Benchmarking Your Way to Mediocrity

An exposé of the pitfalls associated with externally benchmarking your employee engagement metrics

Many organisations feel the urge to put their engagement survey metrics 'into context' by comparing their results against those of other employers. Most research practitioners will step up to the plate to offer you a 'normative' database of results from a wide range of clients, including of course from your own industry sector.

Sadly, these comparisons are fundamentally flawed in a number of ways and can even be perilously misleading if allowed to influence your corporate decision-making. External benchmarking of your survey results can bear little fruit, and the bitter taste of making the wrong decisions will remain long after the sweet taste of believing you are marginally better than your competitors.

FEBRUARY 2015

GRAVITAS
ANALYTICS

.

Liability Notice
This article is presented as is, without warranty of any kind, either express or implied, respecting the contents of this article, including but not limited to implied warranties for the article's quality, performance, or fitness for any particular purpose. Gravitas Analytics shall not be liable to the reader of this article or any other person or entity with respect to liability, loss, or damage caused by, or alleged to have been caused directly or indirectly by this article.

Introduction

The need to compare 'apples with apples' sensibly underpins comparative analysis across all spheres of research and is by no means a new concept. However, when it comes to comparing survey responses in organisations, this wisdom is often forgotten in favour of embracing a basket of mixed fruit.

Employers usually set out to understand what their employees are thinking, how they are feeling, whether they are engaged and aligned to the organisation and, most importantly, what is driving their behaviour. Carrying out an engagement survey is a tried-and-tested way to achieve this, and survey results can provide the key organic information you need to develop your people and culture. The desire to benchmark your results against other employers is understandable, and benchmarking clearly has its place in many business applications, but achieving like-for-like comparisons between survey results is nigh on impossible. Therefore, allowing supposed normative comparisons to fuel decision-making is high risk and can seriously undermine the value of the internal metrics derived from the survey instrument itself.

External benchmarking of survey results is quite simply a bad idea, and this paper summarises why under the headings of Organisational Disparity and Measurement Disparity.

Author
Chris Casson
Founder
Gravitas Analytics

Organisational Disparity

DIFFERENTIATION

The very nature of competition means that organisations in all sectors actively seek to differentiate themselves in the marketplace. Let's face it, **if asked to describe your organisation, would you list all the features that make you the same as everyone else?** If you were preparing a sales pitch you would be focused on listing the unique selling points that differentiate your organisation from the rest of the playing field. This active drive to differentiate produces workplace diversity in organisational structures, job specifications, modus operandi, quality, service offerings, objectives and rewards, to name just a few. Taken together, this diversity shapes the face of every organisation and delivers corporate individuality.

EMPLOYEE EXPECTATIONS

Employees have expectations about all aspects of their worklife, from leadership integrity, learning and development and career opportunities, to workplace relationships, health and well-being and work–life balance. Every organisation will fall short, meet or exceed employee expectations to different degrees and will treat people differently in terms of dignity, respect and involvement. Ironically, the level of expectation for each element of worklife is essentially driven by what you actually deliver as an organisation. In effect, expectations are what employees have come to expect from you, and these may be widely different from one employer to the next. In other words, **survey responses are the result of people's experiences compared with what they expected**. Therefore, assuming an identical workplace experience, employees with lower expectations tend to score more positively than those who have higher expectations. From an external benchmarking perspective, you have no idea how the employees of other organisations expect to be treated, so you cannot realistically draw any comparison.

CORPORATE CULTURE

On a more macro scale, organisations can differ culturally in many ways, and even subtle differences in the baseline fabric of an enterprise can have a significant influence upon your employees' perceptions and their levels of engagement and enablement. The corporate values that you embody, your management style and the way you go about your business will foster an internal branding that directly affects employee characteristics and behaviours. **Ultimately, every organisation develops a unique 'persona', with its employees in the driving seat**. Your people are undoubtedly what makes your organisation different and what drives engagement in one organisation will not necessarily be the same in yours, even if it's a close competitor or a structural clone.

TARGETING MEDIOCRITY

Some eminent business leaders and accomplished shapers define benchmarking as 'clear evidence of the absence of strategy', or an approach for sheep flocking to the middle ground. When your engagement effort aims for the norm then the norm is what you will get. Before aspiring to be just average, consider the time, energy, cost and commitment you have put into your survey initiative and the precious internal data you have just gleaned – don't get 'benchmark complacent' in the final stages of your project. **Most often a benchmark will be calculated by comparing both high-performing and low-performing organisations, giving your establishment, by default, a target of average performance**. If you are not so 'fortunate', your benchmark may be taken from the lower-performing quartile, giving you an even smaller yardstick. The harsh reality, regardless of practitioner claims, is that your benchmark is then highly likely to be pedestrian.

