Beyond Happy Hour: Lessons Learned in BARS (Behaviorally Anchored Rating Scales)

Jennifer K. Phillips, Holly C. Baxter, PhD, Allison K. Hancock, PhD, & Morgan R. Borders
Cognitive Performance Group
Portsmouth, VA
jenni@cognitiveperformancegroup.com, holly@cognitiveperformancegroup.com, allison@cognitiveperformancegroup.com, morgan@cognitiveperformancegroup.com

ABSTRACT

Traditional military assessments often fall short of capturing the dynamic nature of Warfighter performance on complex decision and problem-solving tasks. Behaviorally Anchored Rating Scales (BARS) offer a solution by providing structured, observable criteria to assess performance in contexts ranging from leadership to planning and adaptability. While traditionally used by psychologists and researchers, BARS can be effectively employed by uniformed personnel to evaluate specific cognitive competencies and determine implications for both short-term talent development and long-term career progression. This paper explores the requirements associated with successful application of BARS, drawing on their usage within the U.S. Navy, Marine Corps, and Army, and emphasizing their flexibility to measure performance across various contexts and purposes. Specifically, the paper examines the circumstances under which BARS are particularly beneficial, such as when quantitative data on cognitively complex performance is required, when assessing the quality of skills with no clear right answers, and when performance must be observed to be measured. It also discusses the importance of developing a strong BARS rubric grounded in a model of skill development, built with data collected from domain subject matter experts, and reflecting domain-specific language in clear, concise descriptors. Through an analysis of past BARS applications, the paper highlights lessons learned about rater qualifications, proper training and calibration, and implementation approaches. Examples are provided from military use cases including leader selection and development, communication skills, field exercise performance feedback, and validation of automated measures captured via advanced software. Finally, the paper explores future opportunities to enhance BARS application, including leveraging large language models to augment observations and incorporating artificial intelligence to gain efficiencies in rubric development. By expanding the use of BARS across all levels of leadership, military organizations can improve their ability to assess and develop human performance, ensuring objective, consistent evaluations in support of force readiness.

ABOUT THE AUTHORS

Jennifer K. Phillips is the Chief Executive Officer and a Principal Scientist at the Cognitive Performance Group. Her research interests include skill acquisition, cognitive performance improvement, and the nature of expertise. Ms. Phillips has over 30 years of experience modeling performance across the levels of proficiency, designing learning solutions including decision-centered training scenarios and facilitation techniques, and developing metrics for cognition and decision making.

Holly C. Baxter, Chief Scientist of Cognitive Performance Group, has spent over 25 years specializing in cognitively based Instructional Design, Evaluation Metrics, Organizational Development, and Training in both military and commercial environments. Dr. Baxter has published numerous articles in the fields of cognitively-based training, assessment, and knowledge management; has been an invited speaker at multiple conferences and events; and has given dozens of workshops on Cognitive Task Analysis, Knowledge Capture & Transfer, Knowledge Management Assessment, Intuitive Decision Making, and Leadership Development. Dr. Baxter earned a Ph.D. from Indiana University in Organizational Communication and Human Resource Management.

Allison K. Hancock is a Senior Scientist at the Cognitive Performance Group studying operational military research and human performance assessment. Dr. Hancock has a Ph.D. in Educational Psychology with specialization in Sport and Exercise Psychology and Measurement and Statistics, and a M.A. in Applied Experimental Psychology. She has

over ten years of experience leading research projects with distributed teams for the DoD in the areas of human performance, rapid training development, cognitive skills training, program evaluation, and assessment.

Morgan R. Borders is a Scientist at the Cognitive Performance Group with an MS in Human Factors Psychology. Her research interests include the intersection of the cognitive and social aspects of complex real-world problems, such as the use of social media in influence campaigns. Methodologically, she applies a contextualized multi-level perspective combining lab and field-based methods, such as computational language analysis.

Beyond Happy Hour: Lessons Learned in BARS (Behaviorally Anchored Rating Scales)

Jennifer K. Phillips, Holly C. Baxter, PhD, Allison K. Hancock, PhD, & Morgan R. Borders
Cognitive Performance Group
Portsmouth, VA
jenni@cognitiveperformancegroup.com, holly@cognitiveperformancegroup.com, allison@cognitiveperformancegroup.com, morgan@cognitiveperformancegroup.com

PERFORMANCE ASSESSMENT AND BEHAVIORALLY ANCHORED RATING SCALES

Military organizations have shifted from centralized to decentralized decision making and from an industrial to an information age model of training and development to modernize how we prepare Warfighters to adapt, innovate, and achieve decision superiority. The modern battlespace demands that we prepare every individual down to the lowest echelon to make sense of situations, apply good judgment, and take decisive action. Yet, how do we know whether our Warfighters are trained and prepared to respond to current day decision challenges? We must have tools to assess their cognitive skills, and Behaviorally Anchored Rating Scales, or BARS, are proving to be an effective solution.

