AL/HR-TR-1996-0013

A R M S T R O N G   L A B O R A T O R Y

METHODS FOR CLUSTERING OCCUPATIONAL
TASKS TO SUPPORT TRAINING DECISION MAKING

Bruce M. Perrin
David S. Vaughan
Jimmy L. Mitchell

McDonnell Douglas Aerospace
P. O. Box 516
St. Louis, MO   63166

Robert M. Yadrick
Winston Bennett, Jr.

HUMAN RESOURCES DIRECTORATE
TECHNICAL TRAINING RESEARCH DIVISION
7909 Lindbergh Drive
Brooks AFB TX   78235-5352

June 1996

Interim Technical Report for Period October 1994 - December 1994

DTIC QUALITY INSPECTED 3

AIR FORCE MATERIEL COMMAND
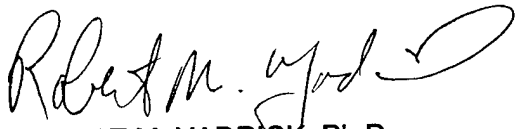## BROOKS AIR FORCE BASE, TEXAS

19960729 033

NOTICES

Publication of this report does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information exchange.
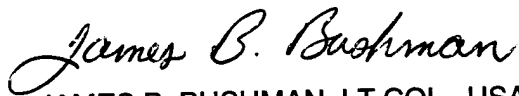
When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

ROBERT M. YADRICK, Ph.D.
Project Scientist
Instructional Systems Research Branch

R. BRUCE GOULD, Ph.D.
Technical Director
Technical Training Research Division

JAMES B. BUSHMAN, LT.COL., USAF
Chief, Technical Training Research Division
Human Resources Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this 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, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE June 1996	3. REPORT TYPE AND DATES COVERED Interim Report - period October 1994 - December 1994

4. TITLE AND SUBTITLE

Methods for Clustering Occupational Tasks to Support Training Decision Making

6. AUTHOR(S)
Perrin, B. M.
Vaughan, D. S.   Yadrick, R. M.
Mitchell, J. L.   Bennett, W. Jr.

5. FUNDING NUMBERS

C - F33615-89-C-0001
PE - 62205F
PR - 1123
TA - C1
WU - 03

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

McDonnell Douglas Aerospace

PO Box 516

St. Louis, MO 63166-0516

8. PERFORMING ORGANIZATION

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Armstrong Laboratory
Human Resources Directorate
Technical Training Research Division
7909 Lindbergh Drive
Brooks AFB, TX 78235-5352

10. SPONSORING/MONITORING

AL/HR-TR-1996-0013

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Estimates of the cost of providing training, in forms ranging from classroom instruction to on-the-job training, are needed to support decisions about who gets trained, when, where, and on what skills. To counter the myriad of uncontrollable factors that may obscure the relationship between manpower, personnel, and training policy changes and organizational outcomes, an organizational simulation of Air Force occupations called the Training Impact Decisions System (TIDES) was developed. An important first step in obtaining these occupation-level outcome estimates in TIDES is to identify groups of tasks with similar knowledge and skill requirements, because economies will be realized when these tasks are trained at the same time. This report compares results from using two different methods to identify groups of Air Force Occupational Survey tasks where these training economies would occur, including methods based on subject matter experts' judgments and statistical clustering using task coperformance. The results from two field applications indicated that the statistical methods could replicate much of the structure of the experts' clusters, and so, could be used to facilitate the process of identifying these task groups. Use of these methods to form task clusters which could be used to support a broad range of training and personnel decisions is also discussed.