INDUSTRY SECTORS

Most organisations embarking upon benchmarking wish to assess their survey results against competitors within their industry sector. However, organisations vary considerably in terms of size, design, sector coverage, portfolio, geographical positioning and maturity, so trying to consolidate survey data around a mean (average) will produce a significant variance. The mean is essentially the benchmark so it follows that the more variance around the mean, the more meaningless the benchmark. Or worse, the comparative sample used may consist only of the provider's own clients (similar or disparate) and, **given the breadth and diversity of organisations in each sector, the normative database used may not be representative of the wider industry**.

On the flip side, by focusing solely on your own industry sector you will deny yourself visibility of employee perceptions across industries that your talent may be considering moving to, or indeed coming from. Employees with transferable skills and working in generic disciplines (human resources, administration, accounts, IT, projects, training etc.) are employable across a broad spectrum of industries, and there is no barrier to mobility in the current environment.

EVOLVING WORKPLACE

Both the public and private sectors have seen many changes over the years, with major advances in technology transforming the workplace, a manufacturing downturn and service uplift. In addition, people-friendly legislation, the influence of social media and serious economic turbulence are all changing the shape of the work environment. There is now a greater reliance on employee performance, and at the same time employees have greater expectations of their employers. The point here is that the workplace is a moving feast, trending in different directions and constantly changing its face. However, in order to increase the sample size and add credibility, **benchmarks are frequently calculated from many years of survey data, which paradoxically has a negative effect on data credibility** as substantial historical trends in the workplace will skew the overall results.

PRIORITISING ACTIONS

The most critical output you should look for in any survey package is the identification of key drivers[1] – what is driving engagement and loyalty. **Benchmarks have no relevance here as key drivers are very much organisation specific**. Key drivers that produce negative responses from the workforce should be top of your list for action, closely followed by preserving the factors that produce favourable responses. Benchmarking can frequently lead to misplaced priorities. For example, if your organisation scores low on communication compared with your competitors, but it was not identified as a key driver in your survey metrics, it is not a priority[2], even though your benchmark may lead you to believe it is. Moreover, if you take action on any item based on benchmark data, without understanding the impact it will have on your own workforce, you could be embarking upon a change management programme that will do more harm than good.

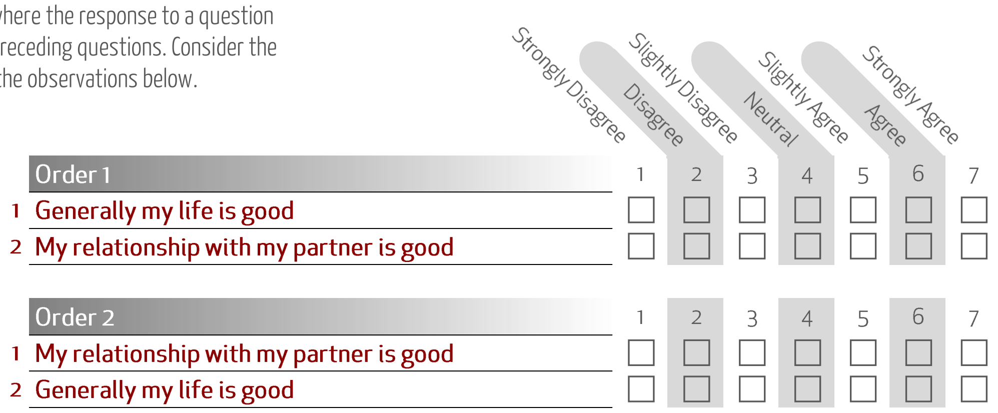
[1] If your chosen research provider does not use advanced inferential techniques within its statistical toolkit to isolate the significant predictors (key drivers) in your organisation, and is relying solely on correlations, or even worse basic descriptives (% pos/neg), then that is a big issue beyond the scope of this paper. Please take a look at the Gravitas white paper 'Survey DIY & Customisation Exposed' for more information.

[2] For clarity, if your survey questionnaire has strong construct validity (i.e. it comprehensively measures what it is supposed to measure) then all survey items will have practical importance. For the purposes of this example communication is sedentary.