The purpose of this paper is to document our best practices from over 15 years of applying BARS with military audiences. Our experience indicates that BARS are highly effective when applied under the right conditions, though early iterations exposed pitfalls in rubric development that we have resolved through iterative refinement. These lessons are worthy of sharing with assessment professionals in the community considering BARS for their own work. We do not describe methods for developing sound rubrics in this paper; readers with interest in our development method are referred to Ross and Phillips (2020) and Phillips et al. (2017). Similarly, we do not discuss the psychometric properties of BARS in detail since each unique rubric requires independent validation. When we have been sponsored to conduct validation studies, the BARS rubrics have shown criterion validity (Phillips et al., 2017) and strong inter-rater reliability when raters participated in 60 minutes of training and calibration (ICC = .90, CI: 0.82-0.97, n = 15).

We begin by briefly describing a BARS rubric and what it looks like, then discuss four use cases to demonstrate the range of BARS applications. We describe the circumstances for which BARS are a good choice as an assessment approach. Then, we move into a section on lessons learned specific to developing BARS rubrics, and another section on implementing those rubrics when military personnel instead of researchers are doing the application. Finally, we end with the future directions we plan to take to tackle specific challenges we are encountering at this stage of our BARS usage, and to make their application more efficient using artificial intelligence.

What is a Behaviorally Anchored Rating Scale?

A BARS is a rubric for scoring performance by associating an observed behavior with a numeric value, with higher numbers reflecting more advanced performance. While similar to a Likert scale, which in practice might range from strongly disagree to strongly agree for every item and include a neutral rating at the center of the scale, a BARS differs in that the scale progresses from novice to highly skilled performance with no neutral rating, and each numeric value is associated with a unique word descriptor so that every item consists of its own distinct scale. Raters applying a BARS watch the individual or team perform and match what they see about the performance to the numeric anchor that best characterizes what they see. Figure 1 illustrates a segment of a BARS rubric, which consists of items organized by category of performance and consisting of their distinct anchors.

BARS IN PRACTICE

As an assessment and developmental tool, BARS have proven quite versatile and easy to implement. We have developed and implemented BARS rubrics for a variety of user communities and purposes, and we share a few of those case studies briefly in the paragraphs to follow.

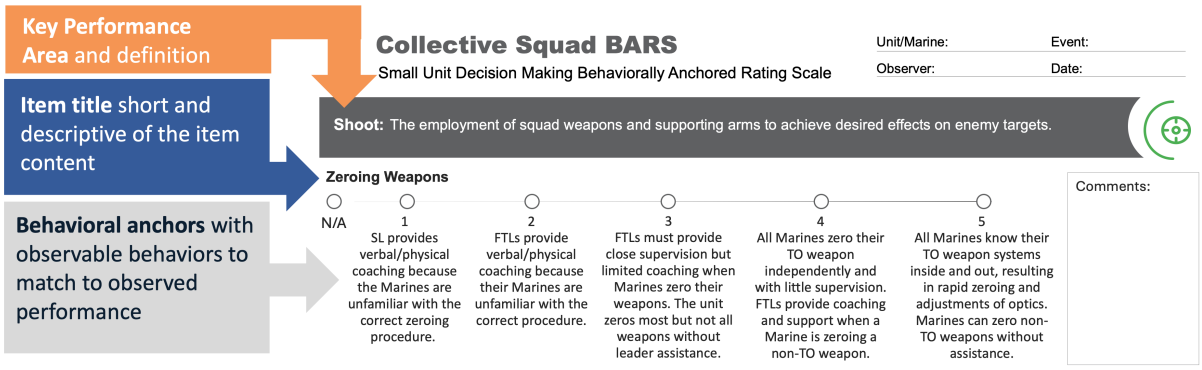

Figure 1. Zeroing Weapons Item from the Collective Squad BARS within the *Shoot* Key Performance Area

Feedback and Individual Skill Development

In 2014, the Marine Corps was interested in attacking the problem of how to rapidly develop the skills of active-duty instructors. While Marines who were selected for instructor duty were recognized as highly proficient in their craft, they typically lacked experience teaching others. And, they were slated to serve no more than three years as a schoolhouse instructor, limiting the time available to develop their instructional skills. The Marine Corps wanted a tool to provide feedback to these new instructors, and they needed the tool to be based on an accepted standard of performance rather than subject to individual opinions about what constitutes good instruction. Moreover, they needed a way to not only tell the rookie instructor how they were doing, but also of helping that individual become more and more effective. And finally, the schoolhouses wanted a means of taking stock of their whole instructor cadre, which presaged the need for objective and quantitative data about where each instructor stood relative to their peers. With these objectives in mind, we co-developed a BARS observation rubric to provide feedback about an instructor's facilitation of a period of instruction, and a BARS supervisor rubric administered like a performance review, where holistic performance as a member of the school staff was assessed on an annual or semiannual basis.

The form and function of the BARS suited all the Marine Corps' objectives. We conducted an analysis that identified the Key Performance Areas (KPAs) for a Marine instructor. The KPAs were similar to competencies and included both in-classroom concepts such as *Communication & Delivery* and *Learning Environment*, as well as outside the classroom requirements including *Planning & Preparation* and *Community of Practice*.[1] The KPAs were one important component of defining the standard for a Marine instructor. The other means by which the BARS rubrics defined the objective standard was in the word-descriptor anchors associated with each item. A rating of 1 on an item described what a novice instructor would do, whereas a rating of 5 described how an expert would perform the same task but with high effectiveness. The anchors not only defined the standard for the observer's rating but also provided an aim point to instructors striving to improve—that is, attempt to do what is described in the next word descriptor up from your current rating. In implementation, BARS were used as part of the "murder board" process to certify that the instructor was ready to teach independently "on the podium," and were applied at various intervals to track improvement over time and give that feedback to the developing instructor.

