

How to Improve the Accuracy and Reduce the Cost of Personnel Selection

Don A. Moore¹

California Management Review

2017, Vol. 60(1) 8–17

© The Regents of the

University of California 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0008125617725288

journals.sagepub.com/home/cm



SUMMARY

One of the clearest lessons to emerge from decades of research on personnel selection is that the traditional face-to-face job interview is terrible for predicting future job performance. The sad truth is that no selection tool is perfect, and future performance cannot be predicted precisely, but the traditional interview is particularly bad. Fortunately, it is easy to improve the predictive validity of the job interview by structuring it around hard-to-fake tests of key skills and abilities that actually predict future performance. There are also other tools as accurate as a structured interview and substantially less expensive to administer.

KEYWORDS: human resources management, organizational behavior, recruitment

Like most people, the MBA students in my class believe that they are good judges of character. Almost all have participated as interviewers and recruiters at their jobs, even before business school. Their companies, like almost all corporations, rely on face-to-face unstructured job interviews to select their employees. So when my students arrive in class at the beginning of their first semester, they believe they can predict work performance from an interview. It is my job to disabuse them of this belief.

I used to offer my students an abundance of evidence from the many hundreds of studies of job interviews, showing that the correlation between interview performance and subsequent job performance was—to say the least—disappointing. Authoritative reviews of the research literature¹ put the correlation between face-to-face job interviews, as they are usually conducted, and subsequent job performance, at around .38. To be sure, that is better than nothing, but it leaves much to be desired. What are some other things with correlations about that size? Weight and height are correlated, across the entire population, at about .4. It is more than zero, but there are plenty of tall skinny people and short heavy people.

¹University of California, Berkeley, Berkeley, CA, USA

Knowing how much someone weighs should not leave you very confident about your ability to guess their height.

After having read the research literature and learned of the poor correlation between interview performance and subsequent job performance, it is not exactly that my students disbelieved the evidence. Many of them were ready to accept the conclusion that the interviewers in these studies could not predict job performance. But for most of them, their faith in their own insight remained unshaken. “Sure, there are plenty of bad interviewers out there,” they told me. “But I know how to ask the right questions and read the responses. I know how to interview.”

I realized I needed to be able to show my students evidence from their own predictions. That is when I started having them interview each other. The assignment: to accurately forecast classmates’ performance on the midterm exam. I gave my students plenty of information about how the exam would work, including copies of all the old midterm exams. I also provided interviewers flexibility to structure their interviews in any way they wanted. The interviews could be long or short. They could ask their interviewees any questions they liked.

I tried to align the incentives to mimic those present in real interviews. Interviewers in my class were rewarded for making more accurate predictions of future performance. The reward for accurate prediction was better grades in the class. I also provided incentives for performance among those being interviewed. Those being interviewed scored more points when their interviewers predicted they would perform well on the exam. The more optimistic the interviewer’s prediction, the better the grade for the one being interviewed.

After the midterm exam, I correlate predictions with actual performance. It is worth noting that this is a luxury most organizations lack. In real organizations, those who perform poorly on interviews simply do not get the chance to perform. They do not get hired and so never get a chance to disprove the interviewers’ grim forecasts for their potential. In order to assess the merits of interviews—or any selection method—it is best to measure subsequent performance in the entire set of candidates, not only those who have performed the best. When it comes to selecting new hires, this would mean hiring everyone, or at least ignoring the interview results when making hiring decisions. Few organizations have the courage to do so, and so few can adequately assess the quality of their selection methods.

It is worth noting another unrealistic feature of my classroom exercise: students know each other. They see each other at school and at social events. This familiarity may provide useful information that would rarely be present in the case of a job applicant. For instance, those taking full advantage of California’s legalization of recreational marijuana usage during their time in Berkeley might not perform at the top of the class. This extra knowledge should serve to increase the correlation between interviewers’ predictions and subsequent exam performance among my students.

