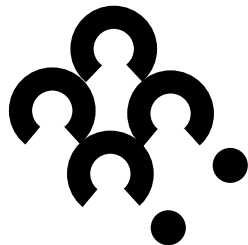

*THE DEVELOPMENT OF TASK-LEVEL
PERFORMANCE STANDARDS FOR APPRENTICE,
JOURNEYMAN, AND MASTER SKILL LEVELS FOR
AEROGRAPHER'S MATE (AG) JOBS IN THE U. S.
NAVY*

Technical Report No. 527

Prepared by:

Jerry W. Hedge, Walter C. Borman, Ulf C. Kubisiak, and Mark J. Bourne

January 2006



Personnel Decisions Research Institutes, Inc.
100 South Ashley Drive, Suite 375
Tampa, FL 33602
Phone: (813) 229-6646
www.pdri.com

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE JAN 2006		2. REPORT TYPE		3. DATES COVERED 00-00-2006 to 00-00-2006	
4. TITLE AND SUBTITLE The Development of Task-Level Performance Standards for Apprentice, Journeyman, and Master Skill Levels for Aerographer's Mate (AG) Jobs in the U.S. Navy				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Personnel Decisions Research Institutes, Inc., 100 South Ashley Drive, Suite 375, Tampa, FL, 33602				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 28	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

Table of Contents

Project Overview.....	1
Introduction	1
Objective 1: Identify/Clarify AG Level 1 Tasks	2
Objective 2: Identify Performance Standards across Skill Levels.....	3
Objective 3: Generate Task-Level Data to Support Short- and Long-term Planning Efforts by the AG Community	3
Section 1: Conduct Job/Task Analysis	5
Step 1: Generate Comprehensive Task List.....	5
Step 2: Develop Job/Task Analysis Questionnaire (JTAQ) and Collect Job Analysis Data.....	5
Construct Task List	5
Identify Measurement Variables and Develop Rating Scales.....	6
Finalize JTAQ.....	7
Step 3: Administer JTAQ and Collect Job Analysis Data	7
Step 4: Analyze JTAQ Data.....	9
Descriptive Statistics.....	9
Reliability of the Responses	9
Section 2: Establish Performance Standards.....	11
Step 1: Develop Methods for Establishing Performance Standards	11
Expert Judgment Task	11
Consensus Workshop.....	12
Step 2: Conduct Standard Setting Expert Judgment Task and Collect Data	13
Step 3: Analyze Expert Judgment Task Data	14
Descriptive and Reliability Statistics	14
Step 4: Conduct Performance Standards Consensus Workshop.....	14
Step 5: Finalize Performance Standards	14
Section 3: Identify Task Difficulty Values for All Level One Tasks	15
Step 1: Develop Task Difficulty Questionnaire (TDQ)	15
Identify Measurement Variables and Develop Rating Scales.....	15
Finalize TDQ.....	16
Step 2: Administer TDQ and Collect Task Difficulty Data	16
Step 3: Analyze TDQ Data	17

Descriptive and Reliability Statistics	17
Section 4: Provide Task-Level Data to Support Short-and Long-Term Planning Efforts by AG Community	19
Section 5: Summary of Findings	21
References	23

List of Tables

Table 1: Demographics of JTAQ Participants	8
Table 2. Intraclass Correlations for Job Analysis Ratings (Multiple-Rater Agreement).....	9
Table 3: Demographics of Standard Setting Expert Judgment Task Participants.....	13
Table 4: Demographics of Task Difficulty Participants	16

Project Overview

Introduction

Over the last several years, the Navy has been developing new strategies for revolutionizing Navy strategic human capital management. This includes a renewed emphasis on approaching a Sailor's career as a lifelong learning continuum gauged toward producing motivated and well-trained Sailors. This perspective suggests that everyone has the capability to grow in some way, at anytime during their career. Thus, an important component of strategic human capital management is understanding and helping each individual realize that potential. At the heart of this initiative is what is known as "The Sailor Continuum," that incorporates five distinct vectors forming the foundation on which the Navy identifies the knowledge, skills, and abilities (KSAs) that Sailors need to succeed in today's Navy. This five-vector model encompasses: professional development, personal development, leadership, certifications and qualifications, and job performance.

