Four Numbers and No Counterfactual: Reading the Artificial Intelligence Case Study in Human Resources

Heather Grizzle · May 23, 2025

Four percentage improvements and no counterfactual. What each standard HR AI case study metric requires to be true, and the questions that take ten minutes.

Contents

The artificial intelligence case study in human resources has settled into a stable form.

The genre

The artificial intelligence case study in human resources has settled into a stable form. An organization describes a deployment, reports three or four percentage improvements, adds a claim about bias mitigation, and concludes that talent has become a source of competitive advantage. The numbers are specific. The methodology is absent. The conclusion invites a conversation.

The form persists because it works as persuasion and because nobody is positioned to check it. Vendors publish it because it sells. Practitioners publish it because it establishes standing. Buyers accept it because the alternative is admitting they have no basis for the decision they are about to make. Everyone in the chain has an interest in the number and nobody has an interest in the denominator.

Everyone in the chain has an interest in the number and nobody has an interest in the denominator.

Novara Consulting Group's position is that these documents should be read as advertising claims rather than as evidence, that the specific figures most often cited are the ones most easily produced without any underlying improvement, and that a buyer capable of asking four questions can distinguish a real result from a reported one in about ten minutes. What follows is that method, applied to the four claims that appear most frequently.

Claim one: reduction in time-to-fill

Time-to-fill is the most commonly reported metric and the most easily moved without changing anything about hiring.

The measurement requires a defined start point and a defined end point, held constant across the comparison periods. In practice the start point varies by organization and often by requisition: it may be the date a requisition was approved, the date it was posted, the date sourcing began, or the date the hiring manager first raised the need. Moving that start point later shortens the measured interval without shortening anything experienced by anyone. So does excluding cancelled or withdrawn requisitions, excluding roles that took longest, or changing the mix of roles in the denominator.

The confound is larger than the effect. Time-to-fill responds strongly to labor market conditions, and the relevant comparison periods for most current case studies span a shift from acute labor scarcity to considerably looser conditions. An organization that changed nothing would show improvement across that window. Attributing the movement to a system deployed during it requires a counterfactual, and no case study in this genre contains one.

The qualifier deserves attention too. Improvement reported for key roles rather than all roles indicates a subset, and where the subset was selected after the results were known, the figure describes the selection rather than the system.

Claim two: reduction in manual transactional activity

This claim requires a baseline measurement of manual activity, and almost no human resources function has one.

Where a figure exists it has usually been derived one of three ways: from a vendor's implementation model, from a survey in which staff estimated their own time allocation, or from a count of tickets or transactions that changed definition when the new system was introduced. None of these supports a percentage comparison across time, because the instrument changed at the same moment as the intervention.

There is also a scope question the metric conceals. Transactional work that disappears from the human resources function frequently reappears elsewhere, most often with employees and line managers through self-service. A sixty percent reduction in transactions handled by the function may represent a genuine efficiency gain, a transfer of the same work to people whose time is not being counted, or some combination. Distinguishing these requires measuring total organizational effort, which nobody does, because the metric exists to justify the function's investment rather than to describe the organization's cost.

Claim three: increase in high-potential retention

This is the most interesting of the four, because the mechanism that produces the number is usually the deployment itself.

High-potential is not an observed characteristic. It is a designation the organization assigns, and an artificial intelligence talent system typically changes how it is assigned. That is the stated purpose. If the system alters which employees are designated high-potential, then the cohort measured after deployment is a different population from the cohort measured before, and a retention comparison between them is not a comparison of the same thing.

The direction of the resulting bias is predictable. A model identifying high potential from performance history, engagement signals, and career trajectory will preferentially designate employees who are engaged, recently promoted, and progressing, all of which are among the strongest available predictors of staying. The system selects people who were going to stay and the organization records their staying as an effect of the system.

Cohort size compounds this. In most organizations the high-potential population is small enough that an eighteen percent change represents a handful of individuals, well inside the variation any small cohort produces year to year. A figure of this kind should be accompanied by the cohort size and by whether the definition changed. It almost never is.

Claim four: bias mitigation

The fourth claim is different from the others in kind, because it is not merely unsupported. It is a legal representation.

Substantiating it requires measuring bias before and after, on defined groups, with cell sizes adequate to support the comparison. It also requires that the groups be known, which is where the claim usually fails. Disability is not measurable this way at all: employers do not know and may not lawfully ask before a conditional offer which applicants are disabled, so no impact-ratio instrument can register disability discrimination. An organization asserting bias mitigation in succession planning has generally measured race and sex, if it has measured anything, and has said nothing about the population most affected by performance-data-driven models.

Succession planning carries a further difficulty. A model identifying readiness or potential from tenure, career stage, time in role, credential recency, and pace of prior advancement is drawing on signals heavily correlated with age. Reporting bias mitigation for such a system, without having examined its age distribution, asserts the opposite of what the inputs predict.