How well are my students able to predict how their classmates will perform? Predictions are correlated about .2 with actual performance. This number

fluctuates from year to year, with some years as low as .15 and other years as high as .27. A .2 correlation does not explain 20% of the variance. To compute the variance accounted for, you have to square the correlation coefficient. So a correlation of .2 means that my students' predictions, based on interviews, are useful for predicting about 4% of the variance in midterm grades.

Objections

Some of my students objected that the predictions they were making were harder than the predictions required in job interviews. Midterm exam performance is a clearer and more objective performance standard than most organizations use in personnel assessment. This objection is well founded. It is true that most organizations assess performance with a more complex, vague, and subjective assessment: supervisor ratings. Whether one gets the promotion or the raise depends on assessment by one's supervisors. And there is good reason to expect supervisors' impressions to be influenced by some of the same factors as are interviews. Traits and behaviors that increase perceived liking and competence in an interview contribute to perceptions of likability and competence beyond the interview.

I responded to my students by noting that perceptions of competence and actual competence are not the same thing. Workers engage in some behavior intended to favorably influence perceptions by supervisors but only weakly related to actual work performance. This may include putting in more face time at the office, flattering the boss, or drawing attention to one's successes through strategic self-promotion. On the other hand, workers do some things that contribute to work performance but may go unnoticed by supervisors, such as helping out a colleague, doing extra work away from the office, or humbly attributing credit to others.

If the same sorts of selfish self-promotion contribute both to performance in interviews and to supervisor ratings, then the two should correlate with each other thanks to what researchers call "mono-method bias." That is, the two measures correlate because they share an underlying methodology: subjective impressions. Indeed, supervisor ratings were the measure of performance used by the majority of the research studies that made the .38 estimate of the correlation between interviews and work performance. We ought to expect lower correlations with actual job performance, assuming we could quantify it objectively. Most organizations have trouble quantifying all the aspects of an individual's performance, which is precisely why they rely on the imperfect proxy of supervisor ratings.

But in my class, I can obtain objective measures of midterm exam performance. Given that this is what my students were trying to predict with their interviews, I think there is reason to think that the .2 correlation I observe in my classes is closer to the true correlation between interviews and subsequent work performance than is the .38 that emerges from studies that rely on supervisor ratings. And of course, most organizations are seeking to maximize actual work performance and not simply maximize successful manipulation of supervisors' impressions.

Some of my students objected because the way I measured their predictions of performance was with a percentile rank. Specifically, the question was, “What percentage of people in the class would perform worse than the person you interviewed?” My students insisted that predicting percentile ranks was harder than predicting absolute performance. They legitimately objected that, since they had not interviewed all the members of the class, it was difficult for them to predict how everyone would perform relative to each other.

Note that asking the interviewers to predict percentage correct on the mid-term exam would not solve this problem. In fact, it might make the prediction more difficult since percentage correct will depend fundamentally on how difficult the exam is. And my students cannot know how hard the test will be before they see it or I have written it. By contrast, the relative abilities of students compared with one other is knowable before seeing the test. So, I started asking them to interview not one but *two* classmates. If interviews allowed the interviewer to compare people, then they should correctly predict which of the two would perform better. If they could do so perfectly, then interviewers should predict the correct order (for the two people they interviewed) 100% of the time. On the other hand, random guessing should get the order right 50% of the time. How did my students do? On average, they got the order right about 55% of the time. They were better than complete chance, but not much.

