5-7-2019

Development and Use of Performance Composite Scores in Dispatch Teams

Christopher R. Bearden

Michael B. Hein

Glenn E. Littlepage

DEVELOPMENT AND USE OF PERFORMANCE COMPOSITE SCORES IN DISPATCH TEAMS

Christopher Ryan Bearden
Michael B. Hein
Glenn E. Littlepage
Middle Tennessee State University
Murfreesboro, TN

Teams perform a variety of functions within organizations and should therefore be evaluated on multiple criteria. This paper argues for the use of a single value. We review the literature on team performance composites and briefly describe two approaches to developing evaluative performance composites in an academic setting by combining performance indicator data: A qualitative approach for performance feedback as well as an empirical approach for research purposes.

Over ten years ago, Mathieu, Maynard, Rapp, and Gilson (2008) suggested that because teams perform multiple functions, a best practice for evaluating teams is to include and combine multiple criteria dimensions to evaluate teams. Theirs' is not the first nor the only call to combine criteria to appear in the research literature (cf. Pritchard, 1990; Salas, Rosen, Held, & Weissmuller, 2009). An argument for a single index of performance can be made on the basis of parsimony. Additionally, a single value can be easily compared across teams, have motivational value, and convey performance data quickly to organizational stakeholders and management (Pritchard, 1990). When evaluating teams, the general recommendations in the literature are clear: criteria should be theoretically-based (Salas, Burke, Fowlkes, & Priest, 2003); criteria measurement should be designed keeping in mind the functions of the team (Mathieu et al., 2008), as well as the purpose and environment of the team (Kendall & Salas, 2004) and the desired outcomes (Rosen, Wildman, Salas, & Rayne, 2012); finally, differentiated criteria should be combined using a formal method (Mathieu et al., 2008).

Literature Review

The argument for a composite measure (i.e., a single criterion variable) begins with the idea that teams need to be evaluated on multiple criteria because they perform a variety of functions (Mathieu et al., 2008). Typically, these dimensions are examined one-by-one; however, it can be difficult to assimilate multiple pieces of information about a team's functioning (Pritchard, 1990). A single value quickly conveys a large amount of information to organizations, researchers, and the teams themselves. Additionally, it provides an evaluative advantage, demonstrating change efforts and allowing for easy between-team comparisons. A composite has motivational value to teams because it clearly demonstrates consequences of effort (e.g., performance increases or decreases; Pritchard, 1990). Differentiated criteria can always be examined for specific reasons (e.g., planning improvements). Table 1 contains example studies to illustrate the creation of various team performance composites.

Many different reasons exist for evaluating a team, such as research, training evaluations, team performance or process diagnostics, or determining team rewards. The reason for evaluation should drive the decisions for selecting the criteria and indicators of the criteria

Table 1.
A Sample of Example Studies Combining Disparate Evaluation Criteria.

Reference	Label	Indicator Dimensions	Theoretical Justification	Procedure
Hiller, Day, & Vance (2006)	Team Effectiveness	Planning and organizing (task) Problem solving (task) Support and consideration (relationship) Development and mentoring (relationship)	All four dimensions adapted from Managerial Practices Survey (Yukl & Lepsinger, 1990). For task versus relationship see Judge, Piccolo, and Ilies (2004)	1, "very ineffective" to 7, "very effective;" averaged; squared to reduce skew.
		Overall effectiveness	None	
Van Der Vegt & Bunderson (2005)	Team Performance	Efficiency Quality Overall achievement	Performance criteria (Ancona & Caldwell, 1992)	Item rated 1, "far below average" to 7, "far above average;" averaged
		Productivity Mission Fulfillment Overall team performance	None; suggested by company	
Mathieu, Gilson, & Ruddy (2006)	Quantitative Performance	Machine reliability Response time Parts expense	(Hyatt & Ruddy, 1997)	Converted to *z* scores; used as indicators for a latent variable
Komaki, Barwick, & Scott (1978)	Safety Performance	Department-specific safety item(s)	Analysis of archival incident reports	Behavior observation scale; score is the ratio of safe actions to total observed
Pearsall & Ellis (2006)	Team Performance	Offensive score Defensive score	(Ellis, Hollenbeck, Ilgen, Porter, West, & Moon, 2003)	Offensive and defensive scores were standardized and summed

(Salas, Reyes, & Woods, 2017). This ensures outcome-measurement congruence. That is to say, measures should capture what is needed to make the generalizations and draw the conclusions needed. Evaluators should not only rely on theory for criteria inclusion but for the justifying the composite itself as well. Choosing a theoretical model, such as Hackman's (1987) team effectiveness framework, will not only guide the measurement strategy but also provide conceptual clarity and lend credence to the approach (Salas et al., 2003). First, a systematic team task analysis should be conducted. Organizational leadership should be consulted for evaluation criteria during this process (Sundstrom, De Meuse, & Futrell, 1990). Consideration should also be given to the behavioral processes and performance criteria identified, defined, and organized through previously taxonomic efforts (e.g., Marks, Mathieu, & Zaccaro, 2001).

