Conceptualization and Measurement of Team Workload: A Critical Need

Gregory J. Funke and Benjamin A. Knott, Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio, Eduardo Salas and Davin Pavlas, University of Central Florida, Orlando, Florida, and Adam J. Strang, Consortium Research Fellows Program, Alexandria, Virginia

Objective: The purpose of this article is to present and expand on current theories and measurement techniques for assessing team workload.

Background: To date, little research has been conducted on the workload experienced by teams. A validated theory describing team workload, which includes an account of its relation to individual workload, has not been articulated.

Method: The authors review several theoretical approaches to team workload. Within the team research literature, attempts to evaluate team workload have typically relied on measures of individual workload. This assumes that such measures retain their validity at the team level of measurement, but empirical research suggests that this method may lack sensitivity to the drivers of team workload.

Results: On the basis of these reviews, the authors advance suggestions concerning a comprehensive theory of team workload and methods for assessing it in team settings. The approaches reviewed include subjective, performance, physiological, and strategy shift measures. Theoretical and statistical difficulties associated with aggregating individual-level workload responses to a team-level measure are discussed.

Conclusion: Conception and measurement of team workload have not significantly matured alongside developments in individual workload.

Application: Team workload remains a complex research area without simple measurement solutions, but as a research domain it remains open for contributions from interested and enterprising researchers.

Keywords: team workload, measurement, subjective measures, performance measures, physiological measures, strategy shift measures

INTRODUCTION

The mental workload experienced by an individual operator performing a task is a vital area of research, informing our understanding and prediction of human performance in complex systems (Parasuraman, Sheridan, & Wickens, 2008). As such, the domain has received significant scientific inquiry, particularly within the past half century (Tsang & Vidulich, 2006). The result has been a proliferation of theories, methods, and metrics designed to evaluate individual workload (e.g., Hancock, Mihaly, Rahimi, & Meshkati, 1988; Tsang & Vidulich, 2006).

By contrast, research on team workload has not matured at the same rate (Bowers & Jentsch, 2005), even though teams have been acknowledged as an integral part of the military, industrial, medical, and public service sectors (e.g., Bowers, Braun, & Morgan, 1997; Salas, Burke, & Goodwin, 2009). A comprehensive, validated theoretical framework for the construct team workload, which includes a description of its relation to individual workload, has not been articulated. Although some interesting work has been initiated (e.g., Bowers et al., 1997; Funke, Knott, Galster, & Brown, 2009), progress in the area has been slow, perhaps because of difficulties defining appropriate team workload dimensions and challenges associated with measurement in team settings.

The purpose of this article is to present and expand on current theories and measurement techniques for assessing team workload—a critical need for optimal human–system integration. We hope this discussion will provide
impetus for more thinking, theorizing, debate, dialogue, and empirical work in this area. In addition, this review serves to identify remaining weaknesses in the literature and provide recommendations for future team workload research.

CURRENT CONCEPTUALIZATIONS OF TEAM WORKLOAD

At the individual level, workload has been proposed as an inferred construct that mediates the relationship of task difficulty, operator skill, and observed task performance (e.g., Gopher & Donchin, 1986; Hart & Staveland, 1988; Moray, 1979). Frequently, this relationship is described using a resource model (e.g., Tsang & Vidulich, 2006; Wickens & Hollands, 2000). Attentional resource theories (e.g., Norman & Bobrow, 1975) posit that information processing and task performance are dependent on the availability of system resources. Typically, resource-based accounts of workload assert that resources exist in a fixed quantity—though there is some evidence to suggest that this quantity may vary with task demands and other factors (e.g., Young & Stanton, 2002a, 2002b), that resources are necessary for task-related information processing, and that resources are expended during task performance (e.g., Norman & Bobrow, 1975). From this perspective, as task demands increase, more effort is required to perform a task, and as a result more resources are expended during performance and workload increases. If available resources are insufficient to meet the demands of the task, performance should be impaired—though skilled operators may adjust their strategy to compensate (e.g., Singleton, 1989). This conceptualization of workload has been quite successful, receiving converging support from subjective, performance, and neurophysiological measures (see, e.g., Tsang & Vidulich, 2006, for a review).

At the team level, conceptualizations of workload have been directly adapted from theories of individual workload. For example, Bowers et al. (1997, p. 90) describe team workload as “the relationship between the finite performance capacities of a team and the demands placed on the team by its performance environment,” and Bowers and Jentsch (2005, p. 57-1) as “an index of the ratio of available team resources to task demands.” From these perspectives, team performance will be maximized (all other factors being equal) when task demands are in reasonable balance with a team’s capabilities and resources. As is the case for individual workload, when demands exceed team resources, teams may alter their performance strategies to compensate (e.g., by changing coordination strategies—Entin & Serfaty, 1999; by reducing team communication—Urban, Weaver, Bowers, & Rhodenizer, 1996; etc.), or task performance may decline and team workload should increase.