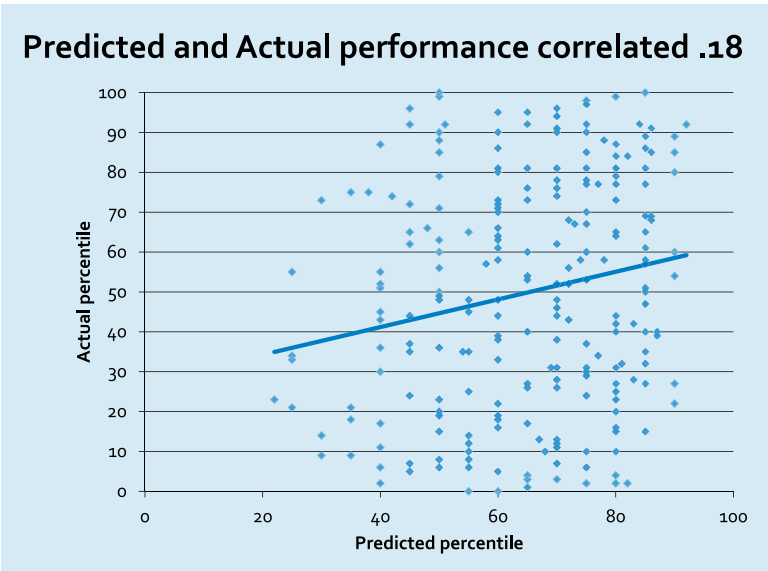
Deception and Manipulation

Another artificial feature of my class is that, in the interests of fair play, I prohibit interviewees from engaging in deception. When candidates tell the truth, it ought to make the interviews more informative and increase their ability to predict future performance. Nevertheless, interviewees had an incentive to manipulate. They had good reasons to want to pull the wool over their interviewers’ eyes. They may have succeeded. The average prediction was that the interviewee’s midterm exam score would beat 66% of their classmates. That is, of course, a statistical impossibility. The average percentile rank, across the entire class, must be 50. The upward bias in my students’ predictions may show that interviewees succeeded in manipulating the perceptions of their interviewers.

Figure 1 shows data from my class. The first striking feature is the poor correlation between predictions and outcomes. Knowing how someone was rated by their interviewer does not allow you to predict their exam performance with much confidence. The second striking feature is how few dots there are on the left side of the plot—no one predicted that their classmates would perform in the bottom 20% of the class. Obviously, this is wrong. However, anyone who analyzed personnel evaluations will recognize this pattern: people rarely use the low end of the scale to evaluate others.

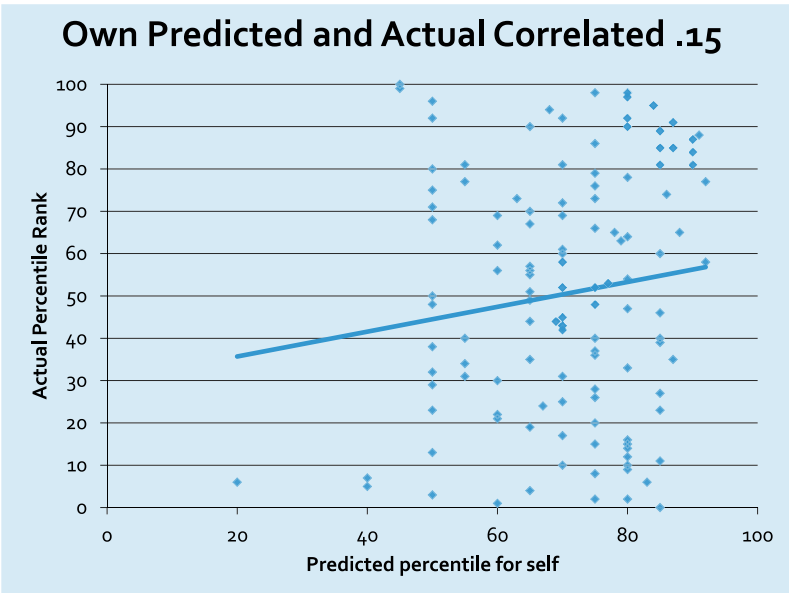
Could my students have asked better interview questions? Maybe. But even if interviewers had been able to inject a truth serum that would have guaranteed honest answers to every question, their predictions are unlikely to have gotten much better. Why not? My students do not know how they will perform

FIGURE 1. Scattergram plotting percentile ranks for actual test performance (y-axis) against predictions by interviewers (x-axis).