This broad vision sees the Sailor Continuum as an executable career roadmap and resume that precisely maps and measures an individual's career progress and identifies learning resources that lead to achieving career milestones, as the Sailor moves through the recruit, apprentice, journeyman, and master career levels. By allowing Sailors to identify and diagram different career paths, they will be better equipped to apply for educational opportunities and future duty assignments. The Sailor Continuum, then, and its five vector model, provide the conceptual framework for designing a process of clearly defined career paths with milestones, that when achieved, lead to career advancement.

Because people are an indispensable element in mission accomplishment, and play a critical role in determining the organization's performance capabilities, it is crucial that the Navy develop systems that enhance their ability to understand who is adding the different kinds of value, and who has the potential to add future value. An ability to both establish performance expectations, and differentiate between proficiency levels, makes it possible for investments in people development to be planned and managed intelligently.

Establishing standards of performance is not a new concept or process; standards exist whether they are discussed or put in writing. When a supervisor views an employee's performance he/she usually makes a judgment about whether that performance is acceptable. How a decision is made about what is acceptable or unacceptable is the philosophy behind establishing performance standards. Standards identify a baseline for measuring performance. Effective performance standards serve as an objective basis for communicating about performance, inform employees of expectations about job performance, and enable employees to differentiate between acceptable and unacceptable results.

Bernardin and Beatty (1984) defined performance standards as levels of performance corresponding to predetermined levels of effectiveness. Bobko and Colella (1994) outlined three components of a standard. Standards: 1) often have an evaluative component; 2) are criteria which are established externally, and imposed on an individual's work task; and 3) are usually considered to remain somewhat stable over time and individuals.

Essentially, then, performance standards are management-approved expressions of the performance threshold, requirement, or expectation that employees must meet to be certified at particular levels of performance. Performance standards are the foundation for a sound evaluation process. Employees must know what is expected of them, and to what degree they will be held accountable for the standards that have been established for their job. In turn, when a minimum performance standard is not reached, a demand signal will be sent for some form of performance remediation. By establishing such a system, the Navy can ensure that they develop the right Sailor, at the right place, at the right time, by accessing and driving "real-time" educational requirements based on actual Sailor performance.

Because the Aerographer's Mate (AG) community was eager to establish performance standards for their Enlisted Sailors at all skill levels – apprentice, journeyman, and master – and by so doing, enhance their ability to gauge current training needs and proficiency levels, a meeting was held in the third quarter of FY2005. Representatives of the AG Professional Development Center (PDC) met with CDR Mark Bourne (Human Performance Center) and project staff from Personnel Decisions Research Institutes, Inc. (PDRI) to discuss research needs and expectations. As a result of this meeting, three primary objectives were established to guide project activities.

While the central purpose of the effort was to design a method for developing task-based performance standards by skill level, use of an up-to-date task inventory was essential. Consequently, the first objective was to verify that a current task list existed, or could be produced in a timely fashion. Once a verified set of tasks was agreed upon, we would be able to then address our second objective – establishing performance standards for each task for apprentice, journeyman, and master skill levels. An important third objective was the design and delivery of task-based data sets and spreadsheets accessible to the AG community to support their short- and long-term planning efforts. Each of these objectives is discussed in further detail below.

Objective 1: Identify/Clarify AG Level 1 Tasks

The AG community has been undergoing significant transformation, as they re-align their business lines and functional activities. Consequently, it was important to examine recent task inventories, and job/task analyses to identify a set of tasks performed by Enlisted personnel representative of all operating directorates. Therefore, we worked with AG representatives to identify, compile, and revise existing Level 1 tasks performed across the Enlisted career field. Once

a final set of Level 1 tasks was identified, and confirmed by subject matter experts (SMEs), we developed a job/task analysis questionnaire (JTAQ) that, once administered, would allow us to establish which tasks were indeed performed within and across jobs, skill levels, and directorates. The JTAQ was administered to a representative sample of AG Enlisted personnel from the various jobs, skill levels, and directorates, and the inventory responses were used to describe the AG Enlisted career field quantitatively in terms of *critical* Level 1 tasks.

Objective 2: Identify Performance Standards across Skill Levels

The second objective of the project was to establish performance standards for all tasks representative of apprentice, journeyman, and master skill levels. With successful completion of work on Objective 1, a target set of tasks existed for which to develop performance standards. We used both an on-line expert judgment task and consensus workshop methodology to gather information from highly experienced AG SMEs. The expert judgment task solicited independent judgments of where performance standards should be established for each task within each skill level. Subsequent data analysis identified a high degree of rater agreement on these standards; only 8% (111 out of 1,377 task ratings) of the tasks had sufficient disagreement to warrant further discussion. This was accomplished in the consensus workshops, where all SMEs, as a group, discussed each task where potential disagreement existed, and reached agreement on a final set of performance standards. In this way, a final set of performance standards was established for task-level performance expectations across skill levels.

