Harvard Business Review

REPRINT H02TF2
PUBLISHED ON HBR.ORG
APRIL 18, 2016

ARTICLE
GENDER

How to Take the Bias Out of Interviews

by Iris Bohnet

Harvard Business Review

GENDER

How to Take the Bias Out of Interviews

by Iris Bohnet

APRIL 18, 2016

HBR STAFF

If you're a hiring manager, you're probably happiest getting a sense of a candidate through unstructured interviews, which allow you to randomly explore details you think are interesting and relevant. (What does the applicant think of her past employer? Does she like Chicago? What does she do in her downtime?) After all, isn't your job to get to know the candidate? But while unstructured interviews consistently receive the highest ratings for perceived effectiveness from hiring managers, dozens of studies have found them to be among the *worst* predictors of actual on-the-job performance — far less reliable than general mental ability tests, aptitude tests, or personality tests.

Why do we stick with a method that so clearly does not work, when decision aids, including tests, structured interviews, and a combination of mechanical predictors, substantially reduce error in predicting employee performance? The organizational psychologist Scott Highhouse called this resistance "the greatest failure of I-O [industrial and organizational] psychology."

The unwillingness to give up a much-loved evaluation approach seems to be driven by two factors: Managers are overconfident about their own expertise and experience, and they dislike deferring to more structured approaches that might outsource human judgment to a machine.

When sociologist Lauren Rivera interviewed bankers, lawyers, and consultants, they reported that they commonly looked for someone like themselves in interviews. Replicating ourselves in hiring contributes to the prevalent gender segregation of jobs, with, for example, male bankers hiring more male bankers and female teachers hiring more female teachers.

Sometimes we have to learn the hard way. A few years ago, Texas legislators realized that the state was short on physicians. To fix the problem, the legislature required the University of Texas Medical School at Houston to increase the class size of entering students from 150 to 200 — *after* the admissions committee had already chosen its preferred 150 students. As most students apply to several medical schools simultaneously, by that time all the top-ranked candidates had already been spoken for. This meant that the pool of still-available students was made up of candidates who had previously received a low ranking from the admissions committee. Of the 50 students finally selected, only seven had received an offer from another medical school.

This government-dictated requirement turned into an eye-opening field study, allowing University of Texas researchers to examine whether the initial ranking mattered for the students' performance in and after medical school. Here's the shocking result: The performance of initially accepted and initially rejected students turned out to be the same. Digging deeper, the researchers found that about three-quarters of the difference in ratings between initially accepted and initially rejected students was due not to more-objective measures, such as grades, but rather to the interviewers' *perceptions* of the candidates in unstructured interviews. Faced with this finding, the researchers ask whether, after an initial assessment of academic and other performance measures, traditional interviews should be replaced by a lottery among the viable applicants.

I don't expect lotteries to replace interviews in most situations, but the evidence against unstructured interviews should make any hiring manager pause. These interviews should not be your evaluation tool of choice; they are fraught with bias and irrelevant information. Instead, managers should invest in tools that have been shown to predict future performance. On the top of your list should be work-sample tests related to the tasks the job candidate will have to perform. For example, Compose, a cloud storage company, decided to completely do away with resumes and instead evaluate job candidates based on how well they solved a job-related problem. Companies including Applied, Blendoor, Edge, GapJumpers, Interviewing.io, Paradigm, and Unitive provide the analytical tools and the software that can bring more structure to hiring procedures.

Still, most companies will want to "see" the candidates after the work-sample test. Companies should rely on a structured interview that standardizes the process among candidates, eliminating much subjectivity. These interviews pose the same set of questions in the same order to all candidates, allowing clearer comparisons between them. This may seem like an obvious approach, but, incredibly, it remains underused. Of course, the flow of conversation during the interview will be slightly more awkward than it already is, but the payoff is worth it.

The protocol should also require the interviewer to score each answer immediately after it is provided. This neutralizes a variety of biases: We are more likely to remember answers with vivid examples, for example, and answers that are most recent. Evaluators who wait until the end of the interview to rate answers risk forgetting an early or less-vivid but high-quality answer, or favoring candidates whose speaking style favors storytelling.

It's also best to compare candidate responses horizontally. That is, if you interview five candidates, compare each of their answers on question one, then each answer on question two, and so on. Many academics, myself included, do this when grading exams. Ideally, evaluators hide their assessment of question one from themselves — literally obscuring it from view — to reduce the chance that the answer will influence scores on subsequent questions. This can be uncomfortable because interviewers often discover that a candidate gives superb answers to some questions but deeply disappointing ones to others. Although it complicates evaluation, identifying this internal inconsistency is worthwhile, especially if there are questions that receive more weight.

Comparative evaluations not only help us calibrate across candidates but also decrease the reflex to rely on stereotypes to guide our impressions. In joint research with Max Bazerman of Harvard Business School and Alexandra van Geen of Erasmus University Rotterdam, we have shown that biases that lead us to expect women to be better at stereotypically female jobs, such as nursing, and men at stereotypically male jobs, such as engineering, tend to kick in when we focus on and vertically evaluate one candidate at a time. In contrast, people are less likely to rely on whether a candidate "looks the part" when evaluating several candidates simultaneously and comparing them systematically.

Structured interviews are not just about discipline in asking questions — some companies, including Google, structure the content of their interviews using data. Their people-analytics departments crunch data to find out which interview questions are more highly correlated with on-the-job success. A candidate's superb answer on such questions can give the evaluator a clue about their future performance, so it makes sense that responses to those questions receive additional weight.

Replacing unstructured with structured interviews is only part of the battle; managers should also abandon panel, or group, interviews altogether. I know of no evidence that they provide a superior gauge of a candidate's future performance. And it's best to keep interviewers as independent of each other as possible. To state the obvious, if you have four interviewers, four data points from four individual interviews trump one data point from one collective interview.

Once everyone has evaluated all candidates, the evaluators should submit their assessments before a meeting to discuss an applicant. This would allow an organization to aggregate answers — those one-to-10 weighted scores on questions asked in the exact same order. Assessments with candidates above a certain threshold should advance for further consideration. Whether the candidate roots for the Red Sox should not even be asked, and of course it's not relevant.

Implementing these adjustments is just the beginning. Having embraced the fact that unstructured approaches are inferior, make sure you keep experimenting to fine-tune what works best in your context. For example, my colleagues and I are working with the government of a large country to improve its talent management. We have run a number of A/B tests to measure the diagnostic power of specific work-sample tests and interview questions for public servants. Data analysis and design innovations such as comparative evaluations can help organizations level the playing field and benefit from 100% of the talent pool.

Like people in most other human endeavors, hiring managers are powerfully and often unwittingly influenced by their biases. While it's exceedingly difficult to remove bias from an individual, it's possible to design organizations in ways that make it harder for biased minds to skew judgment. We should stop wasting resources trying to de-bias mindsets and instead start to de-bias our hiring procedures. Work-sample tests, structured interviews, and comparative evaluation are the smart and the right things to do, allowing us to hire the best talent instead of those who look the part. Smarter design of our hiring practices and procedures may not free our minds from our shortcomings, but it can make our biases powerless, breaking the link between biased beliefs and discriminatory — and often simply just stupid — actions.

Iris Bohnet is a behavioral economist and professor at the Harvard Kennedy School and director of its Women and Public Policy Program. Her new book, "What Works: Gender Equality by Design," was published in March 2016 by Harvard University Press.