Evaluation of Individuals or Teams

Since 2020, the Marine Corps has been interested in measuring performance on the cognitively complex tasks required during tactical operations. Recognizing that the modern battlefield requires agility and innovative thinking to be effective against the adversary, the Office of Naval Research initiated efforts to develop new measurement tools that provide objective data about the decision-making abilities of commanders as well as the readiness of units (Phillips et al., 2022). BARS rubrics have been developed for individuals and teams across echelons: squads, squad leaders, platoon commanders, companies, battalion staffs, battalion commanders, intelligence analysts, and logisticians.

[1] The KPAs that could be observed during a period of instruction were included on the BARS observation rubric; citizenship and other KPAs not directly observable in the classroom were included on the supervisor BARS rubric.

The uses of the BARS rubrics to evaluate individual and unit decision making and readiness in this domain are numerous. In one case, the squad BARS are applied as the primary evaluation tool in the Annual Squad Competitions to determine the highest performing squad within Marine divisions as well as Corps-wide. Similar to the previous instructor example, it is crucial to use an objective and standardized assessment tool to compare units to each other and ultimately select the highest performer. Multiple raters are necessarily involved in this process that spans days and different tactical challenges, and the raters must assess the same components of performance with the same scale to sum performance across days and ranges. One of the core assessment challenges here, as is generally the case for military organizations broadly, is that performance is conducted in a dynamic context where mission accomplishment may be achieved by taking on varying subsets of tasks. First squad may encounter a jammed weapon on their way to the objective and have to adapt in stride. Second squad may have a team leader with a sprained ankle and adjust to accommodate the injury. Furthermore, the squads may encounter the same task but make different decisions, both effective, about how to tackle it. The assessment tool must be flexible enough to be applied evenly despite unique paths and decisions on the way to complete the mission. Since BARS allow for more than one right answer, they prove an effective tool to measure to standard despite variations in problem-solving approaches.

The Marine Corps also seeks a means to longitudinally measure training gains and the readiness of units, such as of battalions prior to deployment. This use case requires a standardized assessment approach that can account for complex performance across echelons (i.e., squad to regiment). Moreover, the assessment approach must withstand constantly changing test scenarios since battalions may be preparing for different mission sets, adversaries, and geographies. The rubric applied to a unit preparing for operations in the Middle East must also suit one readying for combat in the Southeast Asia. Because the BARS are situation-agnostic, they can serve as the primary assessment tool even as the battalions face unique scenarios at each assessment interval. Without a common and objective way of tracking performance trends over the unit's training cycle, the commander (and commander's boss) would have to sift through evaluator notes and comments to extract meaning and would lack the ability to pinpoint areas of improvement, decrement, or no change over time. In addition, by measuring performance at every echelon within the unit, the commander can see how each subordinate unit contributes to or hampers overall battalion effectiveness.

Performance and Promotion Review

In parallel with the Marine Corps' implementation of BARS to assess decision making and readiness, the Navy is investing in an effort to define and evaluate leadership behaviors specific to command roles. Led by the Office of Naval Research, the development of the Navy Leadership Behavioral Model aims to clarify what effective leadership looks like across operational contexts. When complete, this model will include a 360-degree assessment component anchored in BARS to provide granular insight into leadership strengths and developmental needs. This marks a shift toward a more evidence-based, behaviorally-grounded leadership evaluation. By assessing leadership through a standardized and observation-driven rubric, the Navy hopes to not only recognize high-performing leaders but also to guide individual professional development and inform leadership training strategies across the Fleet.

Simultaneously, the Army Research Institute is applying BARS to explore how noncommissioned officers (NCOs) develop interpersonal communication skills, particularly in challenging and variable contexts such as crisis response, inter-rank communication, and external coordination. Recognizing the complexity of communication demands in modern operations, the Army's effort emphasizes adaptability in speaking up, managing downward, collaborating laterally, and engaging with external stakeholders. The BARS being developed for this initiative aim to chart a developmental trajectory for NCOs as communicators and to provide specific behavioral markers for each stage of growth. This structured approach enables targeted feedback, more tailored training, and ultimately supports the cultivation of communication agility as a core leadership skill.

Evaluation of Planners and Analysts in Information Operations

Operations in the Information Environment (OIE) are both a longstanding and increasingly central element of modern warfare, requiring analysts and planners to integrate planning, intelligence, and influence activities to shape perceptions and protect decision-making processes. The information environment (IE) spans the cognitive, informational, and physical domains, and demands the use of advanced technologies and cognitively complex methods to process and act on vast volumes of data to achieve strategic effects. As the complexity of the IE grows, so does the need to evaluate how effectively personnel operate in this space. To meet this need, the Operational Mastery of the information Environment (OMEN) BARS, or O-BARS, was developed to assess analyst and planner performance in

OIE related tasks (Borders et al., 2023). The O-BARS provides standardized, scenario-agnostic measures that evaluate key cognitive skills such as judgment, reasoning, and decision making.