Distributed cognition (Hollan, Hutchins, & Kirsh, 2000) research suggests that team workload is a function not only of team interactions but also of the team’s relationship to the cognitive prostheses and tools that the team works with and is embedded in. When combined with the emerging field of team cognition (Salas & Fiore, 2004), these efforts highlight the complexity of team research. At this point it is useful to distinguish, at a relatively high level of abstraction, the sources of demands within a team setting. Several authors (e.g., Bowers et al., 1997; Glickman et al., 1987; Morgan, Salas, & Glickman, 1994) have suggested that team tasks, examined from the perspective of individual team members, are similar to dual-task performance situations. According to Bowers et al. (1997), team members must effectively balance taskwork, which broadly corresponds to individual efforts to meet task demands, including efforts traditionally associated with individual task performance, and teamwork, which reflects the cooperative efforts of team members required for task performance (i.e., communication, coordination, monitoring, leadership, etc.). Team workload, then, may be expressed in terms of the demands associated with performance of taskwork and teamwork.

Returning to the concept of team workload, few authors examining workload in team settings have explicitly articulated the relationship of individual workload to team workload. In the few instances where a theoretical orientation has been expressed, it has been suggested that team workload is an emergent property of team performance of team tasks. Furthermore, these authors suggest that the relationship of individual and team workload may be complex (i.e.,
multiplicative or nonlinear). For example, Bowers et al. (1997) state,

[I]t is also reasonable to suggest that team workload differs from individual workload by virtue of the complex interactions among the requirements for taskwork and teamwork. That is, although each team member experiences some degree of individual workload, it is hypothesized that—because of the required time sharing of taskwork and teamwork processes—the team experiences overall demands that go beyond the sum of the workloads of individual members. (p. 91)

Similarly, Bowers and Jentsch (2005) argue,

[T]he overall workload capacity of a team is almost certainly not the sum of the task-related resources of the individual members. Rather, each member added to a team brings in not only additional resources, but also represents a process cost. Thus, the goal for a measure of team workload is to estimate the capacity of a team, bearing in mind the workload associated with the team processes required to perform the overall team task. (p. 57-1)

In summary, team workload has been defined as an emergent property of teams, characterized by a nonadditive relationship between the finite performance capacities of a team and the taskwork and teamwork demands placed on the team by its performance environment.

However, the preceding model is not the only possible conceptualization of team workload. From the perspective of parsimony, an additive model should be preferred to the more complex, nonadditive model described earlier. In addition, when specified as an emergent property of teams, measures of team workload will not reflect the distribution of task demands (taskwork or teamwork) across team members, which may be important for diagnostic reasons (e.g., identifying potential sources of errors or excessive demands in a team, etc.). Furthermore, task relevant characteristics (e.g., skills, behaviors, perceptions, etc.) and resources are rarely “equally distributed” among all members of a team (hence the domain of individual differences). Because of the unequal distribution of these factors, to understand team workload, it may be necessary to recognize and retain individual workload and individual performance in our conceptualization. From this perspective, team workload may be characterized as a hypothetical construct that represents the linear aggregate cost incurred by all members of a team of operators to achieve a particular level of team performance, which reflects the interactions of taskwork and teamwork demands and relevant individual characteristics of all team members.

A final view of team workload posited by Kozlowski and Klein (2000) suggests that the relationship of individual and team workload is specified by the task performed and the team’s products that result from performance of those tasks (i.e., how a team achieves the “outcomes” of team processes, by which teams are evaluated to determine whether or not they have met task goals). A common theme of several taxonomies describing team processes and products (e.g., Bell & Kozlowski, 2002; McGrath, 1984; Steiner, 1972; Van de Ven, Delbecq, & Koenig, 1976) is the degree of interdependence required from team members during task performance. If team tasks are viewed on a continuum of interdependence from low to high, team workload in tasks that involve little interdependence among team members may be best characterized using an additive model (Van de Ven et al., 1976), whereas tasks that require intense interdependence may be better represented by multiplicative or nonlinear relationships (Van de Ven et al., 1976).

It should be noted, however, that the definitions outlined above are not without flaws. A primary flaw (if not the primary flaw) of relevance to the current discussion is that the same components of the models used to estimate individual workload are involved in estimates of team workload. In other words, based on the models outlined, estimates of team workload and individual workload will not be independent (in fact, they should covary, perhaps strongly). In addition, as Bolia and Nelson (2007) suggest, the relationship of individual and team workload need not be fixed—it may
change as the situation and team evolve. To date, none of the approaches to team workload discussed herein have been subjected to rigorous scientific inquiry, suggesting that experiments directly comparing the model predictions of each are warranted.