A related effort within this second objective was to create a set of task difficulty values associated with each Level 1 task. This was undertaken to provide additional information about task characteristics that would be useful in later stages of the performance standards work, as well as for AG strategic planning purposes. A relative task difficulty questionnaire was administered to a group of senior AG leaders, and task difficulty values were produced for each of the 459 tasks.

Objective 3: Generate Task-Level Data to Support Short- and Long-term Planning Efforts by the AG Community

The third primary objective of our effort was to provide the AG community with task-level data in a format that would allow them to answer a variety of career field-specific questions. Successful completion of Objective 1 and 2 produced task-level time spent, importance, criticality, difficulty, and performance standards data sets. These were placed in formats and spreadsheets easily retrieved and able to be manipulated in support of short-and long-term planning efforts.

The remainder of this report provides a detailed look at the methods and procedures used to conduct the job/task analysis, establish performance standards for apprentice, journeyman, and master tasks, generate task difficulty data for these tasks, and assemble and organize task-based data for use by the AG community. The report includes the following five sections:

- Section 1: Conduct Job/Task Analysis.
- Section 2: Establish Performance Standards.
- Section 3: Identify Task Difficulty Values for Level 1 Tasks
- Section 4: Provide Task-Level Data to Support AG Planning Efforts
- Section 5: Summary of Findings.

Section 1: Conduct Job/Task Analysis

The development and administration of the job analysis questionnaire involved the following steps:

- Step 1: Generate Comprehensive Task List
- Step 2: Develop Job/Task Analysis Questionnaire (JTAQ).
- Step 3: Administer JTAQ and Collect Job Analysis Data.
- Step 4: Analyze JTAQ Data

Each of these steps is discussed in further detail below.

Step 1: Generate Comprehensive Task List

Our first step was to create a preliminary task list that described the activities performed in AG Enlisted jobs. For this step, we worked closely with personnel from the AG Professional Development Center (PDC). Because the PDC has extensive staff expertise and close contact with AG facilities throughout the world, identifying preliminary tasks was relatively straightforward. The starting point was use of any task lists generated within the last 3 – 5 years. This included a list of Level 1 tasks previously generated by SkillsNet, as well as other more recent and specific lists (i.e., directorate-unique tasks) that should be considered.

These lists were compiled and edited to a common format. They were then reviewed by PDC personnel for accuracy and coverage of jobs and directorates. These actions resulted in a list of 459 Level 1 tasks that described, in a comprehensive manner, the AG Enlisted career field.

Step 2: Develop Job/Task Analysis Questionnaire (JTAQ) and Collect Job Analysis Data

Construct Task List

While the tasks generated in the previous step were considered comprehensive, they were compiled and reviewed by a relatively small number of SMEs, and the information collected did not provide evidence as to the relative criticality of the different tasks for the various jobs, skill levels, and directorates. To ensure that the tasks did comprehensively describe the work performed in the AG Enlisted career field, we constructed a Job/Task Analysis Questionnaire (JTAQ). The JTAQ was prepared by listing all of the 459 tasks identified in Step 1. In addition,

to facilitate the rating process, these tasks were clustered under logical groupings to ease completion of the task ratings.

Identify Measurement Variables and Develop Rating Scales

The next step in developing the JTAQ was to identify the variables to measure and the rating scales to use for this measurement. A basic task variable that is typically operationalized in job analysis surveys is task importance. We chose to include a rating of this variable, defining importance as how important the activity is for successful performance of the job. Another way to consider importance is in terms of the consequences that would result if the task was not performed correctly – the more severe the consequences, the more important the task.

We have found in past research that a five-point rating scale works well for measuring importance (and for most other scaled judgments). Thus, we used the following rating scale for task importance:

- 1 = Minor Importance
- 2 = Some Importance
- 3 = Important
- 4 = Very Important
- 5 = Extremely Important

This scale measures absolute importance. Tasks are compared to a standard (the consequences associated with not performing them correctly, or centrality to mission), rather than to each other. We have found that this type of scale works extremely well for many types of jobs.

However, importance alone does not fully describe a task's criticality for a job. There are tasks that are not as important as others, but which are performed more frequently or on which more time is spent. We therefore included a time variable in the JTAQ.

