

Task difficulty is a complicated concept and defining it is not a simple matter (Christal, 1974; Madden, 1960). A task may be characterized as difficult because it involves one or more of the following conditions (Madden, 1962).

1. The incumbent was not properly trained to perform the task and this training is not available on the job.
2. Performance standards do not exist or are not clearly defined.
3. Proper tools and equipment are not available or are hard to obtain.
4. The task has emotional aspects which are unpleasant to the incumbent.
5. There is some attribute of the task itself which makes it difficult. It may require intense concentration or vigorous physical exertion.
6. There is some attribute of the environment in which the task is performed which makes it difficult. There may be excessive noise or a requirement for extensive cooperation with other workers.
7. The incumbent does not have the ability to perform the task easily even though other workers generally consider it an easy task.

Difficulty as a task attribute may be defined in as many ways. For instance, the definition may be in terms of the amount of training or experience required to perform a task; it may refer to mental or physical effort; or difficulty may mean complexity or monotony, or something else.

It appears that defining difficulty has presented a general problem of critical significance. One approach to dimensionality of difficulty is to ask incumbents to describe specific tasks that are difficult and then to explain why they are difficult. This method was used in a preliminary study conducted by Madden (1960, 1962). Madden found that when incumbents were asked to list tasks which they found difficult to perform or to learn and to state the reasons why they were difficult, reasons given could be classified into 10 separate categories. Following are the ten categories which resulted:

1. training
2. interpersonal relations
3. frustration
4. pressure
5. attention
6. aptitude
7. working conditions
8. forms
9. regulations, technical publications, manuals
10. unclassified

Examples of reasons tasks were categorized as difficult due to "training" included individuals did not receive any training, did not receive enough training, or changes were occurring so rapidly that training was never completed. The "interpersonal relations" category involved reasons associated with difficulty in receiving or giving supervision, relations with peers, and coordination. The "frustration" category included difficulty reasons relating to lack of proper tools, not enough information, unavailable parts or publications, and waiting. Tasks categorized as "pressure" difficulty included reasons such as insufficient time to do the work properly, responsibility for expensive equipment, or rapid changes that require a great deal of adaptability. "Attention" difficulty type tasks included work that required precision, careful naming or labeling, or was very complicated. Tasks that were considered difficult due to "aptitudes" were a result of individuals lacking the appropriate abilities or skills to complete the task properly. An example might be a draftsman who is unable to draw. Climbing ladders or performing dirty work were examples of the "working conditions" category. "Forms" described the difficulties associated with filling them out, getting information for them, and avoiding errors. Difficulties associated with "regulations, technical publications, and manuals" included finding them, understanding and interpreting them, and conflicting interpretations. Finally, the "unclassified" category encompassed irrelevant statements that were not considered real reasons ("It's hard to do."), reasons which were extremely unusual and unlikely to constitute a component of difficulty having operational implications ("I do not speak English well."), and statements which did not seem to fit into any category in which there were at least one other statement.

Nine of these categories represented different definitions of difficulty, and the majority of them were more related to the work or environmental situation in general than to particular tasks. Basically, the "difficult to perform" reasons were those which could appear in any job and were not produced as a result of any peculiar characteristic of a particular job or task. Similarly, only a few "difficult to learn" reasons could be directly identified as derived from the properties of a task. For instance, insufficient training, interpersonal relations with supervisors and subordinates, frustrating factors (waiting for parts), working conditions, aptitude requirements, and completion of forms constituted a large part of all reasons given. Thus, reported difficulty may stem from environmental conditions, personal characteristics, or some factor inherent in the task itself unless the rating scales explicitly directs the rater's response to a specific domain of difficulty.

One way of using the difficulty attribute is to leave it undefined so the incumbent will identify tasks they judge to be difficult about their jobs, whatever the reason. Although Ammerman (cited in Morsh, Madden, & Christal, 1961), Cragun & McCormick (1967), and Madden (1960, 1962) did this with a small degree of success, difficulty is often perceived differently from task to task by the same individual. For instance, a given incumbent may consider task "A" to be difficult for reason "X," but task "B" may be perceived as difficult for reason "Y." Furthermore, interviews of job incumbents have revealed that the same task may be considered difficult for widely varying and unrelated reasons by different incumbents (Madden, 1960, 1962). Judgements and opinions collected using a global task difficulty concept offers little about the type and rationale behind an individual's rating. This information can be misleading and has minimal value. Defining the dimensions of difficulty clearly and accurately and collecting judgements and

opinions regarding specific components of difficulty provide more meaningful information having a higher utility.