**MEASUREMENT OF TEAM WORKLOAD**

Within the individual workload literature, several authors have detailed workload taxonomies enumerating categories of workload assessments. For example, Wierwille and Eggenmeyer (1993; also see, e.g., Tsang & Vidulich, 2006; Wickens & Hollands, 2000) describe three "types" of workload measures: subjective measures (e.g., self-report questionnaires), performance measures (e.g., primary and secondary task outcomes), and physiological measures (hereafter referred to as physio-behavioral measures to include both physiological measures, such as galvanic skin response and heart rate variability, and behavioral measures, such as ocular fixations). An additional interesting measure of workload, suggested by Serfaty, Entin, and Deckert (1993; see also Entin & Serfaty, 1999), may be team-level strategy shifts. These categorizations are conceptually useful in describing potential measures of team workload, and our discussion proceeds as such. A short description of the four categories and a brief assessment of the advantages and disadvantages of each for the assessment of team workload are presented in Table 1. It is worth noting that, in practice, most experimental investigations of team workload have relied on subjective and performance measures.

**Subjective Measures**

For the most part, the approach taken by researchers thus far has been to assess team workload using an existing measure of individual workload, for example, the NASA Task Load Index (TLX; Hart & Staveland, 1988), that has been modified by altering the instruction set, item set, or both to make them more applicable to teams (e.g., Bowers & Jentsch, 2005). This may be in light of the theoretical connection between individual and team workload, or it may stem from more pragmatic concerns (e.g., availability of measures). Typically, team members are asked to rate their individual workload and then researchers aggregate the responses of all team members as a global team workload score; commonly, this aggregation takes the form of a team average (Bowers & Jentsch, 2005).

Presented in Table 2 is a list of studies that have featured attempts to assess team workload using self-report questionnaires. Studies were obtained through computed-based searches for the keywords team and workload in Google Scholar, PsycINFO, and Science Citation Index. Only those articles that described using a subjective assessment of team workload were selected for presentation in the table (as such, articles that discussed team workload without directly assessing it were omitted, as were articles that included manipulations of team-level task demands without a concomitant team-level measure of subjective workload).

As can be seen in the table, the bulk of experiments included have utilized military command and control tasks and the NASA-TLX for individual workload assessment. It is also apparent that a handful of unique measures have been developed and applied to study team workload.

For example, Hildebrand, Pharmer, and Weaver (2003; also see Pharmer, Cropper, McKneely, & Williams, 2004) assessed team workload using a modified TLX that combined the standard six rating scales of the NASA-TLX with an additional four subscales. The supplementary items addressed communication demand, monitoring demand, control demand, and coordination demand and were scored on a scale that ranged from 1 (low) to 100 (high; also see, e.g., Knott, Bolia, Nelson, & Galster, 2006, for a further modification of this team measure).

Schwartz (2008) assessed team workload using a consensus-scored version of the NASA-TLX in a process similar to that proposed by Dickinson and McIntyre (1997). Team members worked collaboratively to assign a score reflecting the team’s overall perceived workload on each of the six dimensions of the NASA-TLX. Participants were told to discuss each trial in conjunction with the six rating scales and to assign a single value to each scale.
**TABLE 1: Categories, Descriptions, and Relative Advantages and Disadvantages of Team Workload Assessment Methodologies**

