Performance Pulse

Generating reliable performance ratings

WHITE PAPER SERIES »

THE MARCUS BUCKINGHAM COMPANY

Introduction

Performance management systems are built to generate reliable data from which the organization can make informed decisions about how to pay, promote, train and deploy each team member. They fail to do this. As built, all ratings-based performance management systems generate unreliable data, which in turn compromises all downstream talent decisions. In this paper, we present Performance Pulse as a solution to this systemic failure.

The Causes of the Failure

There are two main reasons why current ratings-based systems generate unreliable data:1) the idiosyncratic rater effect and 2) the rater insufficiency effect.

Idiosyncratic Rater Effect

Over the last fifteen years, a significant body of research has demonstrated that each of us is a disturbingly unreliable rater of other people's skills and performance (Hoffman et al. 2011; Mount et al. 1998; Ng et al. 2011). The effect that ruins our ability to rate others has a name: the **Idiosyncratic Rater Effect**, which tells us that my rating of you on a quality such as "potential" is driven not by who you are, but instead by my own idiosyncrasies: how I define "potential," how much of it I think I have, how tough a rater I usually am. This effect is resilient: no amount of training seems able to lessen it. And it is large: on average, 62% of my rating of you is a reflection of me (Scullen, Mount & Goff 2000). When I rate you, on anything, my rating reveals to the world far more about me than it does about you.

In the world of psychometrics, this effect has been well documented. The first large study was undertaken in 1998 in *Personnel Psychology*. A second study was published in *The Journal of Applied Psychology* in 2000, and a third confirmatory analysis was published in 2010, again in *Personnel Psychology*. In each of the separate studies, the approach was the same: first ask peers, direct reports, and bosses to rate managers on a number of different performance competencies; and then examine the ratings (more than half a million of them across the three studies) to see what explained why the managers received the ratings they did. These studies found that more than half of the variation in a manager's ratings (71% in the first study, 58% in the second, and 55% in the third) could be explained by the unique rating patterns of the individual doing the rating. No other factor — not the manager's overall performance, not the source of the rating — explained more than 20% of the variance. The bottom line: when we look at a rating, we think it reveals something about the ratee, but it doesn't. Instead, it reveals a lot about the rater.

Rater Insufficiency Effect

The second source of unreliability comes from the **Rater Insufficiency Effect**.

Driven by a desire to understand and evaluate the full spectrum of performance, evaluation builders create evaluations that are ever longer and more complex in the types of competencies and skills being rated. In addition, they have added increasingly detailed rating instructions in an attempt to create consistency in rating scales.

Faced with these complex competency models and scales, team leaders charged with the task of evaluation must probe their long-term memories to think about how each employee did or did not meet expectations on each theoretical competency in the last year (Ghorpade, 2000). Many times, team leaders have insufficient knowledge of the competencies, but still have to provide a rating. Team leaders take shortcuts when the knowledge they possess about a team member is insufficient. Typically, what happens is that a team leader will simplify the process and use a general impression of the employee, ignoring any discrepancies — and thus introducing error into the performance equation (Ghorpade). In addition, team leaders often do not feel able to complete these complex systems of ratings, which in turn introduces self-efficacy error into the evaluation of performance, thus further reducing their ability to rate accurately and effectively (Westerman & Rosse, 1997).

Toward Better Precision

Given that we are all unreliable raters of other people's skills and performance, how can the organization ever create line of sight to the actual performance of each team member? An organization can follow these three simple steps to help minimize the effects of idiosyncratic rater effect and rater insufficiency.

1. Focus the survey on the team leader's intentions.

We should strive to measure, reliably and frequently, what each team leader intends to do with each team member. After all, at its heart the purpose of any performance management system is not to measure perfectly the performance of each person. As the research has shown, this is impossible. Instead, the point is to help the organization gather data to know how it should respond to each person's performance — in the form of increased pay, more training, or a promotion. And the best person to provide these data is the team leader. Will each team leader's intentions be subjective? Yes they will. But another word for "subjectivity" is "judgment," and what is the organization paying its team leaders for if not their judgment?