Although many definitions were considered, task difficulty was initially defined in terms of mental difficulty and physical difficulty (McCormick & Tombrink, 1960). As Madden's (1960, 1962) research predicted, the test-retest reliability was relatively low and the inter-rater consistency indicated differences existed among incumbents in their perceptions of the difficulty of the various tasks.

During the late 1960's, the Air Force Occupational Program collected data using "Difficulty of Task Performance" and "Difficulty of Learning the Task" relative rating scales (Fugill, 1972; Morsh & Archer, 1967; Weeks, 1984). However, as initially discovered by Madden (1962), senior-level technicians did not perceive task performance difficulty when defined in terms of the difficulty of performing a task satisfactorily under normal conditions as an independent task property but rather as dependent on physical working conditions, experience, and interpersonal relations. As a result, senior-level technicians did not always agree on the relative difficulty of a given task. The alternative was to consider a definition which reflects the amount of time it takes for individuals to learn to perform a task adequately. Numerous studies demonstrated that senior-level technicians could achieve high levels of agreement when rating task on learning difficulty (Cragun & McCormick, 1967; Lecznar, 1971; Mead 1970a, 1970b; Mead & Christal, 1970). For example, Christal (1974) found that while supervisors could not agree on a time it takes for workers to learn to perform tasks; supervisors could agree that if other factors are held constant, workers can learn to perform some tasks faster or slower than other tasks. Thus, task-learning difficulty was defined as the time it takes to learn to perform a task satisfactorily (i.e., the higher the learning difficulty, the more time required to learn to perform the task). As a result, this definition was adopted for the purpose of obtaining judgments of task difficulty. Weeks' 1981 study further supported this task difficulty definition by concluding that "knowledgeable judges can reach high levels of agreement concerning the relative learning difficulty of work tasks when learning difficulty is defined in terms of learning time." Finally, Burtch et al. (1982), again provided evidence in support of the reliability and validity of task-level ratings of learning difficulty. The USAF's current definition of task difficulty as "the amount of time needed to learn to do a task satisfactorily" is supported by this 20-year stream of research.

Data Collection

The USAF method of collecting, analyzing, and reporting task difficulty is relatively complex. The essential rationale and research evidence upon which the existing USAFOMS method has been based are comprehensively reported by Morsh et al. (1961). Morsh and Archer (1967) set forth detailed procedures for collecting, organizing, analyzing, and reporting information describing Air Force jobs. Although the analysis routines have changed significantly during the past 30 years and still continue to evolve, the data collection procedures have remained very stable (Phalen et al., 1992).

The Air Force occupational analysis program is designed specifically for large scale administration and operational application. The basic data in the identification of tasks come from survey data routinely collected at USAFOMS. Initially, the job inventory

used in the periodic occupational surveys of active duty jobs are developed by creating a duty outline and a listing of task statements based on job descriptions, course training standards, and other published materials. Related tasks are then organized within duty categories and the task list is revised based on work-site observations of the job and input from technical specialists. When finalized, the job inventory (JI) is administered to a representative sample of job incumbents within a specialty to collect information about the relative amount of work-time spent on tasks which they perform, using a 9-point scale with descriptive bench marks ranging from "a very small amount" to "a very large amount." Specifically, job incumbents are asked to check the tasks they perform in their present job and then rate those tasks in terms of the relative amount of time spent on that task. Relative time spent means the total time spent doing the task compared with time spent on each of the other tasks performed in the present job. These data are compiled in a computer-generated job description to provide, among other information, an estimation of the percentage of incumbents who perform each task and the average percentage of time spent on each task by those in the specialty who perform it. This same information can be reported for any group of individuals who can be defined by available background variables such as Total Active Federal Military Service (TAFMS), grade, education, and time-in-job. The first two sections of a job inventory, a biographical section and a background section, are used to collect general information about the job incumbent and their job. A representative example of items collected in the biographical and background information data collection sections is located in Appendix A.