<table>
<thead>
<tr>
<th>Measure Assessment Categories</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective</td>
<td>Self-report questionnaires that request respondents to estimate their workload or the workload of the team (or both) on dimensions relevant to team task performance. Frequently paired with an existing measure of individual workload.</td>
<td>Most team workload research has utilized this approach; consequently, it includes the largest selection of established measurement options.</td>
<td>Existing measures have not been sufficiently validated. Questions of measure sensitivity are a concern.</td>
</tr>
<tr>
<td>Performance</td>
<td>Monitoring team processes and outcomes for changes occurring in conjunction with manipulation of task demands. Both primary and secondary task measures may be employed.</td>
<td>Frequently included in experiments examining task demand effects on team performance. Important for establishing converging evidence in any such investigation.</td>
<td>Changes in performance may not reflect changes in team resource investments, and it may be difficult to diagnose the source of load (taskwork, teamwork) when a change is detected.</td>
</tr>
<tr>
<td>Physio-behavioral</td>
<td>Physiological- and behavioral-based metrics that are assumed to be correlates of mental work.</td>
<td>Many of these measures can be assessed in real time from multiple participants, providing temporally sensitive assessments of workload during task performance.</td>
<td>Measurement devices may interfere with normal team processes. Establishing baselines may be difficult and time-consuming. It is unclear how to aggregate signals into team-level measures.</td>
</tr>
<tr>
<td>Strategy shift</td>
<td>Indirect indicators of team workload revealed by qualitative and quantitative transitions in team behavior with manipulation of task demands.</td>
<td>Measures could be utilized in near real time without disrupting normal team processes. Emerging analytic techniques may provide additional metrics for assessing team workload.</td>
<td>Methods for assessing team workload have not been fully conceptualized. It may be difficult to diagnose the source of load (taskwork, teamwork) when a strategy shift is detected.</td>
</tr>
<tr>
<td>Team Task</td>
<td>Individual Workload Measure</td>
<td>Team Workload Measure</td>
<td>Sensitive to Manipulation of Task Difficulty?</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------</td>
<td>-----------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>Bailey and Willems (2002)</td>
<td>Simulated air traffic control</td>
<td>NASA-TLX</td>
<td>Global TLX score averaged across team members</td>
</tr>
<tr>
<td>Beith (1987)</td>
<td>Problem solving/pattern recognition</td>
<td>NASA Bipolar Scale</td>
<td>1 question asking participants to estimate the team's workload on a 9-point Likert-type scale</td>
</tr>
<tr>
<td>Bowers, Urban, and Morgan (1992)</td>
<td>Team Performance Assessment Battery</td>
<td>NASA-TLX</td>
<td>Global TLX score averaged across team members</td>
</tr>
<tr>
<td>Funke, Bennett, Nelson, and Galster (2007)</td>
<td>Military C2</td>
<td>NASA-TLX</td>
<td>MT-TLX</td>
</tr>
<tr>
<td>Funke and Galster (2009)</td>
<td>Control of multiple UAVs</td>
<td>NASA-TLX</td>
<td>Global TLX score averaged across team members</td>
</tr>
<tr>
<td>Lin, Hsieh, Tsai, Yang, and Yenn (2011)</td>
<td>Nuclear reactor regulation</td>
<td>NASA-TLX</td>
<td>TWA</td>
</tr>
<tr>
<td>Schwartz (2008)</td>
<td>Military C2</td>
<td>NASA-TLX</td>
<td>Team consensus-scored TLX</td>
</tr>
<tr>
<td>Sebok (2000)</td>
<td>Nuclear reactor regulation</td>
<td>NASA-TLX</td>
<td>Global TLX score averaged across team members</td>
</tr>
<tr>
<td>Shi, Luh, Kleinman, and Serfaty (1991)</td>
<td>Military C2</td>
<td>SWAT</td>
<td>SWAT score averaged across team members</td>
</tr>
</tbody>
</table>
that best reflected the team’s workload. Consensus TLX subscales were scored in the same fashion as those of the NASA-TLX (i.e., 0–100). It is worth acknowledging that this style of team workload assessment requires additional collaboration between team members and is likely to be affected by team and interpersonal dynamics (e.g., team cohesion, efficacy, conflict, etc.; Kozlowski & Ilgen, 2006). Interested readers are directed to Kirkman, Tesluk, and Rosen (2001) for a review of the advantages and disadvantages of aggregation and consensus scoring methods for assessing team-level constructs.

Entin, Serfaty, and Kerrigan (1998; also see Entin & Entin, 2001; MacMillan, Paley, Entin, & Entin, 2005) employed the Team Workload Awareness Questionnaire (TWAQ), a complex, multipart questionnaire designed to assess individual and team workload. Participants used a modified NASA-TLX to rate their own workload, the workload of each of their teammates, and the workload of the team as a whole. Subscales of the TWAQ were scored in the same fashion as those of the NASA-TLX. As the title of the questionnaire suggests, measurements made with the TWAQ may be substantially influenced by team awareness and metacognition, as each team member must be cognizant of other members’ workloads to accurately rate them. In addition, utilizing what is essentially the same measure for assessment at the individual and team levels may introduce bias into participant responses, as raters implicitly or explicitly attempt to coordinate their ratings on each section of the questionnaire.

As an alternative to posttrial subjective measures of workload, the Task Manager unidimensional workload measure (Van Orden, 2001) may offer some advantages for assessment of team workload. The measure requires participants to rate their workload on a scale from 1 to 7 every 2 to 3 min. Although the scale does not provide the full diagnosticity of a multidimensional measure of workload in the sense of ascribing ratings to particular sources or types of demands, the temporal profile provided by the Task Manager measure could allow researchers to characterize changes in workload across team members during task performance. When correlated with task and team events, such as instances of communication or backup behavior, Task Manager profiles may provide further insight into the sources of taskwork and teamwork in a team environment.