We have found in past research that relative time spent is a useful scale for this purpose. Relative time spent ratings compare the total time spent performing a task to the time spent on all other tasks on the job. The focus on relative time spent is important because respondents find it quite difficult to estimate time according to any precise metric (e.g., how many hours per day do you spend on this task?).

The following scale was used to rate relative time spent:

- 1 = Much less time than most other tasks
- 2 = Less time than most other tasks
- 3 = About the same time as most other tasks
- 4 = More time than most other tasks
- 5 = Much more time than most other tasks

The time spent and importance ratings were also used to compute a composite variable that summarizes the “criticality” of each task for each job. This was done by first multiplying the importance rating by two, then adding the time spent rating, and finally dividing the resulting number by three. This essentially gives the importance rating twice as much weight as the time spent rating, and this index has been found useful in previous work as an overall summary of the information contained in these rating scales.

Finalize JTAQ

Once the tasks and rating scales had been prepared, the final step in questionnaire development was preparation of survey instructions, inclusion of a demographics section, and final formatting. Organization of the final survey consisted of a set of general instructions describing the purpose of the study, the requirements associated with completing the questionnaire, and the specific sections to be completed. This was followed by a brief series of background/experience items that allowed us to characterize the workshop participants, then specific instructions for the rating task, and finally the rating task itself.

Specific instructions asked participants to scan the entire task list before starting, and then answer three questions about the tasks. They were asked to decide whether the task was part of their job. Then, for tasks identified as part of their job, they were asked to rate the importance of the task for successful performance of their job, and to rate the amount of time spent on the task relative to other job tasks.

Step 3: Administer JTAQ and Collect Job Analysis Data

Our next step was to administer the JTAQ to a representative sample of AG Enlisted personnel. Because our goal was not only to verify a final set of Level 1 tasks applicable to the Enlisted Meteorological and Oceanographic Services (METOC) community, but to generate task criticality data that would allow the AGs to establish task coverage at the job, skill, and directorate levels, it was important to include representation from all jobs, skill levels, and directorates.

In order to successfully accomplish this data collection effort, we conducted job analysis workshops at METOC facilities in Norfolk, VA, San Diego, CA, and Pearl Harbor, HI over a three-month period. Once workshops at these locations were completed, a preliminary review of participants revealed less than desired numbers in certain jobs within directorates. Working with PDC personnel, we were able to request and receive additional survey completion to supplement the surveys collected at the workshops.

In all, 151 participants completed the JTAQ; 80% were male, 77% were White, with an average tenure in the Navy of almost 10 years. Table 1 presents a more detailed breakdown of demographic data for the JTAQ workshops.

Table 1: Demographics of JTAQ Participants

	<u>N</u>	<u>%</u>
<u>Gender</u>		
Male	120	79.5
Female	31	20.5
<u>Race/Ethnicity</u>		
White	116	76.8
Black/African-American	20	13.2
Spanish/Hispanic/Latino	7	4.6
Asian	2	1.3
Native Hawaiian/Pacific Islander	1	0.7
American Indian/Alaska Native	1	0.7
Other	4	2.6
<u>Current AG Job</u>		
METOC Manager	25	16.6
Meteorological Forecaster	53	35.1
METOC Technician	41	27.2
Missing Data/Other ¹	32	
<u>Directorate</u>		
Maritime Safety	35	23.2
Intelligence Surveillance Recon.	9	6.0
Naval Special Warfare	8	5.3
Aviation Safety	18	11.9
Mine Warfare	30	19.9
Anti-Submarine Warfare	14	9.3
Fleet Operations	27	17.9
Navigation	9	6.0
Missing Data	1	
<u>Tenure in Current Paygrade</u>		
Less than 12 months	31	20.5
12 months to less than 18 months	23	15.2
18 months to less than 24 months	9	6.0
24 months or more	88	58.3
<u>Current Paygrade</u>		
E2 – E4	32	21.2
E5 – E6	68	45.1
E7 – E9	11	7.4
O1 – O5	9	6.0
Missing Data/Other ¹	31	
<u>Skill Level</u>		
Apprentice	42	27.8
Journeyman	76	50.3
Master	30	19.9
Missing Data	3	

¹ Because Mine Warfare has not yet been fully staffed with military personnel, civilian job incumbents performing the same job responded to the survey, but did not identify themselves as holding military jobs or paygrades (N = 29).

