



KEYSTONE RECRUITING

How we evaluate

What the selection research says predicts how someone will perform, how we read it from what people have actually done, and where judgment comes back in.

1 The problem

Most hiring decisions rest on three things that feel reliable and are not.

Interviews feel conclusive and predict little.

An hour of conversation leaves a strong impression: fluent answers, shared references, a confident manner. The impression is real; its link to how the person later performs is weak. In the largest recent re-analysis of the selection research, the unstructured interview, the conversational kind most companies run, comes out at .19 (Sackett, Zhang, Berry & Lievens, 2022). In one experiment, interviewers formed impressions just as confidently when the candidate's answers were random as when they were real, and the interview made their predictions worse (Dana, Dawes & Peterson, 2013).

How to read the numbers. Each figure in this document is a validity: the correlation between what a method measures at hiring and how people later perform in the job. 0 means the method tells you nothing; 1 would mean it predicts perfectly. Most methods in this research fall between .10 and .45.

Resumes are written to be read, not to be true.

A resume is a sales document. It lists titles, employers and years, which are easy to compare and weakly related to performance: years of job experience comes out at .07 in the 2022 re-analysis, and at .18 in the older summary by Schmidt & Hunter (1998). What a person actually built, and how well, is rarely on the page in a form anyone can check.

The pond is the problem.

Even a perfect judge choosing from a short list can only pick the best of that list. A recruiter's network, a job board and the first hundred results of a search tool are a small slice of the market, and they over-represent the people who are actively looking. The best person for a role is usually not in that slice.

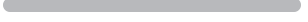
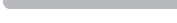
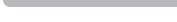
And instinct is hard to give up.

Practitioners keep trusting their own read of a candidate over decision aids that outperform it, a pattern Highhouse (2008) calls a stubborn reliance on intuition. In a survey that drew 959 responses from HR professionals, the gaps between what practitioners believed and what the research shows were especially large in selection, with far less faith in intelligence and personality tests than the research would justify (Rynes, Colbert & Brown, 2002).

2 What predicts performance

Decades of meta-analyses agree on the shape of the answer, even as the exact numbers have been revised.

Schmidt and Hunter (1998) summarised 85 years of research in one table that shaped the field for two decades. In 2022, Sackett and colleagues showed that a common statistical correction had inflated many of those estimates; their re-analysis lowered most of them and changed the order. Structured interviews emerged at the top. Both columns are below. The order matters more than the decimals.

METHOD USED AT HIRING	SACKETT ET AL. 2022	SCHMIDT & HUNTER 1998
Structured interview	 .42	.51
Job knowledge test	 .40	.48
Empirically keyed biodata	 .38	.35
Work sample test	 .33	.54
Cognitive ability test	 .31	.51
Integrity test	 .31	.41
Interests	 .24	.10
Conscientiousness	 .19	.31
Unstructured interview	 .19	.38
Years of job experience	 .07	.18

Validities for overall job performance as reported in each paper. Biodata here means a scored record of past behaviour and experience. Roth, Bobko & McFarland (2005) re-estimated work sample tests at .33 from newer studies, the same value as the 2022 column. Of the personality traits, conscientiousness is the one that showed consistent relations with performance in every occupational group studied by Barrick & Mount (1991).

What people overweight

- **Years of experience.** Near the bottom of both columns. It is the easiest thing to read from a profile and among the weakest things to rank on.
- **Brand names.** The school and the employers on a resume are widely used as shortcuts. Neither appears among the predictors with strong evidence.
- **Polish in conversation.** Appearance and self-presentation sway interview ratings, most of all in unstructured interviews: “what you see may not be what you get” on the job (Barrick, Shaffer & DeGrassi, 2009). At .19, the unstructured interview predicts less than half as well as the structured interview at .42.

3 How we read them

Our scorecards start from this evidence. The question is how to read it from a profile, before anyone has met the person.

Work actually done.

We score what a person has done, not how they describe themselves: the systems they have shipped, the code where there is code, the writing where there is writing. Records of behaviour are what the strongest predictors measure. Adjectives are not.

Evidence with every score.

Each score is written with the passage it rests on, quoted, so anyone can check it.

Confidence levels.

A thin profile cannot support a certain score. Every score carries a confidence that falls when the material is thin.

“Unknown” is an allowed answer.

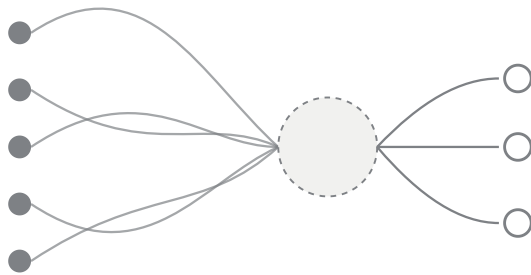
Where the profile is silent, the answer is “unknown”, not a guess. Unknowns become questions for the call.