Note: The line represents the best-fit linear regression line, and illustrates a correlation of .18.

FIGURE 2. Scattergram plotting percentile ranks for actual test performance (y-axis) against predictions by the test-takers themselves (x-axis).



Note: The line represents the best-fit linear regression line, and illustrates a correlation of .15.

on the midterm. As Figure 2 shows, forecasts of their own exam performances are as bad as interviewers' forecasts. The correlation between forecasts and actual scores is about .15, and students rarely forecast that their performances will fall in the bottom third.

Why would people have trouble forecasting their exam performance? One problem is obviously bias: everyone wants to believe they will do well. Numerous studies document “better-than-average” beliefs in which people exaggerate the degree to which they are better than others. Perhaps the most famous example of this is Svenson’s finding that 93% of American drivers believe they are better than the median driver.² As for my students, 83% predict that their exam score will exceed the median.

Another problem that impairs forecast accuracy is unfamiliarity: they do not know exactly what situations and problems they will face. My students note this issue by way of complaint: many of them have been out of school for years, and my midterm is to be the first exam they have taken in a long time. How could they know how they would perform relative to the rest of the class? My response is to plead guilty to the accusation, and note that the same concern applies to most job-selection decisions. Unless the candidate has worked in the exact same job at the same company with the same people, the job is likely to put them in a novel situation facing novel challenges. And no one, not even they themselves, can predict with certainty how they will respond.

The belief that we can predict how people will behave in novel situations rests, in part, on faith in the power of dispositions. Knowing whether someone is honest should help predict their honesty in a new job, right? But decades of personality research suggest we should have low expectations regarding the predictive power of dispositions. The best measures of personality correlate with relevant behaviors only about .2 or .3. For instance, if you measure how conscientious someone is and then measure how clean they keep their room or whether they submit assignments punctually, you are unlikely to get correlations above .3.

Not only do we routinely overestimate the predictive value of personality, but we also tend to infer too much about someone’s personality based on a small sample like a job interview. Indeed, the tendency to over-attribute to personality is so powerful and robust that psychologists have dubbed it the *fundamental* attribution error.³ All of us are tempted to infer, after seeing an interview in which a candidate is prepared, organized, and articulate, that these behaviors are due to the candidate’s personality. We tend to neglect the powerful situational influences that can affect behavior in an interview but might not be present in the workplace. When we observe someone behave clumsily, we are tempted to conclude that they are clumsy and ignore the situational pressures, stresses, and distractions that could contribute to clumsy behavior under the stressful glare of the interviewer but might not predict the same clumsiness in everyday work.

Noisy Measures

Part of the problem with my students’ predictions is that they are biased. Just about no one predicts that their interviewees will score in the bottom third of the class. On the one hand, this seems odd, given that a third of the class will score in the bottom third. On the other hand, this is consistent with the

workplace; supervisors are loath to rate their people as below average. This reluctance renders the lower third of the rating scale as useless; supervisors simply do not use it. One could just cut off the lower third of the scale, but then supervisors would not use the bottom of the remaining scale either. Frustration with the inflation in supervisors' ratings has led some organizations to implement forced ranking systems. Yet these systems are so widely despised that they rarely last long. No one wants to be ranked worst. Most famously, both GE and Microsoft abandoned their stacked ranking systems. Both supervisors and workers hated the system that required 10% of people to be rated in the bottom 10%.

But upward bias cannot by itself account for the low correlation between forecasts and performances. Inflation in forecasts would have restricted their variation, but it need not have suppressed the correlation. Instead, noise and inconsistency in forecasts represents the bigger problem. Like the interviews my students conduct, typical interviews are unstructured. The interviewer has freedom to ask different questions of every candidate, following up on interesting responses and customizing the interview for the candidate's abilities, weaknesses, and quirks.