Note. (N=151)

Step 4: Analyze JTAQ Data

Descriptive Statistics

In order to verify coverage of the task domain, JTAQ data were analyzed using traditional descriptive techniques. Examination of the data included frequency counts, means, and standard deviations across all 459 Level 1 tasks. Overall analyses of the descriptive data for the 459 tasks established that each of the tasks was performed within the AG Enlisted community. Additional analyses were performed in order to link specific tasks to jobs, skill levels, and directorates. These more detailed breakouts were then organized for use by the AG community, as described in Section 4.

Reliability of the Responses

Interrater reliability reflects the extent to which participants agreed in their ratings. We estimated interrater reliabiliies using formulas presented by Shrout and Fleiss (1979). Interrater reliabilities of time spent, task importance, and criticality were calculated across the 459 tasks, as well as for jobs, skill levels and directorates. Results are shown in Table 2.

As can be seen in Table 2, mean reliabilities across all raters for these three scales were quite high (.95, .95, .94). This pattern of reliabilities was also consistently high across jobs, skill levels, and directorates; in fact, the latter reliabilities would have been higher than the Overall reliabilities if as many participants (i.e., 151) had provided ratings.

Table 2. Intraclass Correlations for Job Analysis Ratings (Multiple-Rater Agreement)

Sample Breakdown	Category	Number of Raters	Time Spent	Importance	Criticality
Overall	---	151	0.95	0.95	0.94
By Job	METOC Manager	25	0.68	0.68	0.66
	Meteorological Forecaster	43	0.95	0.95	0.94
	Oceanographic Forecaster	10	0.82	0.82	0.83
	METOC Technician	41	0.86	0.87	0.87
By Skill Level	Apprentice	42	0.79	0.80	0.80
	Journeyman	76	0.93	0.94	0.93
	Master	30	0.74	0.72	0.74
By Directorate	Maritime Safety	35	0.93	0.93	0.94
	Intelligence Surveillance Reconnaissance	9	0.84	0.81	0.50
	Naval Special Warfare	8	0.85	0.78	0.70
	Aviation Safety	18	0.91	0.92	0.92
	Mine Warfare	30	0.88	0.90	0.89
	Anti-Submarine Warfare	14	0.87	0.89	0.89
	Fleet Operations	27	0.87	0.88	0.86
	Navigation	9	0.97	0.99	0.99

Section 2: Establish Performance Standards

The development of performance standards involved the following steps:

- Step 1: Develop Methods for Establishing Performance Standards.
- Step 2: Conduct Standard Setting Expert Judgment Task and Collect Data.
- Step 3: Analyze Expert Judgment Task Data
- Step 4: Conduct Performance Standards Consensus Workshop
- Step 5: Finalize Performance Standards

Each of these steps is discussed in further detail below.

Step 1: Develop Methods for Establishing Performance Standards

Work reported in Section 1 resulted in final identification of the task domain for the AG Enlisted career field. This task list, then, provided the foundation for developing performance standards. In addition, because performance expectations were different depending on skill level, it was also necessary to identify task-level performance standards for apprentices, journeymen, and masters.

Because these performance standards needed to reflect task-level performance requirements that were endorsed by the AG community, and would be applicable across directorates, it was critical that the methodology used rely on input from highly experienced AGs, with broad career field expertise. Consequently, we designed a two-stage process to capture SME expert judgment and produce these task-level performance standards. The first stage required SMEs to identify a performance standard for each task and skill level. The second stage allowed group discussion of rating disagreements across raters and sought consensus on performance standards for those tasks.

Expert Judgment Task

Our first requirement in this standard setting expert judgment task was to decide how to solicit judgments from participants, and the rating scale to use to gather their judgments. We chose an on-line survey method to allow our SMEs to provide independent ratings that could be quickly tabulated and merged with all other ratings in preparation for data analysis that would identify rating discrepancies among raters.

We chose to develop a 10-point proficiency rating scale that allowed raters to make judgments about their performance expectations for an apprentice, journeyman, and master on each of the 459 Level 1 tasks. The rating scale developed for this purpose is shown below:

- 10 - Outstanding; perfectly executed.
- 9
- 8
- 7 - Good; well executed.
- 6
- 5
- 4 - Executed fairly poorly but small adverse consequences.
- 3
- 2
- 1 - Very poorly done or done incorrectly.