DTIC QUALITY INSPECTED 3

14. SUBJECT TERMS
Decision Support
Task Clustering
Training

15. NUMBER OF PAGES
17

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89) Prescribed by ANSI Std Z-39-18
298-102 COMPUTER GENERATED

Table of Contents

List of Tables

PREFACE

METHODS FOR CLUSTERING OCCUPATIONAL TASKS
TO SUPPORT TRAINING DECISION MAKING

INTRODUCTION

Training decision making is a process of balancing, either implicitly or explicitly, a number of separate, and possibly inconsistent objectives. Among these are the goals of providing instruction using the most effective delivery methods available, insuring that adequately trained manpower is available where and when it is needed, and providing this training in the most cost-effective manner possible. Ideally, the instructional delivery method best suited for training a particular skill or knowledge is also the least expensive and is capable of providing sufficient numbers of fully trained personnel when they are required. In practice, instructional, personnel, and financial considerations must be balanced in deciding who gets trained, when, where, and on what skills.

Determining the proper balance of these demands is problematic, however, as research on the effect of training interventions on organizational outcomes such as cost and workforce productivity is rare and typically shows weak, inconsistent effects. Alliger and Janek (1989), for example, identified a sample of only 12 studies which attempted to measure these relationships, producing an average correlation of only 0.19 between behaviors that could be attributed to a training program and organizational results. The observation that the measured impact of training generally declines as evaluation criteria are selected farther from the point of intervention (Goldstein, 1993) can be interpreted as reflecting the influence of a myriad of uncontrolled and unmeasured factors, such as opportunities to perform (Ford, Quinones, Sego, & Sorra, 1992) and resource constraints (Peters & O'Connor, 1980), that occur between the point of intervention and the point at which organizational outcomes are measured.

One response to the difficulty of measuring training impacts on organizational outcomes has been the development of training utility analysis (Cascio, 1989). Under this approach, an estimate of the validity of a training program is essentially translated into dollar figures, reflecting the benefit from productivity gains minus the cost of the program. These methods permit one, for example, to compare programs of differing validities and development costs or to compare the unique impacts of multiple programs. Training utility analysis may also be used to estimate the relative cost of a formal training program compared to on-the-job training (OJT), assuming the loss of productivity of more experienced employees and the cost of resource use for OJT can be estimated.

In a similar approach, the Air Force has developed the Training Impact Decision System (TIDES), which permits tradeoffs in terms of organizational outcomes such as overall costs and resource usage. Specifically, the TIDES holds organizational productivity constant, so that alternative manpower, personnel, and training programs, including different types and levels of formal training and OJT, can be compared. Within TIDES, the demand for training under varying levels of staffing, different job structures and patterns, and different mixes of formal training is derived from an entity-based simulation model that represents the flow of personnel between jobs and training throughout their careers (Mitchell, Yadrick & Bennett, 1993). Proficiency gains from the use of formal training methods, such as classroom lecture or supervised task performance in a laboratory, are represented as a learning curve for a given set of tasks within the occupation. Recognizing that a given training method generally has a point of

1

diminishing returns, statistical models fit to Subject Matter Experts' (SMEs') estimates of proficiency gains under varying allocations of these methods consistently produce the expected, negatively accelerated, learning curves (Perrin, Knight, Mitchell, Vaughan, & Yadrick, 1988; Bennett & Perrin, 1989).

Using data from the simulation, the amount of formal training for each individual across all courses is translated into proficiency using the learning curves, and any shortfall of these courses to achieve the established level of productivity is estimated. The shortfall, in turn, is translated into OJT requirements, using a learning curve for this type of training. Finally, resource requirements for providing both formal training and OJT are compared to the capacities of the organization to provide these types of training, and overall labor and nonlabor costs are estimated (Rueter & Feldsott, 1989; Rueter, Feldsott & Vaughan, 1989). More complete overviews of the TIDES system are available to the interested reader (Vaughan, Mitchell, Yadrick, Perrin, Knight, Eschenbrenner, Rueter, & Feldsott 1989; Mitchell, Vaughan, Knight, Rueter, Fast, Haynes, & Bennett, 1992).