Building on this foundation, a more targeted extension of the tool, the BEND BARS, was created to assess how analysts apply the BEND Framework (Carley, 2021) and use the associated advanced tools and technologies developed to assess the IE. The BEND Framework categorizes influence strategies including 16 maneuvers aimed at shaping narratives and communities. The technologies that support analysis of the IE and BEND allow analysts to track, classify, and respond to adversary influence operations in real time. The BEND BARS evaluates both the cognitive application of the framework as well as the analyst's interaction with these technologies, assessing how well users interpret data, apply influence strategies, and produce actionable assessments. In this use case, BARS provide a means to measure the goodness of a software application over and above the human analyst's performance.

WHEN TO USE BARS FOR ASSESSMENT

As we have seen in the cases above, BARS can be useful for a variety of purposes and circumstances. We have also tested BARS in settings where they were less effective and revealed important limitations. In this section, we describe some of the characteristics of the need or the setting that makes BARS a good choice for performance measurement.

1. When You Need to Measure Cognitive Performance

BARS are best suited when you need to assess performance on complex tasks that call for judgment, sensemaking, reasoning, and decision making. Assessing cognitive skills is quite difficult. Outcome measures can be flawed because good decisions can result in less than stellar outcomes if executed poorly. On the other hand, poor decisions can meet with luck to result in a positive outcome. It is impossible to get inside a person's head to see the rationale behind their decision and it is overly time-consuming to elicit that rationale through questioning. BARS alleviate these factors by identifying observable elements of performance that align with levels of cognitive proficiency. For example, we see novices exhibit behaviors like asking no questions while experts ask insightful questions that anticipate a future event or action. Those observables give us enough of a window into one's cognition to, together with other observables, arrive at a good approximation of the individual's proficiency.

2. When You Need Objectivity Instead of Subjective Judgment

Less experienced raters may judge an individual to be a higher performer than an expert rater who is better qualified to see the gaps and oversights in performance. Even two raters with equivalent resumes are likely to have experienced diverse professional challenges and therefore judge performance quality differently. Leaving the evaluated individual to the mercy of varied, sometimes conflicting, feedback can be detrimental to development. Making policy, readiness, or promotion decisions from subjective rater judgment introduces a host of risks to organizational functions. BARS are a favored measurement technique because raters utilizing BARS are less prone to biases such as the halo effect or positive leniency (Muchinsky, 2003; Riggio, 2000), and calibrated observers show strong inter-rater agreement.

3. When Standardization Is Important Due to Multiple Raters

Tasks that require judgment and decision making , described in the first paragraph, are the sorts where evaluators have different amounts and types of experience and as such, are likely to have biases about what good looks like. Standardization ensures all raters evaluate the same elements of performance. When evaluated individuals and units take on dynamic problems, rater biases emerge as "pet" topics to which they lend more credence and weight. For example, in rating performance on call for fire tasks, one rater with an aviation background may focus on weapon to target match with disproportionate feedback about the air assets that should have been considered, while a second rater with a fire support background may focus on communications protocol and the criticality of correct terminology to prosecute the fire mission. Both raters have good points of feedback, but if the data about the event are not standardized across the two raters, and likewise against past and future raters of other units, the results are of limited use. The units rated by the different raters cannot be compared to each other or previous units, nor do they receive a well-rounded assessment. BARS rubrics are especially useful for military domains where rater backgrounds and experiences influence what they attend to when assessing performance.

4. When You Need Quantitative Data about Cognitive Skills and Abilities

Similar to the first need, BARS are well suited for requirements to obtain quantitative, numeric data about decision making or other cognitive performance. The word-descriptor anchors in BARS items equate an observable behavior

with a level of cognitive proficiency. As a result, an individual or unit rated as a 4 on a 5-point scale can be considered a high performer. Quantitative data are amenable to comparisons across people or units and over time, making BARS an excellent tool for applications like examining the effectiveness of a course on planning, selecting individuals for leadership jobs, or comparing a unit in the middle of its training cycle to the baseline for units at that stage of training. Users in Marine Corps schoolhouses report the implementation of BARS enables them to set a minimal score requirement for instructors cleared to teach, and then speed the development of new instructors to reach those requirements. In addition, quantifying performance with BARS ratings allows for diagnosis of strengths and weaknesses. In our work with Marine Corps squads, we have been able to report the items on which squads across the Fleet score highest and lowest; trainers of squads express the importance of this information, as it reveals areas where training is most and least effective.

5. When Performance Must Be in Context and Dynamically Responsive to the Situation

BARS are specifically designed for assessing the application of knowledge and know-how to a problem in context. Knowing the distance your machinegun team can cover on foot in an hour on a schoolhouse test is quite different than deciding, during the 8th hour of a night operation on hilly terrain with dense vegetation, when to direct your machinegun team to reposition to the next hill without being seen by the enemy so they can provide support by fire in an hour. BARS are not only appropriate for assessing a squad leader on the latter, but they may be the only objective and standardized way of assessing such dynamic performance, where know-how can only be exhibited through *in situ* performance. We have seen BARS applied effectively in schoolhouse settings as part of a broader assessment strategy, but they tend to work better for practical application exercises where students must solve problems by exercising judgments and making decisions. BARS work well for culminating exercises where students conduct a task or produce a product in response to a rich scenario. They don't work effectively for exercises where the students are being taught procedures, such as when they are learning how to conduct a routine sequence of steps. Procedural exercises at the "crawl" level of learning tend to have one correct way to perform them and thus are not suited for BARS.