Thus the challenge becomes not "how can we make all our leaders objective?" but instead "how can we measure, reliably and frequently, the judgment of what a team leader intends to do with each team member?"

And this we can do, because human beings are reliable raters of their own intentions and feelings. So, to see what it should do with each team member, all the organization needs is a short survey that asks team leaders (at least four times a year) a few carefully worded questions about what they intend to do with each of their team members.

2. Neutralize each team leader's unique rating patterns

Even with items reworded to measure a team leader's intention, some biases and unique rating patterns will persist. To further neutralize these, we can measure each team leader's patterns, compare them to those of other leaders, and create an algorithm that modifies the raw data accordingly. If a team leader is excessively lenient, his or her scoring will be adjusted downward; if excessively harsh, it will be adjusted upward.

The benefit of this algorithm is not simply that it produces more precise data, but also that it becomes increasingly precise over time. Each time a team leader deploys the performance survey, the algorithm is fed more data, and thus can calibrate the output more accurately.

3. Calibrate for frequency of interaction

Within organizations, many team members report to multiple team leaders, each of whom should be able to give input into the team member's performance. However, the time spent with one team leader may be significantly different, in both quantity and quality, from time spent with another team leader. To account for this difference, the survey's data output should be weighted according to the frequency of interaction between the team leader and team member. The greater the frequency, the greater the weighting.

How can we measure this frequency? Some organizations ask team members to keep track of where their hours at work are spent; these hours can then be used for calibration. However, most organizations do not count hours. For these organizations, the most effective means of calibration is to record which team members check in with which team leaders, and how many Check-Ins have occurred.

Here are the four questions that the StandOut platform would use for an employee (let's call him Marc, for the purposes of this example):

1. I always go to Marc when I need extraordinary results.

(Asked on a 1–5 scale, this item measures a team leader's judgment of the person's productivity.)

2. I choose to work with Marc as much as I possibly can.

(Asked on a 1–5 scale, this item measures a team leader's judgment of the person's teamwork.)

3. I would promote Marc today if I could.

(Asked on a Y/N scale, this item is the most reliable way to judge potential.)

4. I think Marc has a performance problem that I need to address immediately.

(Asked on a Y/N scale, this item is the most reliable way to reveal the presence of "derailers.")

Aggregating the data from these questions, the organization will see, quarter by quarter, what it should do with each team member, based on the person best qualified to judge this: the team leader.

Below, we present what we have learned thus far from the most recent study of 59 unique organizations who had team leaders complete Performance Pulses on their teams. The teams contained on average 5.3 team members.

What we know so far...

1. Performance Pulse produces reliable data.

Performance Pulse measures one coherent factor that explains 53% of the variance in team member performance. Performance Pulse yields no other coherent factor.

Reliability of a measurement instrument is the extent to which a given instrument produces consistent results and is free from measurement error (Thorndike, 2004). We measured the reliability of the Performance Pulse data using coefficient alpha based on the following hypothesis: taken together, the questions measure one factor we could call "general performance."

The results from our Performance Pulse research are found in Table 1 below:

The sample of team member ratings, as measured by Q1, Q2, and Q3, yielded alphas above a .7, which is within the acceptable threshold for coefficient alphas (Matthews, Deary, and Whiteman, 2003). This provides evidence that team leaders are rating in a consistent manner across the three performance questions. (The fourth question displays a negative relationship to the other three and thus the alpha is lowered when it is added).

A Confirmatory Factor Analysis was performed on these data to test if they indeed measure a "general performance" factor. Even with the negative relationship of the fourth question, the Confirmatory Factor Analysis yielded only one coherent factor that explained 53% of the total variance. Thus, the hypothesis was provisionally confirmed.