A criterion represents an objective or desired outcome, or product or service rendered. Each criteria included in the composite must have at least one indicator, and each indicator must be measurable/measured. When designing the measurement strategy, consideration should be given to the function, purpose, and environment of the team (Kendall & Salas, 2004). Pritchard (1990) makes several recommendations when selecting indicators: Indicators should meaningful to both the purpose of the evaluation and the intended audience; the long-term consequences of improving on the indicators should be considered; the indicators should be under the control of the team; and the indicator should not be contaminated by other units' performance. Additionally, indicators should not be selected if they do not vary between teams.

Indicators can be categorized as objective (e.g., points scored in a simulation game) or subjective (e.g., supervisor judgements). Meta-analytic findings have demonstrated the convergent validity of objective and subjective measures of performance (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995). Consideration should also be given to whether the indicators will be behavior-focused (esp. for training or feedback or rewards) *or* outcome-based. Kozlowski and Bell (2013) argue that team performance itself should be conceptualized as the action(s) the team takes as opposed to the outcomes, which is consistent with the distinction made by Beal, Cohen, Burke, and McLendon (2003) who argue that performance behaviors should be separate from performance outcomes. All decisions should be guided by the purpose of the evaluation. Finally, each indicator should represent the team as a whole not an individual.

Measures of objective outcomes have several advantages, such as possible automatization of data collection, and are also often the most intrinsically interesting to stakeholders. Teams, however, may not be able to control certain outcomes to the same extent they can control their own intrateam processes and behaviors. Subjective indicators are more widely used, in part, because data collection methods are typically easier to design and access. However, subjective measures have their own problems. For one, it can take numerous evaluators to effectively observe a team's performance. Ratings provided by evaluators can also be biased. If subjective indicators are used, raters should receive training and only assess four to five indictors of performance (Smith-Jentsch, Baker, Salas, & Cannon-Bowers, 2001). Ways to avoid this limitation include having raters assess only those indicators with which they are most familiar and increasing the rater pool to include self-report, peers, experts, and supervisors. This could lead to other issues, such as the inability to determine needed interrater agreement on indicators.

After data collection factor analytic methods (e.g., confirmatory factor analysis) can be used to establish construct reliability and provide evidence for construct validity. Often, highly correlated indicators ($r > .70$) are simply averaged. The problem with averaging indicators is that it assumes linearity. In other words, gains in the raw score at any level on any indicator contribute equally to the overall performance. Further, it assumes that deficiencies in one area of performance can be compensated for in another – which is not always the case in applied settings. Statistical methods (e.g., principal components analysis) can be used to inform how or whether to combine data on multiple indicators after the data are collected. Before being combined, indicators can and should be weighted based on their relative value. Weights can be determined through judgement or statistical methods. While this helps, weighting indicators does not solve the problem. Two examples are hereafter provided which do solve this problem.

Two Example Approaches for Developing a Team Performance Composite

Six objective indicators were identified for our team-training simulation of a regional flight dispatch center: number of flights dispatched, number of airline policy violations, total delay time, number of passengers missing connections, pounds of undelivered cargo, and number of airplanes with a tarmac delay fine. Two approaches were undertaken to combine data on the indicators. The first utilizes recommendations from the Productivity Measurement and Enhancement System (ProMES; Pritchard, 1990) to create a composite that provides students with actionable feedback and an ability to set goals. The second utilizes principal components analysis to maximize team differences on the indicators. Both approaches address the non-linear relationships among the indicators and with the criterion.

The ProMES recommends establishing three values for each indicator: a maximum value on the indicator, a lowest possible value, and an expected value on the indicator. See Table 2 for an example using airline policy violations. A raw score of zero represents the best performance. The purpose of this first composite is performance feedback for teams early in training.

Table 2.
Example using Archival Team Performance Data to Determine Indicator Values

		Indicator Value Label											
		Max.	Good		Expected			Poor				Min.	
Raw Score		0	1	2	3	4	5	6	7	8	9	10	11
Teams	*N*	5	9	6	6	3	4	3	1	0	1	3	1
	%	12	21	14	14	7	10	7	2	0	2	7	2

Therefore, using historical data, similar raw scores were grouped in such a manner that *most* teams would perform well or at least as expected (i.e., a negative skew). Some room was left for improvement. Similar values can be grouped considering the reliability of the measurement instrument or can be established using judgement. The points earned for *good* performance versus *expected* versus poor (etc.) can be established using a variety of ways. Here, maximum (Max.) performance is awarded an "A" or a 4.0 out 4.0 points, which uses a frame of reference with which undergraduate aerospace students are familiar. After new values are assigned to the raw scores, the indicator itself (i.e., the 4.0) is weighted relative to its contribution to overall effectiveness. Since policy violations are related to airline safety, and safety is our virtual airlines' number one priority, this indicator is weighted as the most important. This process is repeated for each indicator.