Once the tasks and rating scales had been prepared, the final aspect of survey development was preparation of expert judgment task instructions, inclusion of a demographics section, and final formatting. Organization of the final expert judgment task consisted of a set of general instructions describing the purpose of the study, and the requirements associated with completing the task. This was followed by a brief series of background/experience items that allowed us to characterize the characteristics of the participants, then specific instructions for the expert judgment task, and the judgment task itself.

Specific instructions asked participants to estimate the minimally acceptable level of performing a task for an apprentice, a journeyman, and a master AG. Procedurally, then, the expert judgment task required the SMEs to 1) read the task statement for the apprentice skill level, 2) decide where on the scale should be designated as representative of minimally acceptable proficiency for this skill level on this task, 3) check the appropriate point on the rating scale, and then proceed to the journeyman and master levels and follow the same procedure. In this way, the rater provided a rating for apprentice, journeyman, and master for each task considered applicable for that skill level.

Consensus Workshop

Because it was unlikely that all SMEs would agree on performance standards for each task within skill levels, it was necessary to design a second stage that facilitated reaching consensus where disagreements existed. We chose to use a group consensus process that provided anonymous feedback to the group about ratings on targeted tasks (where discrepancies existed), and asked participants to discuss why these ratings might have been different, and ask them to work toward selection of a performance standard on which all members of the workshop could agree. For this process to be successful, it was necessary to quickly capture, merge, and analyze all data from the expert judgment task, and isolate tasks where disagreement existed.

Step 2: Conduct Standard Setting Expert Judgment Task and Collect Data

Our next step was to conduct the expert judgment task with a representative sample of highly experienced AG personnel. Because our goal was to generate standards applicable for all directorates in the career field, it was essential to include representation from all directorates.

We conducted an expert judgment workshop in San Diego, CA. The workshop was conducted in an Electronic Board Room equipped with networked computers, shareware, and overhead projector capabilities. Participants received general instructions from the session facilitator, and then independently completed the expert judgment task. Because performance standard ratings were required for 1,377 task statements (459 x 3 skill levels), the lists were divided into two parts, with one half being completed by 7 SMEs and the other half completed by the other 6 SMEs. Care was taken to insure that directorate-level coverage existed in both expert judgment groups.

In all, 13 participants completed the expert judgment task; 77% were male, 92% were White, with an average tenure in the Navy of over 18 years. Table 3 presents a more detailed breakdown of demographic data for this workshop.

Table 3: Demographics of Standard Setting Expert Judgment Task Participants

	<u>N</u>	<u>%</u>
<u>Gender</u>		
Male	10	76.9
Female	3	23.1
<u>Race/Ethnicity</u>		
White	12	92.3
Black/African-American	1	7.7
<u>Tenure in Current Paygrade</u>		
Less than 12 months	2	15.4
24 months or more	11	84.6
<u>Current Paygrade</u>		
E6	1	7.7
E7	6	46.2
E8	3	23.1
E9	2	15.4
O3	1	7.7
<u>Skill Level</u>		
Apprentice	0	0.0
Journeyman	1	7.7
Master	12	92.3
Note. (N=13)		

Step 3: Analyze Expert Judgment Task Data

Descriptive and Reliability Statistics

Independent ratings for all expert judgment task participants were merged and analyzed in order to assess agreement about performance standards and to identify tasks where sufficient discrepancies existed to warrant discussion in the consensus workshop. Overall, rater agreement across the 459 tasks was quite high, resulting in a mean reliability value of .94, again using the Shrout and Fleiss (1979) formula. In addition, a closer examination of the data included frequency counts, means, standard deviations, and minimum/maximum values across all 459 Level 1 tasks for apprentice, journeyman, and master skill levels.

Judgment heuristics for selecting tasks for subsequent review included: large standard deviation (> 2.0); outliers (extreme outliers were excluded from future analyses); inconsistencies across apprentice, journeyman, and master levels within tasks (i.e., higher standards for lower skill levels). Where significant agreement was found, the mean value, rounded to the nearest whole number, was established as the performance standard value. Where sufficient discrepancies in ratings were found, task data were extracted and included in a spreadsheet for discussion at the consensus workshop. In all, 1266 of the 1377 tasks required no further examination (92%). The remaining 111 tasks were readied for further discussion.

