What benchmarks are actually for

2026-09-16 · 6 min read

THE SHORT VERSION

- One raw point on a hundred-point engagement index moved an organization from the 60th percentile to the 75th.
- Work the arithmetic backwards and the standard deviation between employers is about 2.4 points, so 95 percent of benchmarked organizations sit inside a ten-point band.
- Two of the three jobs a benchmark does have an internal answer: directional items fix comparability, driver analysis finds the returns.
- Norms still earn their cost on soft organizational items where no better external dataset exists, which includes the one this whole argument started with.

On one engagement survey I worked on, compensation came back as the worst item on the page. Somewhere around 66, against an average of about 77 across everything else we asked. Eleven points below the rest of the survey. From inside the building there is only one sensible reading of that, which is that pay is the problem and pay is where the money should go.

Then we looked at the benchmark. The norm, across millions of responses, was about 64. We were above it. The crisis of that survey sat somewhere around the 75th percentile.

(From memory, and rounded. The gap is the part I am sure about.)

That is a benchmark doing exactly the job a benchmark is for, and it stopped an organization spending real money on something that was not broken. I am not going to pretend otherwise, because the same index did something else two cycles later.

It went up one point. Seventy-six to seventy-seven, on a hundred-point scale.

That one point moved us from the 60th percentile to the 75th.

A percentile is a rank

A percentile tells you how many organizations you passed. It says nothing about how far you traveled to pass them. Those two come apart the moment the organizations are packed closely together, and the arithmetic of my one point says exactly how closely.

Take the norm distribution as roughly normal, which the standard T-score convention assumes by construction. The 60th percentile sits about 0.25 standard deviations above the mean and the 75th

sits about 0.67. So one raw point covered 0.42 of a standard deviation, which puts the standard deviation between organizations at roughly **2.4 points on a hundred-point index**.

> **2.4**
>
> The standard deviation between organizations on a hundred-point engagement index, derived from a single raw point crossing fifteen percentile places.

Sixty-eight percent of the employers in that norm group therefore sit inside a band under five points wide. Ninety-five percent sit inside a band under ten.

> Every benchmarked employer in that database is within about ten points of every other one, and the percentile is doing the work of making a ten-point world look like a hundred-point one.

There is a second problem stacked on top of that, and nobody prints it next to the number. Two organizations half a point apart are about eight percentile places apart. My half point was real: at more than 20,000 respondents per cycle, the standard error on that index is around a tenth of a point. Most organizations in a norm database are nothing like that size. A 2,000-person company's standard error is roughly three times mine, and a 500-person company's is about six times. Their sampling noise moves them eight percentile places as fast as my signal moved me. The ranking reports both with the same confident precision.

The obvious fix is a tighter norm group. That fix kills the tool. A tiny company compared against the Fortune 500 is not apples to apples. Even inside one organization, field against corporate, salaried against hourly, there are group differences big enough to matter. All of which argues for really specific norms. But the tighter the group has to be to be fair, the fewer organizations are in it, the less stable it becomes, and the longer you wait for one. A norm is only trustworthy when it is specific, and specificity is exactly what makes it scarce.

The two jobs people actually buy them for

Strip away the reporting habit and benchmarks do two things worth paying for. Both have an internal answer.

Putting your items on a comparable footing. This is the compensation save, and it is a real one. The reason that item needed a benchmark to be readable, though, is that it was written on an agree-to-disagree scale, which produces a score and no direction. A directional item carries its own instruction: every question answered do-less to do-more, so a negative average means do less of this and a positive one means do more. Read that way, compensation does not need an external

reference at all.

Telling you where the returns are. Driver analysis answers this from inside your own data. Which items actually move your index, crossed with where you currently sit, tells you where effort pays. It is the better answer, because it prioritizes against your drivers rather than against the average employer's.

The second has a real flaw, and it belongs in the open. When nearly everyone in an organization rates an item badly, the variance on that item collapses, and a collapsed variance attenuates its correlation with anything. So the item everybody agrees is broken tends to look like a weak driver, and it looks that way precisely because everybody agrees. This is a live argument in the I-O literature rather than a technicality: the exchange in Industrial and Organizational Psychology between "Survey Key Driver Analysis: Are We Driving Down the Right Road?" and "In Defense of Responsible Survey Key Driver Analysis" is where to read it. Range restriction is the specific blind spot of driver analysis, and it lands precisely on a uniformly low item holding a floor.

The rank order does most of the work

The between-organization standard deviation at the index level is about 2.4 points. Converting that to the item level, which takes an assumption about how tightly the six items hang together, gives something like 2.7 to 3.1 points. My compensation gap was eleven points below my own other items.

So the spread between your own survey topics is three and a half to four times the standard deviation between employers on any one of them. The differences between topics are large. The differences between companies on a topic are small.

Which means knowing the typical rank order of topics captures most of the available signal, and the organization-specific norm is a second-order correction sold as a first-order one. Perceptyx, which sells benchmarking, reports from a 2026 database of more than 23 million responses that pay and benefits are typically the lowest-scoring items and rarely predict who becomes engaged. A benchmarking vendor publishing that the lowest item is usually the same item everywhere, and usually does not matter much, is a better citation than anything I could produce.

And for almost any item where the level does matter, there is a better external dataset than an engagement norm. Compensation has compensation benchmarking. Turnover has industry and government data. Those sources are higher resolution, more specific and more defensible than a percentile on a six-item index.

What a norm is still for

The concession does not vanish. It shrinks to something narrow and real.

A norm still earns its cost on items where the level matters and no better external dataset exists. That is a short list, and it is mostly the soft organizational ones. Trust in senior leadership. Whether change is handled well. Effective collaboration across teams.

I have to sit with that last one, because it is the item I have spent the most time on. "There is effective collaboration across teams in this organization" came back third lowest of fifty on that same survey, with the highest correlation to engagement of anything we asked. There is no Mercer for it. There is no government series. If I want to know whether my people find it unusually hard to work across a boundary, or whether everybody's people do, the norm group is the only thing that answers.

So the argument does not rescue the question I care most about. It rescues most of the others, which is enough to change what you buy. Look at your own lowest item and ask whether a better dataset exists for it. If one does, the engagement percentile was never going to be your best evidence. If one does not, that is the item worth paying a benchmark for, and it is probably not the one you were about to spend the money on.