in collaborative teams. Tetlock and colleagues have spent years studying decision-making and expertise. One of their key findings has been that pooling multiple perspectives can counter the cognitive biases that lead to bad decisions. The BIT drew on Tetlock’s research to help inform their own approach to bias in hiring.

“In fact, researchers have even shown that US defense intelligence analysts with access to classified information can be beaten by some rudimentarily-educated amateurs: largely because they come to conclusions too quickly and struggle to update their opinions in the face of new and conflicting information,” Glazebrook, Fellgett, and Ter Meer explain in their BIT blog post.
Research also suggests that people with varied backgrounds and experiences will tackle problems differently, and this diversity of perspectives can help organizations make better decisions.

A team of psychological scientists led by APS Fellow Adam D. Galinsky (Columbia University) recently summarized empirical arguments for more diverse teams in *Perspectives in Psychological Science*: “Homogeneous groups run the risk of narrow mindedness and groupthink (i.e., premature consensus) through misplaced comfort and overconfidence. Diverse groups, in contrast, are often more innovative and make better decisions, in both cooperative and competitive contexts.”

So when it comes to reviewing resumes and interviewing applicants, how big does the crowd need to be to maximize the benefits?

The BIT designed a simple online experiment in which approximately 400 reviewers rated four hypothetical job candidates based on responses to a generic recruiting prompt (i.e., “Tell me about a time when you used your initiative to resolve a difficult situation.”). The reviewers were given a set of guidelines, similar to those given to conduct a structured interview, to help them assess the quality of responses.

The 400-person crowd had a clear favorite and easily identified the best candidate response.

“We took our data and ran statistical simulations to estimate the probability that different groups could correctly select the best candidate,” Glazebrook and colleagues explain. “We created 1,000 combinations of reviewers in teams of different sizes, ranging from one to seven people. We then pooled them by the size of the group and averaged their chance of selecting the right candidate.”

When there was a gap in quality between the best and second-best responses, an individual picked the less qualified person approximately 16% of the time. However, with a group of three decision-makers, the odds of choosing the lesser candidate dropped to 6%, and with a five-person group, the chance decreased to 1%. When the two candidates were very similar, individuals selected the best candidate approximately 50% of the time — basically, they had the same accuracy as tossing a coin. A crowd of seven, on the other hand, picked the superior candidate more than 70% of the time.

Of course, polling 400 reviewers for every job isn’t very practical. Ultimately, the evidence suggested that three reviewers was the optimal crowd size for recruitment, but more experiments are still in the works.

**Turning the Science Inward**

“The Behavioural Insights Team likes to live by its own principles. When we examined the literature on how organizations can improve their internal practices, we realized we had to apply them to BIT as well,” the team explains.

To this end, the BIT has developed a platform called Applied. The goal of this project is to use findings from behavioral science to reduce the role of bias in the hiring process.
Most job searches start with an applicant submitting their resume or CV along with a cover letter. Someone in human resources then sorts through the pool of applicants, narrowing it down to a set of individuals who will be invited for an interview. But the small experiment described above simply doesn’t support the standard CV sift as a particularly useful hiring tool. For example, a candidate with a degree from a prestigious private university on his or her CV may be chosen over someone equally qualified who attended a state university, or a candidate with a typically masculine name may be assumed to have greater leadership potential compared with a female job candidate.

“With respect to CVs in particular, research argues that CVs typically contain information that is largely irrelevant to a candidate’s performance on the job. Nevertheless, this information has the potential to prey on the unconscious biases of the assessor,” the BIT explains in their 2016 report.

The Applied platform attempts to increase quality and diversity in hiring through implementing four key features: anonymization, chunking, collective intelligence, and predictive assessment.

First, the platform anonymizes applications by scrubbing irrelevant information such as names (which can provide cues about an applicant’s gender, race, age, or ethnic background). The applications then are organized by “chunking” — instead of reading through one full application at a time, reviewers compare a specific question from an application with the same question from other applications. This helps reviewers to identify the overall best responses.

Last, three or more people review the remaining pool of applicants. Agreement of multiple reviewers helps ensure that the best possible candidate is ultimately chosen. Job assessments and situational work tests are chosen based on whether there’s evidence showing that specific tests are “genuinely predictive of performance on the job.”

The Applied platform isn’t just for private organizations and businesses: The BIT has used the platform to improve their own hiring practices.

**Can You Take the Bias Out of Hiring?**

In an experiment to determine whether Applied was doing what it was supposed to, BIT researchers tested the platform against a more traditional “CV sift” during their own 2015–2016 graduate recruitment period.

First, the team designed a parallel A/B test of the 160 candidates who had the best performance on an initial multiple-choice test. The application materials for all candidates were sent through both the automated Applied review and the normal “sift” from a senior HR manager who reviewed CVs and resumes. The resulting pool of successful candidates was then sent through a rigorous set of skill assessments and final in-person interviews.

Ultimately, this process gave applicants two shots to get hired: They could make it through the traditional review process based on an exceptional CV, or they could be chosen based on the scores from the evidence-based hiring tests used by Applied.
“When we pulled all of the data in, lots of things surprised us,” Glazebrook and Ter Meer write in a post on Medium.

There was no correlation between the score for an applicant’s CV and in-person performance in later rounds. Simply having an impressive CV with recognizable schools and fancy titles was a weak predictor for test scores during the other assessments. There was, however, a significant, positive relationship between the Applied scores and the two in-person interview rounds — that is, people with high Applied scores on their application materials also performed well in person.

But did Applied actually come through on delivering a more diverse set of hires? While there wasn’t a significant difference between the two hiring groups on gender, there was evidence that Applied was less biased against people with a disability and people from non-White backgrounds, although the sample sizes were too small to provide a statistically significant conclusion for these measures.

There was a marked difference in the role of educational background between the two groups. While the CV sift favored applicants based on formal educational attainment, those who made it through the Applied sift had a much more diverse educational background — that is, the people who had the most years of higher education didn’t necessarily have the best skills for the job. This finding is in line with trends from companies like Google and IBM, where formal college education or university grades increasingly are viewed as irrelevant predictors of someone’s performance on the job.

“We would never have hired (or even met!) a whopping 60 percent of the candidates we offered jobs to if we’d relied on their CVs alone,” Glazebrook and Ter Meer write.

Of course, more evidence is required to demonstrate that this hiring approach will actually translate into on-the-job performance. The cohort of hires at Applied is too small to use as a meaningful test of the platform’s capabilities in the real world, but the Applied team is looking for opportunities to run a larger test.

On a national scale, the use of this kind of bias-limiting approach could have an enormous impact on helping individuals get the jobs for which they’re qualified. As Glazebrook and Ter Meer explain, “even if 1 in 5 candidates were given jobs that they otherwise wouldn’t have, across the economy, that’s hundreds of thousands of people getting jobs they otherwise wouldn’t have based on merit.”

References