And the claim itself creates exposure. Title VII, as amended in 1991, makes it unlawful in connection with the selection or referral of candidates for employment or promotion to adjust scores, use different cutoff scores, or otherwise alter the results of employment-related tests on the basis of race, color, religion, sex, or national origin. Nothing here suggests these systems do that. The point is that a public statement connecting an automated promotion tool to demographic outcomes is a document the organization wrote, cannot retract, and may be asked to explain.

What is missing from all four

Each of the preceding failures is a version of the same absence. None of these case studies contains a counterfactual, which is to say a description of what would have happened without the intervention.

The organization deployed a system during a period in which the labor market shifted, headcount changed, managers turned over, compensation was adjusted, and other initiatives ran concurrently. It then attributed movement in four metrics to one of those changes. This is not a small methodological quibble. It is the entire question, and its omission is what makes the genre unfalsifiable.

The published evidence on adjacent claims should give any buyer pause. When Harvard Business School and the Burning Glass Institute examined what actually happened at firms that publicly adopted skills-based hiring, they found that dropping degree requirements moved the hiring of non-degreed candidates by roughly three and a half percentage points, amounting to fewer than one in seven hundred hires, with close to half of firms showing no behavioral change at all despite their announcements. The announcements were sincere. The practice diverged completely, and the organizations making them did not know.

The relevant inference is not that people are lying. It is that organizations reliably do not know whether their own initiatives worked, and that a case study is written by the party least positioned to find out.

Why the claimant usually believes it

It is worth stating this directly, because the alternative reading is uncharitable and generally wrong.

These figures are rarely fabricated. They are produced by a measurement environment in which no baseline was established, the definitions moved, the vendor supplied the analytical framework, and the person compiling the results is the person who advocated the purchase. Every one of those conditions biases the output in the same direction, and none of them requires anyone to act in bad faith. The number is real in the sense that someone computed it. It simply does not mean what the sentence around it says.

This is also why the regulatory environment has begun to move. The Federal Trade Commission has been active against unsupported artificial intelligence capability claims, including a January 2025 order requiring an accessibility software provider to pay one million dollars over representations that its automated product could bring any website into conformance with accessibility standards. The relevant principle is not confined to that product category. Performance claims about artificial intelligence require substantiation, and an organization publishing outcome percentages in a marketing document is making a claim of exactly that kind.

What a buyer should ask

Four questions, which take about ten minutes and which the great majority of case studies cannot survive.

What was the baseline, when was it measured, and by whom. If the baseline was constructed after deployment, or supplied by the vendor, the comparison is not one.

Did the definition change. For any metric involving a designated population, such as high-potential or key roles, ask whether the designation criteria were the same before and after. If the system changed them, the populations differ and the comparison is void.

What else changed during the period. Labor market, headcount, compensation, management, concurrent initiatives. An honest respondent will have a list. A respondent

without one has not looked.

What is the denominator, and how many people. Percentages without cohort sizes are uninterpretable, and small cohorts produce large percentages from small movements.

A FIFTH QUESTION

A fifth question is worth asking of any claim touching bias: which groups were measured, at what cell sizes, and what was done about the groups that cannot be measured. The answer to the last part is almost always that nothing was, which is itself the most useful thing the buyer will learn.

What a credible case study would contain

The genre is not beyond repair, and an organization willing to publish the following would immediately distinguish itself in a field where nobody does.

State the baseline and its date. State the definitions and note any that changed. Report cohort sizes alongside percentages. Name the concurrent changes that could account for movement, and say which ones you cannot rule out. Describe what did not work, because a deployment in which everything improved is a deployment in which nothing was measured. And where a bias claim is made, state the groups, the method, the cell sizes, and what remains unmeasured.

That document is less impressive than the current form and considerably more useful. It is also the only version that survives being read by someone with an interest in whether it is true.

Conclusion

The case study is the primary evidentiary artifact in enterprise human resources technology purchasing, and it is the weakest form of evidence in routine professional use. It is single-organization, uncontrolled, self-reported, retrospectively defined, and published by a party with an interest in the finding.

None of that makes it worthless. It makes it a hypothesis. The organizations that will actually realize returns from these systems are the ones that treat other people's numbers as a reason to investigate rather than a reason to buy, and that hold their own numbers to the standard they would want applied to a vendor's. The discipline is not difficult and it is not expensive. It is a baseline, a stable definition, a cohort size, and an honest account of what else was happening at the time.

References

- Federal Trade Commission, *In the Matter of accessiBe, Inc.*, File No. 2223156, complaint and proposed order announced January 3, 2025.
- Fuller, Joseph, et al. *Skills-Based Hiring: The Long Road from Pronouncements to Practice*. Harvard Business School Project on Managing the Future of Work and The Burning Glass Institute, February 2024.
- Title VII of the Civil Rights Act of 1964, as amended, 42 U.S.C. § 2000e-2, including § 2000e-2(l).
- Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. Part 1607, applicable to selection procedures affecting hiring, promotion, and other employment opportunities.