Collectively, the previously described methods of team workload assessment rely on modified individual measures of workload, which

<table>
<thead>
<tr>
<th>Table 2. (continued)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team Task</td>
</tr>
<tr>
<td>Throne et al. (1999)</td>
</tr>
<tr>
<td>Yang, Yenn, and Lin (2010)</td>
</tr>
</tbody>
</table>

Note. TLX = Task Load Index; N/A = not applicable (i.e., the experiment did not include a clear manipulation of task difficulty); military C2 = military command and control; TWAQ = Team Workload Awareness Questionnaire (Entin, Serfaty, & Kerrigan, 1998); MT-TLX = Modified Team Task Load Index (Knott et al., 2006); UAV = unmanned aerial vehicle; TWAS = Team Workload Assessment Scale (Galster & Knott, 2007); T-TLX = Team Task Load Index (Hildebrand et al., 2003); TWA = Team Workload Assessment Measure (Lin et al., 2011); SWAT = subjective workload assessment technique (Reid & Nygren, 1988); C3 + C2 Scale = Communication, Coordination, Cooperation, Plus Command and Control Scale (Yang et al., 2010).
Conceptualization and Measurement of Team Workload

presents several difficulties. First, consideration must be given to how the data from a team of individuals should be combined and interpreted (Bowers et al., 1997). Several different methods have been advanced, including the sum or average workload of all team members, the lowest workload value obtained, and the highest workload value obtained (Bowers & Jentsch, 2005). Experimentally, the results of several studies, all using modified versions of the TLX, provide mixed support for those approaches. Using the Team Performance Assessment Battery, a multicomponent team task, Bowers, Urban, and Morgan (1992) found the highest and lowest individual ratings, as well as the averaged rating, to be significantly correlated with team performance.

Conversely, the results of a flight simulation study by Thornton, Braun, Bowers, and Morgan (1992) revealed no significant relationship between performance and any of the preceding workload indices. Presented in Table 2 is an indication of whether or not the measure of team workload utilized in each experiment was sensitive to a manipulation of task difficulty (team performance was sensitive to task difficulty in all presented cases). From the table, it appears that in five studies the measure was found to be sensitive, and in the remaining four studies the measure displayed mixed or no sensitivity. This suggests that current measures may not be sensitive enough to the sources of team demands to adequately capture team workload.

Second, it is perhaps unreasonable to presume that the psychometric properties of existing measures, determined at the level of the individual, will be unchanged by transitioning to a team level of analysis. The authors are not aware of any research to date specifically examining the psychometric properties of an individual workload measure that has been extended to the team level by aggregation. At best, the aggregate and novel team workload measures described previously have received only limited experimental validation and as such should be employed with caution.

However, the results of two recent studies suggest that researchers investigating team processes and performance may still wish to consider inclusion of a team workload measure in their experiments. Research by Funke, Knott, et al. (2009) and by Lin, Hsieh, Tsai, Yang, and Yenn (2011) examined the incremental increases in total variance explained in team performance by the inclusion of measures of individual and team workload. The results of both studies indicated that the addition of a measure of team workload resulted in small to moderate increases, on the order of 5% to 20%, in the variance accounted for in their models above what was accounted for by a measure of individual workload. This suggests that combining an existing measure of individual workload with an existing measure of team workload, such as the TWAQ or Modified Team Task Load Index, may be a relatively viable approach. Continuing research will be necessary to refine the sensitivity of these measures, to ascertain an appropriate means for aggregating team member data, and to determine the degree of overlap in variance accounted for by each measure.

Alternatively, interested researchers could develop a novel instrument and validate it at individual and team levels of analysis. Such a measure could have several advantages including simplicity of administration, sensitivity and diagnosticity at both levels of assessment, and differentiability of taskwork and teamwork demand sources.

Performance Measures

Performance measures attempt to capture the degree of success in achieving primary (e.g., a pilot maintaining the intended course of an aircraft) and/or secondary (e.g., a pilot holding a conversation with a subordinate while flying) task goals (Tsang & Vidulich, 2006). In team research, performance measures can be quite diverse, ranging from objective (e.g., prosecution time in a tactical air defense task—Strang et al., 2011; maze completion time—Henning, Boucsein, & Gil, 2001) to subjective measures (e.g., mission success rated by a panel of expert judges—Brannick, Prince, Prince, & Salas, 1995).

Although the relationship between workload and primary task performance is thought to be clear (i.e., an increase in workload is expected to result in a decrease in primary task performance, and vice versa; Tsang & Vidulich, 2006), it has been argued that secondary task performance is a better indicator of mental
workload (Vidulich, 2003). As described previously, resource-based accounts of individual workload (e.g., Wickens & Hollands, 2000) posit a negative relationship between task demands and performance; that is, as task demands increase, more cognitive resources are expended during performance, resulting in deficits if available resources are insufficient to meet demands. From this, it has been proposed that secondary task performance, or a change in primary task performance when a secondary task is introduced (Shingledecker, 1987), indirectly indicates the degree of operator mental workload by assessing the amount of, or simply the existence of, spare mental capacity following the meeting of primary task objectives (Tsang & Vidulich, 2006).