Measurement Disparity

QUESTION SEQUENCE

Responses to survey questions can be influenced by previous questions and by previous answers. **The order of survey questions therefore has a direct impact on how respondents will interpret and respond to the overall questionnaire**. When question order is not considered during survey design several problems can occur, most notably bias and priming, where the response to a question is inadvertently conditioned by preceding questions. Consider the example questions in Fig. 1 and the observations below.



	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
Order 1	1	2	3	4	5	6	7
1 Generally my life is good	□	□	□	□	□	□	□
2 My relationship with my partner is good	□	□	□	□	□	□	□
Order 2	1	2	3	4	5	6	7
1 My relationship with my partner is good	□	□	□	□	□	□	□
2 Generally my life is good	□	□	□	□	□	□	□

Fig. 1

1st Observation

In Order 1, 'Generally my life is good' will be interpreted as including the relationship context because it is asked prior to 'My relationship with my partner is good' and so is not influenced in any way. Conversely, in Order 2, 'Generally my life is good' may be interpreted as excluding the relationship context because it is asked after 'My relationship with my partner is good'.

2nd Observation

In Order 2, by asking respondents an explicit question about a relationship first, we have narrowed their perception when it comes to answering the wider question about life in general. Consequently, the response to 'Generally my life is good' may not be as open and objective as hoped and may also be influenced by any positive or negative exchanges they have had recently with their partner.

3rd Observation

The 1st and 2nd observations show how a question can influence the response to another, pushing it up or down the agreement scale. Another factor is how question order can affect the correlation between responses, which is an important feature of statistical modelling. In Order 2, given the influence of relationship over life in general, you would expect to get similar answers from different respondents, i.e. they would be more correlated. Research has shown this to be the case. In our example, the correlation of answers for each question order was:

Order 1	.18	[low correlation]
Order 2	.67	[high correlation]

It follows that, even if two surveys appear well structured in terms of their question order, there may be nuances that produce clear differences in the responses. **The only way of ensuring meaningful comparison across surveys is by making them identical**. From a benchmarking perspective this raises two issues: (a) you may be benchmarking against results from surveys that are poorly designed; (b) the chances of all your benchmark results coming from surveys that are identical to yours is extremely low, even when you are using results from your practitioner's own normative database.

QUESTION PHRASING

Survey questions can be worded in many different ways, and there are many different ways in which survey questions can be worded poorly. These include biased, ambiguous, leading or framed questions, to name a few. Apart from the obvious possibility that you may be benchmarking your own well-structured survey against results generated from poorly worded surveys, there are also some important subtleties to consider. It could be that your benchmark surveys were in themselves suitably worded, but the questions in your own survey instrument, while exploring similar areas, were not phrased in exactly the same way. **Differences in the wording of questions will elicit different responses, which means that drawing comparisons between 'equivalent' responses is unreliable**. Small changes to question wording can have a substantial impact upon the responses received (25%+ variance).

Consider the classic statistical examples in Fig. 2. Although not directly relevant to employee engagement, these examples show how phrasing can significantly influence survey responses.

	Wording 1	MORE GOOD	MORE HARM
1	Do you think that, in the long run, atomic energy will do more good than harm?	69%	31%
2	Do you think that, in the long run, atomic energy will do more harm than good?	56%	44%

	Wording 2	YES	NO
1	Do you think the US will go into the war before it is over?	41%	33%
2	Do you think the US will succeed in staying out of the war?	44%	30%

[Don't Know 26%]

Fig. 2

Wording 1

Both questions in Wording 1 are effectively asking the respondent to make the same decision on harm or good, but the questions elicit different responses. People who are uncertain about how to answer a question (because of its sensitivity, difficulty or complexity) will often choose what seems to be the most socially acceptable answer. In this case, the complex subject of atomic energy is met with a bias towards the socially palatable 'more good'. The question is also posed in a way that makes it easier for those who are uncertain to respond with a simple 'yes' to the leading proposition presented to them, hence the increase in 'harm' responses to the 2nd question.

Wording 2

This is a good example of where changing the wording of a question can completely reverse the outcome of your enquiry. The number of people who believed that the US would, or would not, go into the war was dependent upon how this question was worded.