6. When There Is a Range of Performance Goodness

BARS are well suited for tasks where there are multiple right answers, where more than one approach can achieve a suitable outcome, and where there are gradations in the quality of responses. We worked with instructors at a school who were frustrated with their inability to differentiate high performers in the class who exceeded the standard from mediocre performers who merely met the standard. The instructors knew who the top students were but were unable to reflect that excellence in the gradebook, where entries were allowable only for pass and fail, or only for trained, partially trained, and untrained. Since BARS use a scale ranging from beginner to expert levels of performance, they enable raters to capture when individuals complete the task in a more or less satisfactory manner. BARS also are characterized by behavioral anchors agnostic to the specific response by the performer, therefore allowing a range of appropriate decisions to all be characterized as effective.

7. When Test Problems Vary across Time or Performers

Outside schoolhouse settings, and sometimes even in the schoolhouse, military organizations often customize the "test" scenarios to be responsive to training and operational needs. Such flexibility is important but can present a problem standardizing assessments across them. BARS rubrics work effectively under these conditions because they are agnostic to contextual elements and decision-making requirements. For example, the BARS for instructors is applicable whether the individual is teaching aircraft maintenance or maneuver tactics. The BARS for battalion staffs is appropriate whether the unit is being assessed on a Middle East or Southeast Asia scenario.

HOW TO DEVELOP A STRONG BARS RUBRIC

BARS rubrics vary in quality, and their effectiveness depends on several key factors. In this section we outline some critical lessons learned that influence rubric quality across three main stages of development: content collection, data analysis and rubric generation, and validation stages.

Content Collection

The content, or data, collection phase of rubric development obtains the subject-matter expertise needed to identify the items along which to assess and the behavioral anchors, or word descriptors, for the scales.

1. Use a Developmental Model as the Rubric Foundation
In our implementation, we use BARS exclusively to provide a window into the cognitive proficiency level of the assessed person or group. Therefore, the underpinning of the rubric must be a developmental model. We favor the Dreyfus and Dreyfus (1986) model of cognitive skill acquisition because it articulates how individuals progressively advance from a beginner to a master in naturalistic domains characterized by incomplete information, vague goals, multiple acceptable courses of action, and dynamic settings. It also has been validated across multiple domains. The Dreyfus and Dreyfus model identifies five stages of development from novice to expert (master is a 6th stage which we and others drop from our work for practical reasons); we have found five points to be a fitting number of scale anchors to use when differentiating performance. More anchors would be onerous and arguably impossible to generate for every rubric item. Fewer anchors—many developmental models favor three levels of performance—miss out on the important variations in performance that occur within the intermediate stages of growth, where most people function. Identifying a foundational developmental model for your BARS provides guidelines for the behavioral anchors you will create. For example, our foundational model dictates that a "1" on the scale reflects rule-based performance without considering how situational factors impact application of the rules, and a "3" reflects deliberate and analytical planning but hesitancy to veer away from the initial plan. In other words, the underlying model provides the skeleton for your BARS rubric and informs your anchor writing.

2. Conduct Knowledge Elicitation Interviews
Rubric development can start with a blank slate or can re-use content from other products. However, we strongly recommend BARS rubric content to come from first-hand knowledge of how to perform the task being measured as opposed to doctrinal writings or manuals. Doctrine tends to be too general for use as assessment fodder. Manuals tend to be too procedural. Neither source provides insight into how performance on the task varies with levels of proficiency. At best, you can identify behavioral anchors at the middle of the scale, that is, competent but not expert performance, using doctrine and manuals, but you will not be able to create content for the low or high ends of your scales. Instead of using written material, conduct interviews to elicit knowledge from subject-matter experts (SMEs) who have witnessed, and ideally coached, individuals performing the task well and poorly. We recommend the Developmental Progression Interview Technique (Phillips et al., 2013) which elicits domain-specific descriptions of how individuals perform at each stage of your foundational model. The number of interviewees required depends on the size of the job you seek to assess with your BARS rubric. If you want to assess a discrete job or task like conducting security for a particular site, you will need 4-6 SMEs. For more sprawling jobs like leadership across Navy organizations we recommend 20-24 SMEs.

3. Frame Data Collection around the Types of Individuals Who Will Be Evaluated
While it may seem intuitive to focus data collection on the target evaluation population, one project highlighted the consequences of misalignment between rubric scope and audience. We set out to help a schoolhouse better prepare its students for tasks they would encounter once assigned to a unit. We collected data from individuals operating in a variety of units and captured the progression of development from incoming to seasoned Warfighter. After all, this defined the expectations for course graduates. When we went back to the schoolhouse with our BARS rubric to help them assess students, we encountered two problems with our approach. First, the students in the classroom had no opportunity to perform many of the tasks elicited from unit personnel. Second, all the students in the entry level class performed as novices. The five-point rubric returned average scores ranging from 1 to 1.5 for all the students, rendering it useless as a tool for differentiating student performance. The rubric was somewhat more useful for students in advanced classes, but it still fell short because the course objectives were not nicely aligned with the BARS items. Our lesson learned was to ensure our rubrics were customized to the specific population for evaluation.