The TIDES simulation and learning curve models are built from AF Occupational Survey (OS) tasks, typically a set of 500 to 1200 behavioral statements that are used to describe work within an occupation. As behavioral statements, a given OS task may share skills and knowledge with other tasks in the same occupation. Consequently, an estimate of the number of the hours required to train all of the tasks in an occupation individually may overstate training requirements, if similar tasks are trained at the same time. If the similarities between tasks are represented, the TIDES model can capture the savings realized by these economies. For example, if all or a portion of a group of related tasks is trained at the same time in a formal course, the hours devoted to the group of tasks and the proficiency achieved are estimated and applied in TIDES. Similarly, economies from conducting OJT on related tasks are accounted for in TIDES when simulated job incumbents perform two or more of the related tasks as part of their job responsibilities. Thus, a starting point for the development of a TIDES model of an occupation is the identification of groups of tasks for which economies would result if they were trained at the same time.

Although we focused our efforts on methods suitable for identifying task groups for TIDES, our findings are also potentially of broad interest to personnel psychologists, job analysts, and training managers. For example, Goldstein (1993) discusses the usefulness of task clusters in training needs analysis and training planning. Similarly, Cranny and Doherty (1988) indicate that job analysts form task clusters for a variety of purposes, including personnel selection. Methods for forming these clusters, however, have remained elusive. Goldstein (1993) summarizes the status of work in the training area by noting that there are "questions on the appropriate techniques to use in developing clusters" (p. 60). Cranny and Doherty (1988) argue persuasively that a common approach to forming tasks clusters, factor analysis of task similarity ratings, is generally inappropriate for any purpose. They conclude their article by proposing a number of alternative approaches, including cluster analyzing Subject Matter Experts (SMEs) ratings of task similarity or having SMEs directly sort task statements into piles (clusters) on the basis of hypothesized similarities. Some years after Cranny and Doherty's (1988) paper, however, we know of no published reports of attempts to develop or evaluate alternative approaches to task clustering.

This report describes our evaluation of two approaches for grouping tasks that should be trained together to support TIDES analysis of training costs in an organization. The methods were applied in two, very different occupations -- Avionics Inertial and Radar Navigation Systems Maintenance (hereafter referred to as navigation systems maintenance) and Law Enforcement/ Security Police (hereafter referred to as military police). The first method involved having SMEs sort tasks into groups which they believed should be trained together. Although time-consuming and expensive, this method directly elicits the task groupings we sought, and consequently, the groups derived from this method

2

provided the standard against which the other method could be compared. The second method involved statistically clustering tasks based on task coperformance.

METHOD 1 -- SME CARD SORTING

Perhaps one of the most straightforward ways to classify information is simply to have people sort examples into categories. The groups that result constitute the categories. This approach has a long history in psychology for research in cognitive modeling and comparing expert and novice's perceptions (e.g., Schoenfeld & Herrmann, 1982). In the training area, this method is commonly used to derive clusters for describing jobs and assessing training needs, and has been termed the rational clustering exercise by Goldstein (1993).

To make the job of sorting all of the tasks in an occupation more manageable, we wanted to provide some initial structure in the form of starter groups. These piles were not to limit the sorting of tasks in any way, but merely to provide groups of manageable size based on reasonable divisions of the work. Air Force training and operational managers already organize tasks into groups in a document called Specialty Training Standards (STSs), used to promote discussions and reach agreements on the level of training to be provided for each group of tasks. We believed that STS groups might make acceptable starter piles. A second set of starter piles, based on task coperformance clustering, was provided in order to determine if initial structure unduly affected the final results. Coperformance clustering is described in the second study.

SMEs worked as teams and were asked to rearrange the cards into piles that represented groups of tasks that "should be trained together". The directions to the SMEs stressed the importance of using their expertise, and the instructions specifically noted that the resulting groups could include the same task in two or more groups and could be of any size, including single tasks. In both occupations (navigation systems maintenance and military police), enough SMEs were available to form two or more teams. These teams worked independently, proceeding at their own pace and using their own strategies to sort the tasks. When two SME teams were satisfied with their task clusters, they met to reconcile any differences between them. This reconciliation phase allowed the SMEs to compare strategies, decide upon the best criteria, and produce a final set of task clusters.