RESPONSE FORMAT

Even if your own survey questions match those in your benchmarked surveys, employees might have been asked to respond on a different scale. There are many different scales available, and used, within the research industry, which seriously challenges the statistical reliability of survey benchmarking. **The notion of manipulating and normalising survey responses using different scales to create comparable indexes simply does not wash**. Even where common approaches are used, such as the agreement scale in Fig. 1, the number of points on the scale could be 3, 5, 7 or 10 – there is no standard number. An approach that has become all too commonplace within the world of employee surveys is using the 5-point agreement (or Likert) scale. Employee survey practitioners have adopted this scale as it is widely used across all spectrums of research and so it is the easy option. While the 5-point Likert scale is used throughout social and psychological research to great effect, when it comes to employee surveys it quite simply fails to perform and often proves to be the Achilles' heel of many engagement endeavours. In order to appreciate the importance of this issue it is covered in more detail below.

The 5-point scale was developed for scenarios where there was a reasonable chance of data being normally distributed, i.e. with the prospect of people scoring an equal number of 1s and 5s (Strongly Disagree and Strongly Agree). However, **because your respondents are employed by you, your organisation must be of value to them**, which makes the respondent pool positively biased. It is reasonable to assume that any disgruntled employees who might have scored highly unfavourable responses have already thrown in the towel and moved on. This then leaves your organisation with a pool of respondents who are, by nature, likely to be scoring more 3s, 4s and 5s than 1s and 2s, making the survey results positively skewed. Assuming that 1 responses are not going to feature heavily at all and that we know 3s (Neither Agree nor Disagree) offer comfort to the respondent but nothing to you statistically, you are essentially left with a 3-point scale. The 5-point scale is quite simply not sensitive enough to capture a true evaluation from the respondents, and indeed they can become frustrated at not being able to express their opinions properly – despondent respondents!

The 5-point agreement scale will leave you with nothing more than a coarse estimate, rather than the accurate subjective evaluation you are looking for.

Extensive usability research on the sensitivity of Likert scales has concluded that 5-point scales are unable to capture the subtlety of opinion that participants want to express. Tests have shown that people will 'interpolate' when confronted with a 5-point scale, i.e. they have an opinion that sits between two options on the scale, say 3.5, and are then forced to select either 3 or 4. This interpolation can occur across 20% of the respondent pool, which significantly reduces the accuracy of an already insensitive scale.

So, if a 5-point scale does not cut the mustard, what does? Tests have shown that although a 10-point scale offers much greater sensitivity and reliability, it suffers in several key areas: there is no comforting mid-point on the scale; people tend to psychologically migrate to a 'marks out of 10' mindset; and the scale has too many options. Furthermore, reliability of the scale plateaus at around 7 and extending beyond this point is of no real value.

The conclusion from the scientific research is that a 7-point Likert scale is the 'sweet spot' for employee surveys and achieves the sensitivity required in the face of any positive bias. The 7-point scale reduces data loss because respondents who want to choose 3.5 can do so and that 0.5 is not lost by being forced to choose 3 or 4. (In usability tests respondent interpolations on a 7-point scale were 0%.) Also, 7-point scales have proved to be more accurate and easier to use, more statistically reliable and a much closer reflection of people's opinions. They also enable better year-on-year tracking of the steady progress you are making as your results do not suffer from a limited response range. Finally, and most importantly, all this adds up to being able to get your hands on more accurate and informed metrics to fuel smarter decision-making across your organisation.

Appointing a practitioner with a survey instrument scaled at 7 points is the first step in the right direction. Be aware, however, that the bulk of employee surveys available in the marketplace are based on the less reliable 5-point scale.

CONSTRUCT VALIDITY

For many years employee surveys have been designed to measure generic attitudes and opinions in organisations and to report on levels of job satisfaction. All too often **employee satisfaction surveys have simply been repackaged or relabelled as engagement surveys without due consideration of the structure of the enquiry**. These do not target the affective characteristics of psychological engagement, such as drive, energy and passion, and most certainly won't identify what is required to help you leverage engagement and ultimately competitive advantage. The content of your survey must focus on measuring the 'construct of interest'. For example, if engagement is your target then the survey instrument must comprehensively map the construct of engagement: does your survey data and your benchmarked data measure levels of psychological engagement, transformational leadership, job-motivating potential and cognitive alignment? The answer should be yes.