If every interviewer asks different questions of every candidate, the result is inconsistency across interviews that reduces their comparability. If some questions allow candidates to shine more and some questions are more likely to reveal weaknesses, then asking different questions of different candidates will reduce the reliability of an interview as a true measure of a candidate's abilities. Structured interviews have the same interviewers ask all candidates the same set of questions. Research shows that structuring interviews this way is the single most important thing organizations can do to increase their predictive validity.

Improving Predictions of Work Performance

How much better are structured interviews than unstructured interviews? Schmidt and Hunter's definitive review of the research literature gives us reason to expect that they could improve the correlation between interviews and work performance to .52. That is impressive, given that .52 is nearly as good as more costly tools like job tryouts. Why do more organizations not structure their interviews? The answer is apparent to anyone who has taken part in a structured interview: they are more boring for all parties involved. The questions are driven not by what the interviewer wants to ask about or where the conversation naturally flows. Instead, they are designed to zero in on precisely those qualities that the organization believes are most predictive of job performance. The interviewer loses the ability to follow up on interesting responses and instead must provide a numerical assessment of the candidate's responses to each question.

Are there better predictors of job performance than structured interviews? Yes, but they are substantially more costly. The best predictor of future work performance is, unsurprisingly, performance at that same job. So job tryouts or probationary periods are better than interviews at predicting on-the-job performance. How

accurate can predictions about future work performance get, once you know how someone has performed at that job in the past? Job performance from one time period to the next correlates with itself about as well as performance on my mid-term exam and performance on my final exam: .54. That correlation provides a sobering splash of cold water on the dream of perfectly predicting job performance. The hard truth is that job performance is, and will remain, profoundly unpredictable. A correlation of .54, when squared, indicates that it accounts for less than a third of the variance. That leaves the majority of work performance unaccounted for. People learn and change. They get injured and they innovate. Their level of dedication or level of distraction varies as they fall in love, have children, or start taking Ritalin. Even if the person remains exactly the same from one time period to the next, the precise set of tasks, people, issues, and challenges at work can change. And that will always make it difficult to predict their future performance.

Fortunately, the research literature identifies other assessment tools that are less expensive to administer, less biased, and more accurate than interviews. One of these is tests of general mental ability. In other words: intelligence. It turns out that intelligent people perform better at their jobs, whether they are designing microchips or mopping floors. And there are intelligence tests that do not discriminate on the basis of language, culture, or ethnicity. These tests, such as Raven's Progressive Matrices, are simple to administer and considerably less costly in time, effort, and organizational resources than are face-to-face interviews. General mental ability, as measured by scores on the standardized GMAT test, is the one thing that does consistently predict my students' midterm exam performances.

Many people worry—justifiably—that intelligence tests fail to take into account many things that they value in a work colleague, including kindness, conscientiousness, and wisdom. That is true, but the evidence suggests that these things are also hard to diagnose based on an interview. Moreover, intelligence tests may be just as useful for something that they omit: intelligence tests are not affected by the biases of current employees, conscious or unconscious, regarding gender, ethnicity, or age. Most organizations would prefer to avoid discrimination on these categories. However, the evidence suggests that it is difficult for individuals to completely set aside their cultural attitudes when assessing others' potential cultural fit in the organization. All of us tend to engage in pattern matching, hiring more people that look like those who have previously been successful, with the result that hiring decisions perpetuate the current demographic makeup of the organization.

Perhaps the biggest single lesson from the research on personnel selection decisions is that performance is difficult to predict.⁴ No method of selection is perfect. There will always be false positives—people you hire whose performance is disappointing. And there will always be false negatives—people you reject who would have been great—although this latter group will probably be invisible to you. Selection is important, and getting a little better at it is often a worthwhile investment. But perfect performance prediction is a fantasy.