Data Analysis and Rubric Generation

Once content is available from the data collection step, analysis and rubric creation can take place.

1. Organize the Rubric by Meaningful Categories
Our analytic approach is to conduct a thematic analysis that defines KPAs, which are similar to competencies. The KPAs serve two purposes. First, they support rater cognitive load by directing their focus to the performance they have observed and wish to rate during an evaluation. Second, they are informative as to the performance strengths and weaknesses of the evaluated person or group, enabling subsequent actions to conduct more training on areas of weakness, contribute explanations for other performance metrics, and enable trends to be tracked over time. The best KPAs are short, clear descriptions of the desired performance that are meaningful to the user community. For example,

our squad BARS consist of five KPAs—*Shoot*, *Move*, *Communicate*, *Decide*, and *Esprit de Corps*—which reflect the language of small unit operations.

2. Restrict the Number of Items in the Rubric

An ideal rubric consists of 15-25 rated items. More items inflict a burden on raters, who in our experience are generally active-duty domain practitioners who wear many hats and are not assessment or learning professionals. It is acceptable and sometimes necessary to generate more than 25 items to sufficiently reflect the requirements of a job. For one domain, we developed 124 items across 9 KPAs. In those cases, we recommend treating your items as a pool from which the practitioners can draw instead of a set that must be evaluated for every assessed event. Different scenarios are usually designed to exercise different subsets of the job, and the 15-25 BARS items most relevant to the scenario should be extracted from the larger pool of items. In addition, when the task being observed follows a typical temporal sequence, we recommend sequencing the BARS items on the rubric in the order in which they are likely to be observed.

3. Show Clear Progression of Development in the Behavioral Anchors

The purpose of BARS items is to identify the cognitive proficiency of the rated person or group; therefore, the anchors should clearly reflect how individuals improve as they mature. The progression should be crystal clear to the raters which means all the anchors across an item must be internally consistent, following the same formula or cadence in their wording. Subsequent anchors either add on to the prior anchor or replace one behavior with a new, more advanced behavior. The thread from one anchor to the next must be clear and use the same terminology to minimize cognitive load. Figure 2 provides an example of an effective thread showing progression of development along an item.

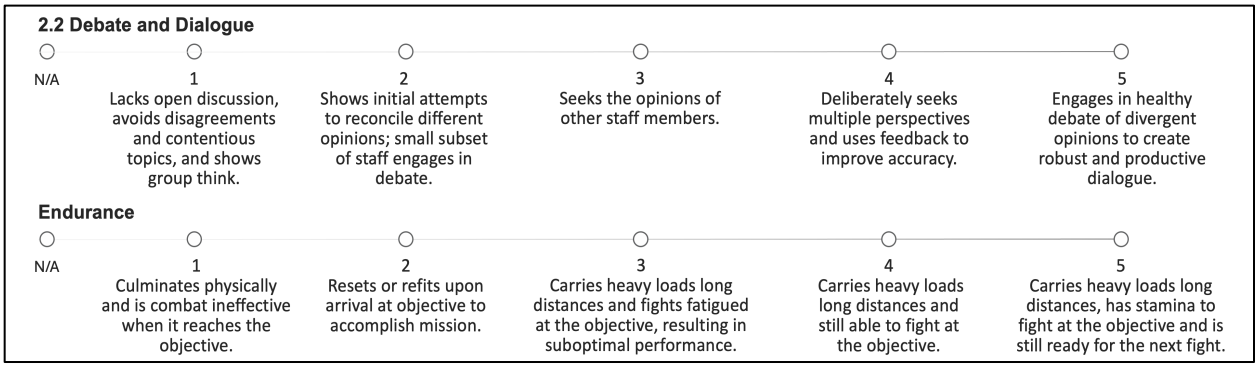

Figure 2. Two Examples of Consistent Terminology across the Behavioral Anchors

4. Be Brief and Concise

The shorter the item title and anchor, the better. Like all assessment tools, it is critical that each item includes only a single concept to rate. The use of "and" in behavioral anchors can be dangerous; when raters observe one of the behaviors described but not the second behavior they don't know whether the performer has earned that rating or not. In addition, brevity in item titles and anchors serves to minimize the cognitive load of raters, who in our experience are generally tired and time-pressured when they are completing their BARS assessments.

5. Use Domain Terminology

As psychologists or learning professionals, it is tempting to use big words and academic speak. After all, words have meaning, and big words can help us be more concise. However, it is imperative that the lingo used by domain practitioners, even their slang, is present in the KPAs, item titles, and word descriptors. We have learned to use word-for-word descriptions from our interview data when possible. The overriding goal in writing the rubric is to ensure the rater readily understands the meaning behind the items.

6. Validate With Domain Experts

Validation is a necessary step in rubric development, as with any assessment, though the approach can vary depending on context and available resources. One critical, yet often challenging, aspect is obtaining solid feedback from SMEs. This step is essential for confirming the clarity, relevance, and accuracy of rubric content. While we do not prescribe a universal method, we recommend confirming face validity through SME review and assessing inter-rater reliability

when feasible to ensure the rubric performs as intended. Beyond these steps, validation methods should be tailored to the specific goals and constraints of the use case.