The same duty/task list is administered to approximately 75 senior NCOs, usually supervisors, who are asked to rate the tasks on training emphasis (TE), based on how much structured training is required for first-term personnel. Structured training includes training such as basic resident training, formal OJT, and first-term career development courses. TE raters first check tasks requiring structured training and then they recommend the amount of first-term training emphasis needed based on a 9-point scale ranging from "1" indicating extremely low training emphasis to "9" indicating extremely high training emphasis. TE data are important for the determination of initial skills training requirements by aiding decisions about which job tasks should be trained and to what degree.

Another 75 senior NCOs are asked to rate the same duty/task list on task difficulty, based on how much time is required to learn the tasks. The instructions for completing the task difficulty inventories are located in Figure 5. TD raters are asked to first develop a frame of reference for rating task difficulty by scanning the entire list of tasks. Then they are requested to estimate the task difficulty ratings for each task compared with other tasks in the inventory on a 9-point scale ranging from "1" indicating extremely low difficulty to "9" indicating extremely high difficulty. These ratings are used to compute an estimate of the task difficulty of each task compared with other tasks in the inventory. Christal (1974), Mead and Christal (1970), Ruck, Thompson, and Stacy (1987), and Ruck, Thompson, and Thompson (1978) determined task factor values could be reliably obtained from as few as 20 to 40 raters. However, USAFOMS historical records indicate approximately only 60 percent of task factor booklets are returned from a general administration. Accordingly, task difficulty is typically administered to 75 raters.

INSTRUCTIONS FOR RATING TASK DIFFICULTY

As a senior technician, you have been selected to provide needed information pertaining to the difficulty of tasks performed in your career ladder. This information will be of value to the Air Force in the improvement of training, classification, and testing programs. To accomplish this rating, follow the procedure listed below.

NOTE: To obtain the maximum response possible, it is requested that you rate each task of which you have any knowledge. Rate those tasks you presently perform or supervise, those tasks which you have performed at a prior time in your career, and those tasks which you have observed or supervised while being performed by others. Most personnel with your experience and background will be able to rate the majority of the tasks listed and in many cases to rate all of them.

STEP 1. Develop a frame of reference for rating task difficulty. For this survey, task difficulty is defined as the amount of time needed to learn to do each task satisfactorily. To develop a frame of reference, scan the entire listing of tasks. Pick out some easy tasks and some difficult tasks. Then, find some tasks which fall between these extremes that are of average difficulty. Use these tasks at or near the middle of the range as a reference point for judging the difficulty of all tasks in the inventory. This frame of reference will be used for completing STEP 2.

STEP 2. Estimate the task difficulty rating for each task compared with other tasks in this inventory. Use the scale shown here and at the top of each page to rate each task.

1. Extremely Low
2. Very Low
3. Low
4. Below Average
5. Average
6. Above Average
7. High
8. Very High
9. Extremely High

Begin with the first task in the booklet and give each task of which you have knowledge a difficulty rating from 1 to 9; record the value opposite the task statement in the column titled "TASK DIFFICULTY." Try to rate every task on each page. Remember (from STEP 1) that you are comparing each task with the other tasks in the career field.

STEP 3. The last page of the booklet is available to add any tasks you do now which are not listed. Your constructive suggestions in improving the job inventory will be useful.

STEP 4. Review the booklet to see that you have rated the DIFFICULTY of all tasks possible. Each task can be given only one rating.

Figure 5. Instructions for Rating Task Difficulty

Analysis

These data, once collected, are then analyzed using the Comprehensive Occupational Data Analysis Program. The CODAP programs serve as the basic analytical technique for presenting a job analyst with readily interpretable information on the content of Air Force jobs and specialties. However, the validity of these program's products obviously assume and require accurate input data. To the extent that a subject matter expert cannot provide specific reliable information concerning the correct type of difficulty, errors may occur in the resulting recommendations.