Participants

The card-sorting method in the navigation systems maintenance occupation was applied at Keesler AFB, MS. Ten SMEs, who were technical trainers at the base, served as subjects. The average level of the SMEs' experience in the career field was almost 11 years and varied from about 7 to over 15 years. Initial card sorts required about one and one-half days for one team and approximately two days for the other team. Reconciliation of the results from the two teams was completed on the third day.

The card-sorting method for the military police career field was applied at Lackland AFB, TX. Fourteen SMEs participated in the exercise. Eight were currently trainers at the base, while the remaining six were assigned to operational units. Experience in the career field varied from just under 5 years to over 24 years, an average of slightly over 12 years. Because both training and operational units were represented, individuals were assigned to teams differing in both background (field or training) and starter pile (STS or coperformance). The four resulting teams will be identified by the type of starter pile and background as follows: STS/school; STS/field; coperformance/school; and coperformance/field.

The two teams with coperformance starter piles finished sorting midway through the second day. They started reconciling their results at that time, and completed their work by the end of the third day. The two teams with STS starter piles did not finish their initial sorts until the morning of the third day, but also completed their reconciliation by the end of the day.

Results

Of the six task sorts in the military police occupation (four team results and two reconciliation sorts), descriptive statistics on five of the sorts were very similar. Across these five results, the number of task groups was approximately the same, ranging from 65 to 75 groups, and the average group size ranged from about 9 to 11 tasks. Similarly, few task groups were formed with only a single task in these five sorts, and the largest task group in these sorts varied from 30 to 49 tasks. The same task was seldom placed in two or more groups by these SMEs (a maximum of 46 tasks were duplicated) and most of the tasks were classified (a maximum of 12 tasks were unclassified across the five sets of results). The sixth sort, the one from the STS/school team, however, showed fewer (only 33) and larger groups (averaging over 20 tasks per group). Their results involved more duplicate (129 duplicate tasks) and unclassified tasks (117 tasks not classified). Additionally, their largest group contained 109 tasks.

The results for the card sorts from the navigation systems maintenance teams were, on the surface, also quite dissimilar. Groups formed from the STS starter piles averaged only about one-fourth as many tasks per group as the groups formed from the coperformance starter piles (5.6 tasks per group compared to 23.6 tasks), although the largest task group produced by both SME teams was not substantially different (78 tasks compared to 89 tasks). Additionally, neither team used many duplicate tasks and most tasks were classified. The results of the reconciliation sort by these SMEs reflected compromise on all measures. The number (75) and average size of the groups (10.1), the number of single-task groups (4), and the number of duplicate (7) and unclassified tasks (26) following the reconciliation sort were all greater than the same measure for one team's results and less than it for the other.

For our purposes, however, the consistency with which the tasks are grouped is more germane than characteristics such as number or size of the groups, although the two are related. A number of statistics are available to compare groupings, based on pairwise classification of cases for the two solutions. One such statistic is the Fowlkes and Mallows (1983). If A indicates the number of pairs of tasks grouped by both solutions, and B and C indicate frequencies of disagreement in which pairs are grouped by one method, but not the other, the Fowlkes and Mallows (F&M) is computed as follows:

$$F\&M = \frac{A}{\sqrt{(A+B)*(A+C)}}$$

The F&M has an upper bound of 1.00 when the two solutions agree perfectly, and a lower bound of 0.0 when A is zero. Additionally, it is undefined when cells A, B, and C are all equal to zero, but this would occur only when the number of groups equals the number of cases for both solutions (i.e., no clustering had occurred).

A sampling distribution for the F&M was generated by randomly assigning tasks to groups, then computing the agreement statistic, to determine the level of agreement that would occur by chance for solutions representative of the number and size of our task groups. Comparisons of two solutions with relatively large task groups showed higher levels of chance agreement than comparisons involving solutions with smaller groups. For comparisons of solutions with large groups (averaging over 30 tasks per group), the F&M statistic exceeded 0.094 in only 1 case out of 100. We adopted this number as the critical value of the F&M statistic at the .01 significance level. Selection of this value provides a very

conservative test, as the F&Ms for comparisons of solutions more typical of our results (about 9 tasks per group, on average) exceeded 0.027 in only 1 of 100 cases.