Step 4: Conduct Performance Standards Consensus Workshop

The same individuals who had participated in the expert judgment task the previous day again met as a group to discuss the 111 tasks that required continued discussion to establish agreed-upon performance standards. Using the Electronic Board Room, we described the results from the expert judgment study analyses, and the objective of the current exercise. Then, projecting summary data (e.g., mean, minimum, maximum values) from the analyses on the overhead screen, we facilitated discussion of each task, one by one. If previously agreed-upon standards were available for other skill levels for the task under discussion, we also provided that information as a frame-of-reference. Lively discussion ensued, but in every instance, SMEs arrived at an agreed upon performance standard for a task before proceeding to the next task.

Step 5: Finalize Performance Standards

After consensus was reached on the proper performance standard for the 111 tasks where individual disagreement existed, these tasks (and their associated performance scores) were combined with the 1266 tasks and standards identified in the expert judgment task. In this way, a final set of performance standards was established for task-level performance expectations across skill levels. Final performance standards for all 459 tasks for each of the three skill levels are provided in an addendum to this report.

Section 3: Identify Task Difficulty Values for All Level One Tasks

The development and administration of the task difficulty questionnaire involved the following steps:

- Step 1: Develop Task Difficulty Questionnaire (TDQ).
- Step 2: Administer TDQ and Collect Task Difficulty Data.
- Step 3: Analyze TDQ Data

Each of these steps is discussed in further detail below.

Step 1: Develop Task Difficulty Questionnaire (TDQ)

Identify Measurement Variables and Develop Rating Scales

To gain a more complete understanding of the AG career field, it is useful to gather as much information about the task characteristics that comprise the career field as feasible. Therefore, in addition to the time spent and importance information already obtained, we gathered information related to task difficulty.

Task difficulty has been used for many years, especially within the Air Force's occupational measurement program, for such diverse purposes as prioritizing task training and setting enlistment standards (Boyce & Gould, 1996). Numerous studies have demonstrated that senior-level personnel can achieve high levels of agreement when rating the "learning difficulty of tasks" (see Burtch, Lipscomb, & Wassman, 1982).

We chose to use a similar approach in our work, defining task difficulty as "the amount of time needed to learn to do a task satisfactorily." In addition, because Christal (1974) found that supervisors could not agree on the absolute time it takes to learn to perform tasks, but could agree on relative time estimates, we used the following scale to rate relative time spent:

- 1 = Much less difficult than most other tasks
- 2 = Less difficult than most other tasks
- 3 = About the same difficulty as most other tasks
- 4 = More difficult than most other tasks
- 5 = Much more difficult than most other tasks

Finalize TDQ

Once the tasks and rating scales had been prepared, the final component of questionnaire development was preparation of survey instructions, inclusion of a demographics section, and final formatting. Organization of the final survey consisted of a set of general instructions describing the purpose of the study, the requirements associated with completing the questionnaire, and the specific sections to be completed. This was followed by a brief series of background/experience items that allowed us to summarize the characteristics of the participants, then specific instructions for the rating task, and finally the rating task itself.

Specific instructions defined task difficulty as the amount of time needed to learn to do each task satisfactorily. Then participants were asked to develop a frame of reference by scanning the entire listing of tasks. They were encouraged to pick out some of the easy tasks and some difficult tasks, and then to find some tasks which fell between these extremes that were of average difficulty. They were to use the tasks at or near the middle of the range as a reference point for judging the difficulty of all tasks in the inventory. Once they had done this, they were asked to estimate the task difficulty rating for each task compared with other tasks in this inventory.

Step 2: Administer TDQ and Collect Task Difficulty Data

Our next step was to administer the TDQ to a representative sample of AG Enlisted personnel. Because our goal was to establish task difficulty values for each of the Level 1 tasks across all jobs, skill levels, and directorates, it was important to include representation from all directorates.

We conducted a task difficulty workshop in San Diego, CA. In all, 9 senior-level SMEs, with broad career-field expertise completed the TDQ; 78% were male, 89% were White, with an average tenure in the Navy of more than 18 years. Table 5 presents a more detailed breakdown of demographic data for the JTAQ workshops.