One means of minimizing the effects of inaccurate data is through a CODAP program which identifies and removes divergent raters. A divergent rater is one whose rating behavior demonstrates that the rater did not take the task seriously or one who unintentionally rated improperly, such as one who reverses the meaning of the rating scale. Once divergent raters are identified, they are examined to determine if there are any systematic similarities among them. Similarities may suggest the presence of multiple policies in the AFS. The reliability of a single rater is viewed for the expected correlation between that randomly chosen rater from a sample and another rater randomly chosen from the same sample. The reliability of a composite of raters is the expected correlation between the mean task value for a set of raters drawn from the sample and the average task means of an equivalent set of raters drawn at random from the population of raters from which the sample of raters was drawn. The CODAP GRPREL program tells how many raters of the same type in the sample would be needed to achieve a reliability of a composite of raters that would equal .90. Therefore, when reliability of a composite of raters equals or exceeds .90, it is determined there exists high interrater reliability among raters (Christal & Weismuller, 1976; Goody, 1976).

Another means of minimizing the effects of poor data involves a quality review of incoming data. Specifically the occupational analysts from the Airman Analysis Branch (USAFOMS/OMYO) are responsible for reviewing at a minimum 10 percent of all returned job inventories and 100 percent of all task factor booklets. This quality control procedure is one method of ensuring job incumbents are following instructions for filling out the booklets and that the data look realistic.

The occupational analysts are also responsible for analyzing and making recommendations based on occupational survey data for the enlisted career ladders. Of primary interest, The "1990 OMYO Self-Inspection Checklist" located in the *USAFOMS/OMYO Occupational Analyst Handbook* (1990) and recreated in Figure 6, instructs analysts to "Analyze task difficulty data." Section 4 of the Handbook, "Procedural guide for writing OSRs for Enlisted AFSCs," contains the guidelines for reviewing task difficulty (TD) data when writing the related portions of the OSR. These guidelines provide three questions to be asked by the analyst about tasks with high task difficulty ratings: What are the tasks?; Who performs the tasks?; and Is there a trend?.

**The Procedural Guide for Writing OSRs for Enlisted AFSCs:
TD Survey Data**

"Objectives: To be able to provide valuable information for decision-makers regarding training decision, we must first gather data that are reliable and analyzable. To this end, we collect secondary task factor data in the form of task difficulty and training emphasis. Each of these collection instruments provide very specific and definitive kinds of data. For example, task difficulty is a measure of how long it takes to learn how to do a particular task.

"Analysis Procedures: The primary purpose of this analysis is to provide information to technical training center personnel which may be used to review and update current training programs. With this in mind, the analyst should leave no stone unturned in his/her search for answers that will support a sound training policy. It is incumbent upon the analyst to use approved statistical techniques, as well as sound judgement in performing the analysis process, based on the data collected. Once the analyst has percent performing data, task factor data, and properly matched the STS and POI, he/she can then compare that information to all pertinent documents."

For the "Task Difficulty" subsection, the following questions should be addressed:

- 1) What tasks are rated highest in TD?
- 2) Are the highest rated tasks performed by high percentages of first-term airman, 7-skill level personnel, or both?
- 3) Is there a pattern found for tasks rated highest in TD?

Figure 6. Task Difficulty Analysis

The Problem

The utility of a technique to determine the difficulty level of Air Force jobs based on a time to learn definition is not in debate. The basic question is whether the task difficulty data currently collected by USAFOMS is in fact "time it takes to learn to perform a task satisfactorily" or are raters providing different interpretations of task difficulty in their ratings?

As shown in Figure 5, the current instructions provided to the NCOs by USAFOMS for rating task difficulty do not emphasize the "learning" dimensionality of task difficulty. The instructions state the definition only once without bold-face or underlining of the term "learning" to highlight or draw attention to this important distinction. Furthermore, the rating scale defined at the top of each task-rating page states only "task difficulty" (see Appendix B). When task difficulty is used without any qualifiers, individuals may think of performance difficulty or how difficult it is to perform the task. In which case, USAFOMS may be essentially providing instructions which collect task

"performance" difficulty or a "global" difficulty rather than "learning" difficulty ratings. If so, improper guidance may be given to training developers and other policy makers. One consequence might be that classification personnel establishing high aptitude requirements for specialties which have tasks that are very hard to complete or perform but whose technical learning requirements are not difficult. While interrater agreement could be high, it only means that raters agreed on difficulty, not that they rated only learning difficulty. Past research clearly indicates "time to learn" is a valid and reliable means of collecting and understanding task difficulty (Burtch et al., 1982; Christal, 1974; Lecznar, 1971; Mead 1970a, 1970b; Mead & Christal, 1970; Weeks, 1981). The concept of task learning difficulty is key to ensuring proper data are collected from the raters. The users of task difficulty information base decisions with the assumption the data measure the task learning difficulty. There is a strong need to know if task difficulty data currently being collected are actually task learning difficulty, and if not, to identify what was being collected, as well as a possible method to ensure future task learning difficulty instructions are clearly understood and accurately reported by subject matter experts.