The same rules for everyone.

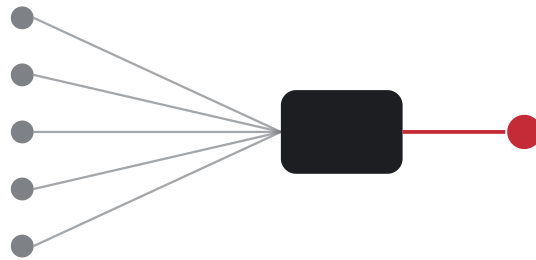
The rules that disqualify someone are written down and applied first, to every profile, the same way.

4 Why mechanical combination beats judgment, and where judgment comes back in

When the same information is combined by a fixed rule rather than by an expert's overall impression, the rule usually predicts better.



The same facts, combined by overall impression: the answer depends on who is looking, and when.



The same facts, combined by a fixed rule: one answer, every time, open to checking.

- Across the studies reviewed by Grove and colleagues (2000), mechanical prediction was about 10% more accurate on average. It was substantially better in 33% to 47% of studies, and substantially worse in 6% to 16%.
- In selection and admissions decisions, combining the same data mechanically rather than holistically improved the prediction of job performance by more than 50%, and the loss from holistic combination held even for experts (Kuncel, Klieger, Connelly & Ones, 2013).
- Even crude linear models, with equal weights, outperform clinical intuition (Dawes, 1979).
- Managers who hired against test recommendations ended up with worse hires on average (Hoffman, Kahn & Li, 2018); different interviewers at the same firm make different decisions about indistinguishable applicants (Kahneman, Sibony & Sunstein, 2021).

So scoring is done the same way for every profile, and a feeling does not reorder the ranking. Judgment comes back where it belongs: a person reads the top of the ranking, compares the closest candidates head to head, calls them and writes the brief. The system ranks; it never decides.

5 The structured interview

The strongest single predictor in the 2022 re-analysis is the structured interview, at .42. The conversational kind comes out at .19. Structure is the difference.

The pattern is old and stable. McDaniel, Whetzel, Schmidt and Maurer (1994) found structured interviews more valid than unstructured ones a generation ago (.44 against .33 for job performance). Campion, Palmer and Campion (1997) identified fifteen components of structure, some shaping what is asked and some shaping how answers are evaluated. Huffcutt and Arthur (1994) found validity rising with structure up to a ceiling, past which more structure adds essentially nothing. Levashina, Hartwell, Morgeson and Campion (2014) reviewed the research that followed.

Same questions, same order

Built from the role's scorecard and asked of every candidate in the same way, inside a fixed time.

Anchored scoring

Each answer is rated during the call against written descriptions of weak, adequate and strong answers. In studies reviewed by Levashina et al., anchored rating scales went with higher validity (.35 against .26).

Planned follow-ups

Probes are written in advance and used the same way for everyone, only when an answer is thin.

A transcript

Calls are recorded with consent and transcribed, so a rating can be checked against what was said.

Two independent ratings

The interviewer rates during the call; a second rater scores the transcript without seeing those scores. Disagreements beyond a set gap go to a person, never averaged away.

A blind interviewer

The person running the call does not see the profile score, so the ranking cannot colour the rating.

The call takes twenty minutes and happens before you spend an hour. It does not replace your interviews; it decides who is worth them.

6 How we check ourselves

A method that is never checked against outcomes is an opinion. These are the checks.

1

After each shortlist

The hiring manager rates every candidate we presented and tells us who they chose to interview. Those ratings are compared with our ranking.

2

At ninety days

Whether the hire is still there, and how the manager rates their work. If they have left, the replacement search runs at no charge.

3

Weights reviewed against results

Scoring weights move only when enough outcomes have accumulated to say something, by small steps, and every change is logged with the numbers behind it. One hire going well or badly changes nothing.

Keystone is new, and these checks begin with our first search. Until outcomes accumulate, the method rests on the published research summarised here, which is why we cite it.

What this document leaves out

Which attributes we score, at which weights, read from which signals. That is the method. Everything else is on this page and on our website, including our terms.

7 The research

- Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, 94(6), 1394–1411.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1), 1–26.
- Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, 50(3), 655–702.
- Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making*, 8(5), 512–520.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30.
- Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, 1(3), 333–342.
- Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, 133(2), 765–800.
- Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, 79(2), 184–190.
- Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.
- Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, 98(6), 1060–1072.
- Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, 67(1), 241–293.
- McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79(4), 599–616.
- Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, 58(4), 1009–1037.
- Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management*, 41(2), 149–174.
- Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, 107(11), 2040–2068.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262–274.

Every figure in this document was checked against the source paper (abstract, table or full text) before publication. Where a figure could not be checked, the finding is stated without it.