So how can your company use the research to improve the process by which it selects people? There are three straightforward lessons, and forward-thinking companies have successfully implemented many of them. First, if you

are going to continue using interviews, structure them. Determine which skills and abilities are actually necessary to succeed on the job, and then identify hard-to-fake ways of assessing those skills and abilities. Daniel Kahneman, a psychologist who later won the Nobel Prize in Economics for his foundational work, successfully introduced structured interviews in the Israeli military when he was doing his military service as a young man. In so doing, he helped improve their predictive validity of the army's selection of people into leadership positions. Likewise, the People Innovation group at Google uses a structured set of work-relevant questions in their interviews.⁵ It actively discourages interviewers from using their own intuitions to guide their interviews. Although many interviewers enjoy asking "mini-case" questions—such as, "How many rats are there in New York City?" as Laszlo Bock writes about in his book *Work Rules!*—these mini cases turn out to not be very useful for predicting subsequent work performance. They allow the interviewer to seem smart, but that's about it.

Each interview question should have a point. It ought to be able to assess some work-relevant ability or behavior. As such, it should be easy for interviewers to specify beforehand what a good answer might look like. Actual responses should be scored against these criteria. These scores should then be averaged to establish some numerical assessment of interview performance. If you are concerned that the scored questions seem to omit an intuitive holistic assessment of the candidate, then you can consider adding a final holistic assessment, and this can be averaged in with the scores for the specific interview questions.

Second, identify the key qualifications or prior experience that you think sets someone up for success at the job. Attach scores to each of the qualifications and include them in the hiring decision. The method Kahneman introduced in the Israeli military had interviewers score the candidate's resume, and these scores were then combined with the ratings of interview performance. References are also potentially useful. If you can get an honest assessment out of former colleagues, bosses, or subordinates, get it. These are helpful because they reflect a much longer time span than a brief interview. The selection process, during which a candidate is on his or her best behavior, is only a brief snippet of time. More revealing information comes out over months or years of work.

Third, consider combining interview performance with other measures, such as tests of intelligence. Standardized test scores are also useful tests of mental aptitude. It is also worth noting that you need not spend too much time worrying about how to weight these different components (resume qualifications, references, interview, and test scores). It is unlikely you will be able to compute the optimal regression equation that will tell you how to weight each one. Studies show that just averaging them is a pretty good approach that will be nearly as good as the optimal model in its ability to predict work performance.⁶

And finally, don't bother holding long meetings where you discuss the merits of all the candidates and argue about how to interpret this response or that test score. Too often, those meetings are dominated by the opinions of the person who happens to talk the most. Consequently, they introduce more arbitrary influences into the hiring process. Instead, the best way to capitalize on the wisdom of

the crowd is to simply average the scores for each candidate.⁷ The candidate with the highest average score is the one whom your evaluation system says is the best bet. Hire that one.

These recommendations will likely both improve the predictive validity of your company's selection process and also reduce the time and effort you put into recruiting. A hiring process that is both cheaper and more effective at the same time? That is a contender worth selecting.

Author Biography

Don A. Moore is the Lorraine Tyson Mitchell Chair in Leadership and Communication at the Haas School of Business at UC Berkeley (email: dmoore@haas.berkeley.edu).

Notes

1. F. L. Schmidt and J. E. Hunter, "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings," *Psychological Bulletin*, 124/2 (September 1998): 262-274.
2. O. Svenson, "Are We Less Risky and More Skillful than Our Fellow Drivers?" *Acta Psychologica*, 47/2 (February 1981): 143-151.
3. L. Ross, "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process," in *Advances in Experimental Social Psychology, Volume 10*, ed. L. Berkowitz (New York, NY: Academic Press, 1977), pp. 173-220.
4. S. Highhouse, "Stubborn Reliance on Intuition and Subjectivity in Employee Selection," *Industrial and Organizational Psychology*, 1/3 (September 2008): 333-342.
5. L. Bock, *Work Rules!* (New York, NY: Hachette, 2015).
6. R. M. Dawes, "The Robust Beauty of Improper Linear Models in Decision Making," *American Psychologist*, 4/7 (July 1979): 571-582.
7. W. Thorngate, R. M. Dawes, and M. Foddy, *Judging Merit* (New York, NY: Psychology Press, 2008).

Copyright of California Management Review is the property of California Management Review and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.