IMPLEMENTING BARS WITH A USER COMMUNITY

When bringing BARS to a community of practitioners to implement, expect to scaffold your support to the community in the areas of their performance observation (or data collection) process, the data management approach, and their analytics and usage of the BARS data.

Performance Observation (Data Collection)

1. Define the Performance Event To Be Assessed
Raters must be given clear parameters for what performance they should evaluate. BARS are typically applied during a learning event like a course practical application or culminating exercise, a training event like a unit's field training scenario or a force-on-force battle, or an instance of job performance such as an instructor conducting a period of instruction. They can also be used during performance reviews and initial training assessments to determine where leaders need additional skill growth. Within those events, it is critical to articulate what constitutes complete performance. Is it every time the rated individual performs an item on the rubric? Is it the last best version of that performance? Is it the average of all performance across the event? While we as researchers would like to have data for every time the rated performs (e.g., for an instructor, every instance when they attempted to engage a student), in practice we have found that goal to be unsupportable. We generally work with the raters to define, for their use case, how they should identify a single rating for multiple instances of the observed behavior during the event.

2. Define Meaningful Performance Increments
Work with the user community to identify meaningful start and end times for the performance being rated. It is desirable to have raters conduct multiple BARS ratings (i.e., complete an entire rubric consisting of 15-25 items multiple times) over the course of a multi-day performance event. Our practice is to work with the practitioners prior to the event to identify where natural breaks are likely to occur. Sometimes the natural breaks are clear, such as the instructors' standard 50-minute period of instruction where raters complete one rubric per period. In Marine Corps squad competitions, the squads take on four different ranges over the 72-hour event and observers complete one rubric for every range. When natural breaks are unclear, like in a 96-hour force-on-force exercise, we recommend a single rubric per 12- or 24-hour period. Gathering multiple assessments from a single event can be valuable, for example, to show whether fatigue impacts performance as evidenced by decreasing scores over time, or to see how well a unit is equipped for night versus daytime operations.

3. Identify an Appropriate Data Collection Format and Process
Selecting the right data collection format is essential for successful implementation. We recommend supporting both paper-based and digital options to accommodate different user needs. Paper forms (e.g., PDFs) work well when the goal is to provide feedback but are not ideal for analytics due to the need for manual data entry. Electronic formats, such as tablets or survey tools (e.g., Microsoft Forms), streamline data management and enable efficient analysis across raters and events. The most effective approach is for raters to take observational notes during the event and complete the BARS afterward, when they can reflect on performance more accurately and apply the scale consistently.

4. Select, Train, and Calibrate Raters in Advance
Effective use of BARS requires careful selection, training, and calibration of raters. Ideally, raters should have enough experience to interpret high-level performance descriptors, though less-experienced individuals can be trained to apply the rubric reliably. A focused training session, typically one hour, should cover the purpose of BARS, structure of the rubric, and how to derive ratings from observed behavior. A separate calibration session is essential to ensure consistency. Without calibration exercises, we have found inter-rater reliability to range from .47 to .56 (Cohen's kappa; Phillips et al., 2017). With calibration sessions, inter-rater reliability has ranged from .72 to .87. Calibration sessions should include activities such as video scoring, simulated scenarios, and dress rehearsals. These steps help align rater interpretations, improve inter-rater reliability, and ensure that BARS data are both accurate and actionable.

5. Allow Observers to Add Comments
Many of the organizations with whom we have worked recognize the importance of observer/evaluator comments, and in some cases those comments formed the basis of the assessments prior to the introduction of BARS. These rater comments are critical for feedback to the performer, and we recommend all BARS scales give raters the option to add comments associated with individual items as well as summary comments. In analyzing BARS results, the observer comments also provide insight into how raters interpret and apply the scale, which is valuable for future revisions.

6. Standardize a Dashboard for Analysis
BARS lend themselves well to analysis through standardized dashboards due to their straightforward scoring structure. Users can examine performance at the item level, calculate average scores for each KPA, and derive an overall score by averaging KPA scores. While this approach can unintentionally weight smaller KPAs more heavily, we accept this trade-off given that all KPAs are selected for their equal importance. Customized weighting is also an option. Dashboards should support trend analysis over time, particularly for multi-day exercises or repeated assessments. We recommend Power BI for its interactivity, visual quality, and low development overhead, though any tool capable of basic statistical functions is sufficient. Tools must also be compatible with DoD systems to ensure accessibility for end users. Finally, incorporating inter-rater reliability metrics and, potentially, sentiment analysis of rater comments can further enhance the utility of the dashboard for both feedback and future calibration.

FUTURE DIRECTIONS

One current challenge in our application of BARS rubrics is whether and how they can be employed to support an individual's career progression, i.e., a 20+ year span of time. An enduring performance assessment capability would prove valuable for consistent, career-long support of skill development as well as a standard approach to evaluating complex cognitive task performance akin to using the Fitness Report or Officer Evaluation Report for annual reviews. The challenge lies in the fact that military personnel, more so than others, change job duties with frequency. The Marine Corps' instructor BARS would certainly maintain relevance for a career military instructor, since it is focused on the job of teaching. However, when Warfighters take on increasing levels of responsibility over their career—the responsibility of a tactician leading a platoon is different from one leading a company—and encounter new knowledge requirements at higher echelons of duty—such as new asset sets and increased coordination requirements—it becomes more difficult to create a BARS to cover one's career. However, we hypothesize that a set of core capabilities, like communication or command presence, maintain relevance to a military occupational specialty even as individuals take on more and different job duties. In our work with the Army Research Institute, we will test that hypothesis. We are currently finalizing a Communication BARS to be applied across the career of an NCO, that is, from E4 to E-9. We believe this line of research will yield important new best practices for identifying how BARS rubrics can support assessment, and importantly, skill development, across one's career.