The second approach uses non-linear principal components analysis (PCA; see Linting, Meulman, Groenen, & van der Koojj, 2007). Non-linear PCA is suitable for all measurement levels, so indicators could be ordinal, nominal, numeric, or any combination thereof. Non-linear PCA reproduces more variance than traditional PCA – even in the unlikely cases where the relationships are linear. The second composite is to be used as a criterion for research purposes, therefore maximizing the differences between teams on the indicators is useful. The first step is to determine the overall number of components. Typically, this is one, but more may be needed. Second, the indicators are rescaled to account for non-linearity (this can be accomplished using the PRINQUAL function in SAS or SPSS's optimal scaling function). Third, a PCA is conducted on the rescaled indicator variables. The component score(s) produced maximizes the differences between teams while accounting for the non-linear relationships. Limitations of this approach include the need for large amounts of historical data, increased complexity of interpretations, and automation requires sophisticated information technology skills complex.

Conclusion

Several clear recommendations should aid in the design and interpretation of performance composites. Teams should be evaluated on multiple dimensions, which cover their functions and purpose (i.e., content validity; Mathieu et al., 2008); the specific criteria selected should fit within a theoretical framework (i.e., construct validity; Salas et al., 2003); the criteria must match outcomes (i.e., criterion relevance and criterion validity; Sundstrom et al., 1990). Each criterion must have at least one measurable, controllable, and uncontaminated indicator. When measuring behaviors and processes, these can be split into task and relationship (Judge, Piccolo, & Ilies, 2004), which could add more conceptual clarity and aid in interpretation. Indicators should be carefully combined using a formally articulated method. Methods of combining indicator data should account for the (potential) non-linear relationships among indictors and between the indicators and the evaluative criterion.

References

Beal, D. J., Cohen, R. R., Burke, M. J., & McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology*, *88*(6), 989. DOI: 10.1037/0021-9010.88.6.989

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel psychology*, *48*(3), 587-605. DOI: 10.1111/j.1744-6570.1995.tb01772.x

Hackman, J. R. (1987). *The design of work teams*. In J. Lorsch (Ed.), Handbook of organizational behavior (pp. 315–342). New York, NY, USA: Prentice Hall.

Judge, T. A., Piccolo, R. F., & Ilies, R. (2004). The forgotten ones? The validity of consideration and initiating structure in leadership research. *Journal of Applied Psychology*, *89*(1), 36. DOI: 10.1037/0021-9010.89.1.36

Kendall, D. L., & Salas, E. (2004). Measuring team performance: Review of current methods and consideration of future needs. In J. W. Ness, V. Tepe, & D. R. Ritzer (Eds.) *The* *science and simulation of human performance*, *Vol. 5* (pp. 307-326). Emerald Group Publishing Limited.

Kozlowski, S. W. J., & Bell, B. S. (2013). *Work groups and teams in organizations: Review update* [Electronic version]. Retrieved from Cornell University, School of Industrial and Labor Relations site: http://digitalcommons.ilr.cornell.edu/articles/927

Linting, M., Meulman, J. J., Groenen, P. J., & van der Koojj, A. J. (2007). Nonlinear principal components analysis: introduction and application. *Psychological Methods*, *12*(3), 336. DOI: 10.1037/1082-989X.12.3.336

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356-376.

Mathieu, J., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of Management*, *34*(3), 410-476. DOI: 10.1177/0149206308316061

Pritchard, R. D. (1990). *Measuring and improving organizational productivity: A practical guide*. Westport, CT, USA: Greenwood Publishing Group.

Rosen, M. A., Wildman, J. L., Salas, E., & Rayne, S. (2012). Measuring team dynamics in the wild. In A. Hollingshead & M. S. Poole (Eds.), *Research methods for studying groups: A guide to approaches, tools, and technologies* (pp. 386–417). New York, NY, USA: Taylor & Francis.

Salas, E., Burke, C. S., Fowlkes, J. E., & Priest, H. A. (2004). On measuring teamwork skills. In J. C. Thomas (Ed.), *Comprehensive handbook of psychological assessment, Vol. 4. Industrial and organizational assessment* (pp. 427-442). Hoboken, NJ, US: John Wiley & Sons Inc.

Salas, E., Reyes, D. L., & Woods, A. L. (2017). The Assessment of Team Performance: Observations and Needs. In *Innovative Assessment of Collaboration* (pp. 21-36). Springer, Cham.

Salas, E., Rosen, M. A., Held, J. D., & Weissmuller, J. J. (2009). Performance measurement in simulation-based training: A review and best practices. *Simulation & Gaming*, *40*(3), 328-376. DOI: 10.1177/1046878108326734

Smith-Jentsch, K. A., Baker, D. P., Salas, E., & Cannon-Bowers, J. A. (2001). Uncovering differences in team competency requirements: The case of air traffic control teams. In E. Salas, C. A. Bowers, & E. Edens (Eds.), *Improving teamwork in organizations: Applications of resource management training* (pp. 31-54). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Sundstrom, E., De Meuse, K. P., & Futrell, D. (1990). Work teams: Applications and effectiveness. *American Psychologist*, *45*(2), 120. DOI: 10.1037/0003-066X.45.2.120