Guidance from the statistical literature was equivocal as to how to treat tasks not classified and how to treat duplicate task statements. In the first instance, unclassified cases are generally omitted from statistical comparisons, when the available information does not support clustering them in any group. Tasks were left unclassified in the card sorts primarily because the SMEs felt these tasks were no longer part of the responsibilities of the occupation, not because of an inability to place them with related tasks. Nonetheless, we decided to omit these tasks from the analysis. In the case of duplicate tasks, each instance of the task was tabulated individually. That is, one grouping of the duplicate may have agreed with the second solution (and be counted in A), while a second sorting would disagree with the other solution (and be counted in B or C).

The results of the SME card sorts are reported in Table 1. All comparisons are F&M agreement statistics, and all are statistically significant at the .01 level. The statistics indicated by an asterisk (*) are not independent comparisons; that is, they are comparisons involving a reconciliation sort and the solution from one of the two teams reconciling their results.

The F&Ms ranged from 0.199 to 0.304 for the 11 independent comparisons involving groupings by the military police. The two task groups formed by the field SMEs agreed more closely than any other set of comparisons, even though these SMEs worked from different types of starter piles. Additionally, the reconciliation sorts were more consistent with the field SMEs' groupings than the groupings formed by the SMEs involved in training.

Only one comparison is independent for the navigation systems maintenance task sorts. The F&M

Table 1
Comparisons of SME card sorts

			GROUP		
	STS reconcil.	Coperf. school	STS school	Coperf. field	STS field
Military Police					
Coperf. reconcil.	0.293	0.293*	0.199	0.716*	0.300
STS reconcil.		0.290	0.227*	0.294	0.757*
Coperf. school			0.233	0.270	0.298
STS school				0.199	0.219
Coperf. field					0.304
Navigation Systems Maintenance					
Reconciliation		0.520*	0.787*		
Coperformance school			0.476		

statistic for this comparison was quite high (0.476), greater than for any comparison between independent military police SME card sorts.

Discussion

Although the F&Ms are all statistically significant, one would like to develop an impression of the level of agreement expressed in these F&Ms, so that their practical importance could be evaluated. Obtaining this insight is difficult, given the number of cases (tasks) and groups produced and the complexity of the relationships between solutions, although some inferences can be drawn. For example, if all of the solutions involved task groups of 9 (near the average for most of our results), an F&M of 0.426 would result if 6 of 9 tasks in one solution were also grouped in the second solution. This result is quite comparable to that obtained by the navigation systems maintenance SMEs, with an F&M of 0.476 for the comparison of their independent sorts. The card sorting results are not this simple, of course, as group sizes varied widely within a given sort.

The F&M, like other agreement statistics of this type, is substantially affected by differences in the size of groups being compared. Assume, for example, that four of the task groups from the navigation systems maintenance SMEs who started with STS based piles (which averaged only about 6 tasks) could be combined to form one task group from the SMEs who started with coperformance piles (which averaged nearly 24 tasks in each). These differences in group size alone would reduce the F&M to about 0.43, less than the observed F&M for this comparison. Even for solutions in which the average group sizes were the same, observation of the SMEs during card sorting suggested that differences in task group size in different areas of the occupation were common, due to differences in expertise and emphasis. Where one team of SMEs might cluster 12 tasks, another might break these tasks into two groups of six. With variation in specificity from area to area, solutions could average nine tasks per group, yet show substantial differences in specificity across areas of specialization.