With regard to team research, although experimental manipulations of primary task difficulty are relatively common, changes in primary task performance have not been used as direct indicators of team workload. Although shifts in primary task team performance in response to changes in task demands are frequently discussed in terms of “team workload,” this reference term is often given without thoughtful consideration of the construct it represents.

Moreover, primary and secondary performance measures used to assess team workload are difficult to interpret for the reasons enumerated by Gopher and Donchin (1986, p. 41-25) regarding individual workload, namely, that performance measures do not adequately reflect variations in the resource investments of teams and team members because of changes in task difficulty. This means that performance measures may not allow diagnosis of the source of the load (taskwork, teamwork, or some combination thereof) or permit a systematic conversion of performance units into measures of the relative demands or load on the team and team members. Because of the complex nature of the processes underlying team performance, it is difficult to ascribe changes observed in performance after manipulation of task demands only to team workload—equally valid explanations could include differences in team mental models (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000), reductions in individual and team situation awareness (Bolstad & Endsley, 1999), insufficient or incomplete training (Cannon-Bowers, Salas, Blickensderfer, & Bowers, 1998), and so on. Although it is certain that manipulations of primary and secondary task demands will continue to be an important part of team research, their utility for assessment of team workload is debatable.

**Physio-Behavioral Measures**

Physio-behavioral measures such as echocardiography (ECG), eye tracking, electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and transcranial doppler ultrasonography (TCD) have been relatively successful as indices of individual workload—though this viewpoint is somewhat contentious in light of research identifying dissociations between and among physio-behavioral and subjective workload measures (see, e.g., Tsang & Vidulich, 2006, for a recent review). Occasionally, simultaneous physio-behavioral measurement of multiple team members has been attempted in team settings, for example in assessing the workload of flight crews during aircraft testing and evaluation (Bonner & Wilson, 2002) and ship bridge teams under several team configurations (Murai, Okazaki, & Hayashi, 2004). However, the focus of such research has been on categorizing the individual workload of team members rather than assessing the team’s workload.

Several reasons for this situation suggest themselves. First, some of these measures may interfere with, or disrupt, team process that are critical to team tasks such as communication (e.g., speech during EEG and TCD measurement introduces artifacts) and movement (e.g., eye trackers, most brain imaging techniques, and any recording equipment with cables that “tether” the participant to a machine restrict movement inherent to many cooperative team tasks). To some extent, these problems may be overcome by the widening array of commercially available wireless monitoring devices, but this is not a matter that seems likely to be completely overcome soon.

Second, most physio-behavioral processes require baseline estimates to draw inferential comparisons (e.g., if a researcher wants to demonstrate that a manipulation of task demands resulted in an increase in heart rate variability,
baseline variability estimates for each team member are required; Stern, Ray, & Quigley, 2001). However, establishing an appropriate baseline for physio-behavioral responses is not a simple process, as the quality of those estimates is influenced by the state and condition under which the baseline is observed. For example, Fishel, Muth, and Hoover (2007) demonstrated that detection of arousal state changes in ECG heart rate variability (HRV) may be biased by the state of the operator during baseline assessment; high arousal states are easier to detect when HRV baseline estimates are recorded from relaxed operators, and vice versa. Consequently, it has been suggested that the “average” physio-behavioral response observed under a range of task conditions (e.g., average EEG theta response observed across relaxed, active, bored, and stress states) might be the best (i.e., least biased) comparator in experiments examining positive and negative state changes with task manipulations (Belyavin, 2005). However, a more practical approach might be to establish a comparison measure that represents the average physio-behavioral response over an entire experimental session for each participant (e.g., relaxed state, practice or warm-up trials, and experimental trials). This latter procedure has been shown to be relative effective and unbiased for detecting individual workload changes in ECG HRV (Fishel et al., 2007) and would not be overly cumbersome for team researchers.