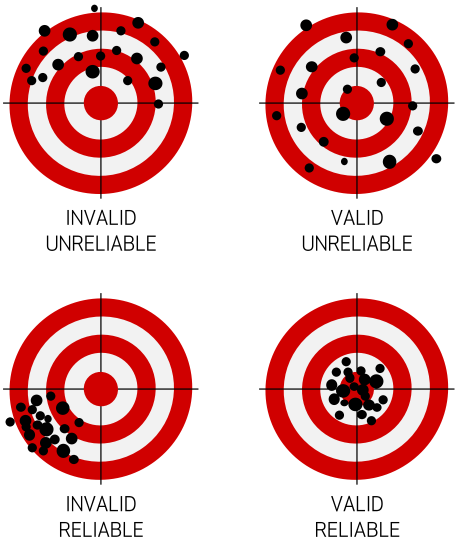
In the marketplace you will find countless definitions of engagement, with many consultants attempting to brand their own 'engagement index' and laying claim to a comparative benchmarking database. Be very wary of such claims. If you are serious about understanding the true construct of engagement, its conceptual framework and how to measure it, you need to understand the scientific and academic research behind it. You can save yourself several years of in-depth enquiry by taking a look at the Gravitas white paper 'The Engagement Imperative'. The lack of critical thinking in this area and the conceptual disparity among consultants means that (a) it is likely that your benchmarks are not measuring real engagement, and (b) it is likely that your survey is measuring a different construct to that of your benchmarks.

it is likely that your benchmarks are not measuring real engagement.

FACTOR RELIABILITY

Credible survey design means your questions must be psychometrically sound and provide proper coverage of the construct and sub-constructs being measured. Examples of sub-constructs of engagement might be: leadership, development, recognition or communication. These are the sections or factors that provide structure to your survey. **If the factors in your benchmarked surveys are poorly designed, and the questions don't measure anything accurately, then any associated benchmarks will be meaningless**. Individual questions must be precise, actionable, of practical importance and provide a meaningful measure of the sub-construct. Most likely you won't get sight of the benchmarked questions, so ask your survey provider for copies of published, peer-reviewed research supporting the validity and reliability of the engagement construct and sub-constructs in the benchmarked surveys. How the consultant responds should prove interesting! Fig. 3 shows how validity and reliability can affect your target. Think of the centre of the target as the concept that you are trying to measure.

Fig. 3



Reliability
Precision & consistency of survey items covering the construct; measuring signal as opposed to noise

Validity
Overall coverage of the construct's theoretical domain; the survey instrument measuring what is should be

Conclusion

Organisations can easily become distracted by comparing survey results against so-called normative benchmarks. The harsh reality is that this will do little more than bolster your survey practitioner's revenue stream and mask the important issues identified in your own engagement metrics. The most valuable benchmarks to fuel decision-making will come from your own organisation's statistical output, such as year-on-year progression, interdepartmental differences and variations in demographic opinion. All of these are accurate, unambiguous, relevant and reliable.

As a final thought, remember that the most critical output from an employee engagement survey is the identification of the key drivers, and these are completely specific to your organisation. So put external benchmarking on the back seat where it belongs and don't be sold on the assurance of comparing apples with apples – your benchmark 'apple' is highly likely to be another fruit altogether.

high-end employee engagement metrics for the discerning organisation

gravitasanalytics.com

References

Bishop, G.F. & Smith, A.E. (1997) 'Response-order Effects in Public Opinion Surveys: The Plausibility of Rival Hypotheses', paper presented to annual conference of the American Association for Public Opinion Research, Norfolk VA.

Cummins, R.A. & Gullone, E. (2000) 'Why we should not use 5-point Likert scales: the case for subjective quality of life measurement', in Proceedings of the second international conference on quality of life in cities, 74–93.

Diefenbach, M.A., Weinstein, N.D. & O'Reilly, J. (1993) 'Scales for assessing perceptions of health hazard susceptibility', Health Education Research, 8(2), 181–192.

Finstad, K. (2010) 'Response interpolation and scale sensitivity: evidence against 5-point scales', Journal of Usability Studies, 5(3), 104–110.

Miller, G.A. (1956) 'The magical number seven, plus or minus two: some limits on our capacity for processing information', Psychological Review, 63(2), 81.

Nunnally, J.C. (1978) Psychometric Theory, New York: McGraw-Hill.

Rugg, D. & Cantril, H. (1944) 'The Wording of Questions', in H. Cantril (ed.) Gauging Public Opinion, Princeton University Press.

Russell, C.J. & Bobko, P. (1992) 'Moderated regression analysis and Likert scales: too coarse for comfort', Journal of Applied Psychology, 77(3), 336.

Schwarz, N. & Strack, F. (2003) 'Reports of Subjective Well-Being: Judgmental Processes and Their Methodological Implications', in D. Kahneman, E. Diener & N. Schwarz (eds) Well-Being: The Foundations of Hedonic Psychology, New York: Russell Sage Foundation.