Table 4: Demographics of Task Difficulty Participants

	<u>N</u>	<u>%</u>
<u>Gender</u>		
Male	7	77.8
Female	2	22.2
<u>Race/Ethnicity</u>		
White	8	88.9
Black/African-American	1	11.1
<u>Tenure in Current Paygrade</u>		
Less than 12 months	2	22.2
24 months or more	7	77.8

Table 4: Demographics of Task Difficulty Participants (cont'd)

	<u>N</u>	<u>%</u>
<u>Current Paygrade</u>		
E6	1	11.1
E7	4	44.4
E8	2	22.2
E9	1	11.1
O3	1	11.1
<u>Skill Level</u>		
Apprentice	0	0.0
Journeyman	1	11.1
Master	8	88.9
Note. (N=9)		

Step 3: Analyze TDQ Data

Descriptive and Reliability Statistics

Means and standard deviations were computed for each of the 459 Level 1 tasks. In addition, interrater reliabilities were calculated across the 459 tasks using formulas by Shrout and Fleiss (1979), and showed high levels of agreement across tasks (mean = .92). As a result of these analyses, mean values were adopted as the task difficulty values for all Level 1 tasks.

These task difficulty values are included in the addendum to the report.

Section 4: Provide Task-Level Data to Support Short-and Long-Term Planning Efforts by AG Community

In our initial planning meeting with AG representatives, prior to beginning work on the project, emphasis was placed on the importance to the AG community of data collected during the course of our work being available for their use. Certainly, the ability to examine – and even manipulate – data files could provide a wide variety of immediately useful insights for the AG community. At a minimum, these data should help to support personnel management and training decisions, both short- and long-term.

Consequently, data gathered at every stage of the project were accumulated in user-friendly data bases and spreadsheets, that with a minimum of training would be easy to examine, manipulate, and re-examine from a variety of different perspectives. For example:

- Time spent data could be compared to importance data for each task;
- Criticality data could be sorted to identify the most critical tasks in the career field;
- Tasks comprising the Aviation Safety directorate could be compared to Fleet Operations tasks;
- Apprentice tasks within the Maritime Safety directorate could be compared to journeyman tasks in that directorate;
- The most difficult tasks in the AG Enlisted career field could be isolated for additional study;
- Performance standard scores could be sorted to reveal those that were identified as having highest or lowest performance expectations for each skill level

Many other combinations and permutations are possible; the point is that the AG community will be able to use these data sets and spreadsheets to support various career field initiatives as their needs arise. Consequently, job/task analysis, task difficulty, and performance standards data sets and spreadsheets were delivered in conjunction with this final report.

Section 5: Summary of Findings

The primary objective of the project was to establish performance standards for tasks representative of apprentice, journeyman, and master skill levels. In order to satisfy this objective, we first had to identify a set of tasks performed by Enlisted personnel representative of all operating directorates. Once a final set of Level 1 tasks was identified, a job/task analysis questionnaire was administered to a representative sample of AG Enlisted personnel from the various jobs, skill levels, and directorates. In this way, we verified the current task domain for the AG Enlisted career field.

With successful completion of the job/task analysis work, a target set of tasks existed from which to develop performance standards. Then, we used expert judgment and consensus workshop methodologies to gather information from highly experienced AG SMEs to establish performance standards for each task for all three skill levels. This resulted in a final set of performance standards established for task-level performance expectations across skill levels.

We also generated a set of task difficulty values associated with each Level 1 task. This was undertaken to provide additional task information that would be useful in later stages of the performance standards work and for AG strategic planning purposes. A relative task difficulty questionnaire was administered to a group of senior AG leaders, and task difficulty values were produced for each of the 459 tasks.

Finally, we organized all data collected during the study, and placed the data in formats and spreadsheets that could be easily retrieved and manipulated. This will allow the AG community to answer a variety of career field-specific questions, and support of short- and long-term planning efforts. The AG community has been undergoing significant transformation as they re-align their business lines and functional activities. The work reported here will help the AGs not only to keep pace with this transformation, but to manage these changes proactively with data-driven insights.

References

- Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston, MA: Kent Publishing.
- Bobko, P., & Colella, A. (1994). Employee reactions to performance standards: A review and research propositions. *Personnel Psychology, 47*, 1-29.
- Boyce, L. A., & Gould, R. B. (1999). *Task difficulty measurement by the United States Air Force Occupational Measurement Program (AL/HR-TR-1996-0151)*, Brooks Air Force base: United States Air Force Research Laboratory, Human Effectiveness Directorate.
- Burtch, L. D., Lipscomb, M. S., & Wassman, D. J. (1982). *Aptitude requirements based on task difficulty: Methodology for evaluation (AFHRL-TR-81-34)*. Brooks Air Force Base: Air Force Human Resources Laboratory.
- Christal, R. E. (1974). *The United States Air Force occupational research project (AFHRL-TR-73-75)*. Brooks Air Force Base: Air Force Human Resources Laboratory.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