Third, issues of aggregation are as important for physio-behavioral measures of team workload as for subjective measures. It is not immediately clear how measures of team members’ individual physiological processes may be aggregated to form a meaningful collective measure of team workload. However, recent research concerning neurophysiological synchronization (e.g., Stevens, Galloway, Berka, & Sprang, 2009; Tognoli, Lagarde, DeGuzman, & Kelso, 2007; also referred to as social-psychophysiological compliance—Henning, Armstead, & Ferris, 2009—and physio-behavioral synchronicity—Strang, Funke, Knott, & Warm, 2011), which occurs when team members exhibit synchronous changes in observed physio-behavioral states during performance of a team task, may provide additional, alternative methods for assessment of team workload. It has been suggested that synchronicity may index the degree of interpersonal coordination between teammates (e.g., Tognoli et al., 2007) and that greater synchronicity may enhance team performance (e.g., Elkins et al., 2009; Henning et al., 2001). As such, fluctuations in synchronicity may serve as useful indicators of changes in team workload. This viewpoint is currently quite speculative, as a strong body of corroborating research has not yet been established.

It is also worth noting that most of the computational techniques that have been (e.g., cross-correlation and cross-spectral analysis; Henning et al., 2001) or potentially could be (e.g., cross-recurrence analysis—Webber & Zbilut, 2005; cross-sample entropy—Richman & Moorman, 2000; average mutual information—Abarbanel, 1996) employed as indices of physio-behavioral synchronicity are limited to examining the relationship between two data streams (e.g., paired streams of EEG activity from two participants); multi-source methods have not yet appeared in the research literature. Thus, the validity and diagnosticity of these indices within the context of physio-behavioral measures and workload have not been established and comparisons are currently limited to dyads.

Last, it may be relatively difficult, using measures of team workload derived from physio-behavioral data, to differentiate workload into taskwork and teamwork components, though this issue may at least partially be addressed by careful temporal correlation between team behaviors and physiologic events. Overall, it is the opinion of the authors that physiological measurement of team workload is not impossible but that substantial research is required to bridge the issues outlined.

**Strategy Shift Measures**

Strategy shift measures exploit changes in team processes as indirect indicators of team workload. In this way, strategy measures are similar to behavioral observation techniques, as examining team performance episodes may reveal qualitative and quantitative transitions in team behavior with changes in task demands (e.g., Cooke, Stout, & Salas, 2001; Marks, Mathieu, & Zaccaro, 2001). To date, strategy
shifts have been examined only peripherally within the context of team workload, with research indicating that teams are likely to select lower workload strategies as the time remaining on a task decreases (Rouse, Cannon-Bowers, & Salas, 1992) and that changes in team communication may signal shifts in team workload (e.g., MacMillan, Entin, & Serfaty, 2004; Urban et al., 1996). In addition, some research has focused on training teams to shift strategies as a response to increased task demands (Entin & Serfaty, 1999), which benefits researchers by clearly defining the types of behavioral changes that signal a strategy shift though there is some risk of capturing only the types of shifts that have been trained.

Recent developments in nonlinear analytic techniques support further interesting possibilities for behaviorally and physiologically based team workload assessment. Research in this area suggests that aspects of team processes, such as patterns of communication (Dale & Spivey, 2006) and postural sway during interpersonal communication (Shockley, 2005), may be amenable to nonlinear assessment using recurrence and cross recurrence quantification analysis (among other techniques). These methods allow researchers to quantify complex patterns of team behaviors, and, potentially, shifts in those patterns may be useful as indicators of changes in team workload. This supposition is highly speculative, however, as research examining the utility of nonlinear methods for quantification of team workload using behavioral or physiological data has not yet been explored. Interested readers are referred to Riley and Van Orden (2005) for an introduction to nonlinear methods applied to the social and behavioral sciences.

**Theoretical and Statistical Considerations**

Although most of the previously discussed team workload measures rely on aggregation of individual-level data to achieve a team-level estimate, important theoretical and statistical considerations must be made to guide that process. From a theoretical standpoint, when describing and testing a team-level attribute or process, researchers must first demonstrate that the team-level construct exists (e.g., Kirkman et al., 2001). Kirkman et al. (2001) argue that to meet this requirement, researchers must establish that (a) the construct reflects the team as a whole, rather than individuals as separate units; (b) team members exhibit agreement with regard to the construct assessed; (c) the construct discriminates among teams; and (d) the origin of the construct reflects the processes of interaction that occur within the team. The last point suggests that the conceptual definition of team workload adopted should guide the measurement procedures utilized. In other words, if team workload is considered to be an emergent property of teams, then it should be assessed at the team level rather than from aggregated individual-level data. Failure to closely tie the measure to the construct definition adopted risks insensitivity to the emergent dynamics that govern the phenomenon.

With regard to statistical concerns, researchers must be prepared to address issues of non-independence of observations in their data. Improperly addressing (or simply ignoring) these issues may lead to statistical outcomes that are not reliable (e.g., Bonito, 2002). Interested readers are directed to Murray, Varnell, and Blitstein (2004) for a recent review of statistical methodologies to address these problems.