Table 1: Reliability Coefficient Alphas for Performance Pulse

Questions Included	Reliability Coefficient
Q1, Q2	.86
Q1, Q2, Q3	.817*
Q1, Q2, Q3, Q4	.604*

**Coefficient alpha tends to underestimate reliability of dichotomously scored items. Q1 and Q2 are scored using a five-level Likert scale and Q3 and Q4 are dichotomously scored. When Q3 and Q4 were removed, the alpha increased.*

2. Team leaders are using the entire scale for Q1 and Q2 to rate their team members.

Differences exist in the performance of individuals across organizations and within teams. To reflect this performance range, a measurement tool should be able to produce unforced distribution on a continuum. In this study, team leaders had no preconceived expectations about how they should rate their team members. A natural "unforced" variation occurred.

Table 2: Descriptive Statistics

	N	Min.	Max.	Medium	Std. Deviation
Q1	371	1	5	3.66	.998
Q2	371	1	5	3.74	.964
Q3	371	0	1	.43	.496
Q4	371	0	1	.20	.400

3. Q1 and Q2 on the Performance Pulse are highly but not perfectly correlated.

As anticipated, a strong positive correlation exists between Q1 and Q2, and yet these correlations are not perfect. This signals that they are not measuring exactly the same judgments on the part of the team leaders. For example, it is possible for team leaders to *Strongly Agree* that their team members provide "extraordinary results," and yet not want to work with them "as much as I possibly can." From this study, we see the inverse as well: some team leaders *Strongly Agree* that they choose to work with a team member "as much as they possibly can" but do not count on that team member for "extraordinary results."

Table 3: Correlations

	Q1	Q2
Q1	1	.759(**)
Q2	.759(**)	1

*** Correlation is significant at the 0.01 level (2-tailed).*

4. A high score on Q1 or Q2 does not necessarily lead to a Yes on Q3 or a No on Q4.

While productivity (Q1) and teamwork (Q2) ratings may be highly correlated, a high rating on these items does not guarantee that a team leader feels a team member is ready for promotion (Q3). It is also true that team members given low ratings on productivity (Q1) and teamwork (Q2) do not necessarily have a performance problem (Q4).

Furthermore, we see team members who are high on Q1 and Q2 with identified performance problems (Q4) — according to these data, even high performers appear to have performance problems from time to time, which are often missed when performance is measured merely once a year.

Table 4: Correlations

	Q1	Q2	Q3	Q4
Q1	1	.759(**)	.605(**)	-.386(**)
Q2	.759(**)	1	.559(**)	-.355(**)
Q3	.605(**)	.559(**)	1	-.203(**)
Q4	-.386(**)	-.355(**)	-.203(**)	1

*** Correlation is significant at the 0.01 level (2-tailed).*

Conclusion

The Performance Pulse contains four questions: one for productivity, one for teamwork, one for promotion, and one for performance problems. Our research shows that the 70 team leaders who participated in this study were able to discriminate between high and low performers, those ready for promotion, and those who have performance problems needing to be addressed. These team leaders used the full range of the scale on each item regardless of what they recorded on any of the other items. Over time, these data can be aggregated across the organization, quarter by quarter, so that the organization can have better real-time information about what to do with each team member.

References

Ghorpade, J. (2000). Managing five paradoxes of 360-Degree feedback. *The Academy of Management Executive* (1993–2005), 14, 140–150.

Hoffman, B., Lance, C.E., Bynum, B., & Gentry, W.A. (2010). Rater source effects are alive and well after all. *Personnel Psychology*, 63, 119–151.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A.(1998). Trait, rater, and level effects in 360-degree performance ratings. *Personnel Psychology*, 51, 557–576.

Ng, K., Koh, C., Ang, S., Kennedy, J.C., & Chan, K. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology*, 96, 1033–1044.

Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 83, 956–970.

Westerman, J., & Rosse, J. (1997). Reducing the threat of rater nonparticipation in 360-degree feedback systems: An exploratory examination of antecedents to participation in upward ratings. *Group and Organization Management*, 22, 288–309.

THE
MARCUS
BUCKINGHAM
COMPANY