If the ratings are not "pure" learning difficulty measures, the required corrective actions might be minimal. Re-titling the cover page as Task Learning Difficulty, highlighting and emphasizing the definition and instructions, and titling the scales located at the top of each page as "Task Learning Difficulty" might be sufficient changes. However, these changes may have a serious impact on validity and reliability of both past and future data (Demetriades, Knoll, & Boyce, 1990). Research is required to demonstrate the relationship between task difficulty data, as collected by USAFOMS, and data which emphasize the learning aspect of difficulty.

Therefore, this study investigated a new data collection procedure which modified the current techniques by using a clearer more concise cover page statement, instructions, and difficulty rating scale headings with the expectations that:

- 1) Significant differences will be identified between different rating methods for the same tasks. Specifically, there will be a significant difference between task learning difficulty and task difficulty/task performance difficulty ratings.
- 2) Task learning difficulty data will have greater rater reliability because of a more focused definition. That is, task learning difficulty will have fewer divergent raters as well as higher intra- and inter-rater reliability than the other two rating procedures.
- 3) Task learning difficulty will appear more valid as a measure of learning difficulty through the specificity of its technique and relationship to other task data, such as percent time spent, time functions (seniority and experience of high grade, longer service time, and greater time in career field), and training emphasis of tasks performed by incumbents in their first jobs.

METHOD

The method used to investigate the reliability and validity of task difficulty data was three-fold and similar to the standard procedures currently used to collect task difficulty data. Three equivalent samples of members in an Air Force Specialty (AFS) were surveyed in a single administration. Each rater received one booklet, either the current task difficulty (TD) survey booklet, a new experimental task learning difficulty (TLD) survey booklet, or an experimental task performance difficulty (TPD) booklet. Training emphasis (TE) booklets and job inventory (JI) surveys were also administered to AFS personnel according to standard USAFOMS procedures. Raters had approximately 3 months to complete their survey booklet. As necessary, follow up telephone interviews with a sample of raters were coordinated to obtain subjective estimates of rating scale differences.

Materials

The Air Traffic Control career ladder was selected as the specialty to be used for this study for three reasons. First, the job inventory was in the final stages of completion at the time of this study's initiation. Second, the population size was large enough for administration of the additional survey booklets; and thirdly, the nature of the Air Traffic Control job was considered suitable for investigating task difficulty with results being generalizable to other AF specialties. The Air Force Specialty Code (AFSC) 272X0 Air Traffic Control career ladder job inventory task list, dated June 1992, was provided to senior NCOs. The Job Inventory was prepared by an inventory developer after carefully reviewing pertinent documents, such as previous task lists and training documents. This task list was refined and validated through personal interviews with 34 subject-matter-experts representing five operational bases. This process resulted in a final job inventory containing 514 tasks organized under 10 duty headings.

Three types of difficulty inventory booklets were constructed. One booklet received a brown cover page with the standard "Task Difficulty" title and contained the current instructions and scale headings (see Appendix B). Another inventory used a pink cover page with the revised title, "Task Learning Difficulty," and revised instruction page and scale headings (see Appendix C). The third inventory used a purple cover page with the revised title, "Task Performance Difficulty," and revised instruction page and scale headings (see Appendix D). All three difficulty booklets contained a duplicate page to assess internal consistency or intra-rater reliability. Page 11 (tasks 229 through 252) was chosen for duplication and was located following page 10 and again following page 13 in each task factor booklet (see Appendix E).

Subjects

All eligible senior noncommissioned officers holding a Duty Air Force Specialty Code (DAFSC) 27270 designation were identified using the Uniform Airman Record (UAR), provided by Brooks Air Force Base (AFB) Armstrong Laboratory. The UAR is maintained by the USAF Military Personnel Center (USAFMPC) at Randolph AFB, Texas. From the 1,307 eligible personnel, three lists of 75 names randomly selected by the