Finally, care must be taken to avoid making ecological fallacies (e.g., George & James, 1993; also referred to as fallacies of the wrong level—Glick, 1985). These occur when the unit of analysis is inconsistent with the unit of a theory (i.e., using individual data to make inferences about teams or, conversely, using aggregate data to make inferences about individuals), which may lead researchers to make...
Conceptualization and Measurement of Team Workload

erroneous inferences about the relationships studied.

Methods for Aggregating Indices of Team Workload

As mentioned previously, several different methods have been suggested as potentially useful aggregated indices of team workload, including the sum or average workload of all team members, the observed variance in team workload estimates, and the lowest and highest workload values reported (Bowers & Jentsch, 2005). The decision to adopt any of these approaches should be dependent on the research questions addressed and the team task employed (Kozlowski & Klein, 2000). The sum of all team member ratings may be the most versatile method for estimating team workload and is potentially useful whenever the goal of research is to compare team workload under varying degrees of task demands and team characteristics (though it may be less useful when the experiment includes a manipulation of the number of team members). Mean- and variance-based measures are of more questionable utility; they may be interpreted as indices of the average workload experienced by all team members and the distribution of workload across team members, respectively, but these are inherently more individual-level-focused research questions, suggesting that an individual-level measure of workload is more appropriate for addressing those questions. Finally, measures of the highest or lowest reported team workload ratings may be appropriate in situations where team success or failure may be determined by a single team member (i.e., where an inordinate degree of success or failure is attributable to a single team member, rather than to the team as a whole).

CONCLUSION

Although empirical results of team workload methods have been mixed, team workload remains a promising area of research. In 1997, Bowers et al. suggested that researchers develop measures of team workload that allow them to understand the demands placed on team members. This recommendation remains meaningful, as current measures of team workload have not significantly matured. For the time being, it appears as though subjective measures adapted from individual-level measures have received the greatest research and validation in the existing literature.

Perhaps of greater concern to the overall field is the lack of a unifying model or theory through which team workload can be described and measured. Clearly, more research is required to address this need. Although team workload remains a complex research area without simple measurement solutions, as a research domain it remains open for significant contributions from interested and enterprising researchers. We hope this discussion motivates more theory and metric development as well as empirical research in this important area of human–system integration.

ACKNOWLEDGMENTS

The authors would like to thank F. Thomas Eggemeier, Michael A. Vidulich, and Joel S. Warm for their reviews of an early draft of this article. The authors would also like to thank their reviewers for their detailed and useful feedback.

KEY POINTS

- A validated theoretical framework describing team workload has not yet been articulated.
- Most attempts to evaluate team workload rely on measures of individual workload, but this approach may not adequately capture the drivers of team workload.
- The potential exists to assess team workload using subjective, physiological, and strategy shift measures. These methods require further development and validation before they will be research ready.
- A successful measure of team workload will need to address difficult theoretical and statistical issues associated with aggregating individual-level workload responses to a team-level measure.

REFERENCES

Bailey, L. L., & Willems, B. F. (2002). The moderator effects of taskload on the interplay between on route intra-sector team communications, situation awareness, and mental workload (Rep. No. DOT/FAA/AM-02/18, Department of Transportation, Federal Aviation Administration, Office of Aerospace

Downloaded from hfs.sagepub.com by guest on July 14, 2015


Gregory J. Funke received his PhD in human factors psychology from the University of Cincinnati in 2008 and is a research psychologist in the Applied Neuroscience Branch of the U.S. Air Force Research Laboratory’s Human Effectiveness Directorate at Wright-Patterson Air Force Base in Dayton, Ohio.

Benjamin A. Knott is a research psychologist in the Applied Neuroscience Branch of the U.S. Air Force Research Laboratory’s Human Effectiveness Directorate in Dayton, Ohio. He received his PhD in applied experimental psychology from The Catholic University of America in 1999.

Eduardo Salas is trustee chair and professor of psychology at the University of Central Florida, where he also holds an appointment as program director for the Human Systems Integration Research Department at the Institute for Simulation and Training. He earned his PhD in 1984 in industrial and organizational psychology from Old Dominion University.

Davin Pavlas is a graduate of the Applied Experimental & Human Factors Psychology doctoral program at the University of Central Florida. To date, his research has examined the relationship between aesthetics and usability in user interfaces, the nature of flow state in games, the attributes of serious games...
that contribute to learning, and the role of play experience in learning.

Adam J. Strang received his PhD in psychology from Miami University in 2010 and is a postdoctoral fellow with the Consortium Research Fellows Program. He is currently in collaboration with the U.S. Air Force Research Laboratory to develop metrics of team coordination and synchrony from physiological (e.g., heart rhythm) and behavioral (e.g., body movement) responses.

Date received: April 8, 2011
Date accepted: September 13, 2011