Building a good BARS rubric admittedly relies on researcher experience to do it well the first time. Moreover, the process can be time-consuming. In some cases, we have found it difficult to access all the SMEs required for the knowledge elicitation step. As an innovative approach to making the development process more efficient and less subject to SME availability, we have introduced the use of generative artificial intelligence (AI) to support rubric development. In our Army NCO communication project, interviewees with highly specific communication experience, such as external engagement or crisis interaction, were unavailable. We applied generative AI to synthesize insights from existing literature and related behavioral data. This approach allowed us to fill gaps in the behavioral anchors with plausible, evidence-informed examples while maintaining alignment with known patterns of NCO development. Importantly, AI was not used to replace human judgment but rather to augment the process, generating initial drafts that were then reviewed, refined, and validated by human experts. This was our first experience integrating generative AI into the rubric development process, and it has demonstrated that AI can serve as a valuable tool for creating more robust and comprehensive behavioral rubrics in military settings.

The ethical considerations associated with generative AI usage are important. We offer that AI can be used ethically when applied to fill gaps in one's dataset using only that set of materials identified by the researcher as relevant and approved by the customer or user community. Generative AI should not be tasked to pull from colloquial sources (i.e., general, open-web ChatGPT databases), nor should it generate the rubric content in its entirety. In addition, AI-assisted content must be transparently called out to the client and clearly visible to reviewers of the rubrics.

We speculate that technology tools and AI might also support rubric implementation, making the measurement approach more scalable and decreasing the requirements for human observers. In our current O-BARS work, we are augmenting human observation with automated data collection. This hybrid method includes capturing real-time system interactions and tool use alongside evaluator observations of verbalizations, behaviors, and team dynamics. Given the complexity of many measures of performance, early findings suggest that a combined approach offers the greatest potential. Automated data can capture detailed behaviors that are difficult for observers to track in real time, while human ratings provide valuable context for interpreting decision-making behaviors and collaborative processes. In the future, we speculate that application of large language models (LLMs) may be effective in augmenting or replacing the human rater by deriving meaning from the performers' utterances. An LLM-assisted approach would depend on the ability to record and obtain digital and verbal communication, and training of the LLM to approximate the human reasoning that occurs when a rater judges which of the anchors best matches the behavior that is seen. Practically speaking, it is more likely an LLM could recommend or draw a human's attention to certain verbalizations, thereby augmenting rather than replacing the human observer. Yet, we see this line of research as an important future direction for making assessments of cognitive skills more prolific and user-friendly. The data governance issues associated with LLM application would include following approved human subjects protocols to protect performers and following strict security protocols to protect the content of potentially sensitive communications and events.

BARS have emerged as a powerful tool for capturing the complexity of human performance in today's military. Over 15 years of application have shown their value not just in assessing individuals and teams, but in driving meaningful growth and readiness across the Force. By turning invisible cognition into observable behaviors, BARS replace speculation and subjectivity with objective measurement. As we look to the future, tools like generative AI and LLMs offer exciting new frontiers for making these assessments even more scalable, responsive, and impactful.

ACKNOWLEDGEMENTS

The authors wish to thank the Office of Naval Research and specifically Dr. Peter Squire, LCDR Michael Natali, CDR Jacob Norris, and Dr. Rebecca Goolsby; and the Army Research Institute, specifically Dr. Larry Golba, for their guidance, support, and sponsorship.

REFERENCES

Borders, M.R., Ross, W.A., and Williams, M.L. (2023). Assessing information maneuver performance and effectiveness. *Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, Orlando.

Carley, K.M. (2020). Social cybersecurity: an emerging science. *Computational & Mathematical Organizational Theory*, 26, 365–381. https://doi.org/10.1007/s10588-020-09322-9.

Dreyfus, S. E., & Dreyfus, H. L. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York: The Free Press.

Muchinsky, P. M. (2003). *Psychology applied to work.* Melmont, CA: Wadsworth/Thomson Learning.

Phillips, J. K., Ross, K. G., Rivera, I. D., & Knarr, K. A. (2013). Squad leader mastery: A model underlying cognitive readiness interventions. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*. NTSA.

Phillips, J. K., Ross, K. G., & Rosopa, P. J. (2017). Assessment instruments in support of Marine instructor development. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*. NTSA.

Phillips, J. K. and Ross, K. G. (2020). Developing mastery models to support the acquisition and assessment of expertise. In P. Ward, J. M. Schraagen, J. Gore, and E. Roth, *The Oxford Handbook of Expertise* (pp. 312-332). New York: Oxford University Press.

Riggio, R.E. (2000). *Introduction to industrial/organizational psychology*. Upper Saddle River, NJ: Prentice Hall.