Although no systematic analysis of differences in specificity was performed (due to the overall size of the task lists and complexity of the group structures), we believe a substantial part of the observed shrinkage in the F&M statistics can be attributed directly to this cause. If task groups could be equated for specificity, we would expect a level of overlap of 70 percent or more between individual task groups. Across the occupations, we believe SMEs were able to achieve considerable agreement on the tasks that should be trained at the same time.

METHOD 2 -- STATISTICAL CLUSTERING

Statistical clustering has a long history in military occupational analysis. For more than 30 years, job analysts in the Armed Services have used case cluster diagrams to assist in identifying jobs -- groups of people performing similar sets of tasks. Over that period of time, standardized data collection instruments (Occupational Survey task inventories) have been developed and refined; computing algorithms and diagnostic statistics have been devised, tested, and implemented in the Comprehensive Occupational Data Analysis Programs (CODAP); important background characteristics have been identified and incorporated to aid in the analysis of the structure of work; and an extensive body of research and application has accrued (Christal, 1974; Christal & Weissmuller, 1988).

By comparison, the notion of using cluster analysis to derive task groups to support training decision making is relatively new, while the research on the use of other statistical techniques, notably factor analysis, has not been encouraging. Factor analysis has not been found to replicate critical job dimensions identified by SMEs, nor have the factors derived from this analysis always been interpretable

as job dimensions (Cranny & Doherty, 1988). As a result, Goldstein's rationale clustering exercise, which is equivalent to card sorting, is cited as the most tenable method (Goldstein, 1993), while other researchers have cited the need for additional research on statistical methods for grouping tasks to support training decision making (e.g., Schmitt, 1987). Research on the use of cluster analysis for these purposes was begun in only the mid-1980s. Nonetheless, a variety of algorithms has been added to CODAP for the analyst interested in statistically clustering tasks (Phalen, Mitchell & Hand, 1990).

CODAP uses the average linkage clustering procedure (Ward, 1963). This procedure has performed well in empirical studies that have compared various clustering algorithms (e.g., Milligan, 1981), and consequently, was selected for use in task clustering. We used task coperformance as the similarity index in this analysis. While job typing involves grouping persons who perform the same (or similar) sets of tasks, task clustering using coperformance similarity produces sets of tasks which tend to be performed as a group; that is, if individuals perform one task in the group, it is likely that they perform many of the others. Tasks that are coperformed may be similar with respect to requisite skills and knowledge, and, more pertinently, will result in economies in training cost and from the sharing of resources during OJT.

To identify task groups from a coperformance cluster diagram, we elected to use the same methods that a job analyst would typically use to interpret a case cluster diagram. The heart of this approach is to identify clusters that maximize the variability between groups, while minimizing the variability within groups (Archer, 1966). We tried some of the informal guidelines that job analysts have developed over the years. For example, the 35-50 rule is a general rule-of-thumb that occupational analysts use to identify initial jobs, and which indicates that a good starting point for identifying jobs is to select case clusters with a homogeneity index within the group of about 35.0 or less and a between-group homogeneity index of 50.0 or more. Final jobs often deviate substantially from this guideline, but the value of the rule as an heuristic is unquestioned. It was apparent from the outset, however, that this heuristic was not applicable to task cluster analysis (a similar but more stringent rule could perhaps be adopted for task clustering purposes.) Additionally, many software products that an analyst may employ to identify jobs are not available to interpret a task cluster diagram.

Results

The analysts who interpreted each of the task coperformance cluster diagrams were experienced in occupational analysis and were also somewhat familiar with the occupation they were evaluating. They found that the most useful information for identifying task groups was the pattern of changes in homogeneity, rather that the actual levels. Specifically, the points where homogeneity dropped significantly as tasks were combined tended to mark distinct task groups, a heuristic also used to identify jobs. The analyst who interpreted the military police cluster diagram identified 67 task groups, resulting in an average group size of about eight tasks. The largest task cluster contained 34 tasks and no groups were composed of a single task (although several involved a pair of tasks). Additionally, 115 tasks could not be classified solely on the basis of the cluster diagram.

Working with the navigation systems maintenance task cluster diagram, the analyst identified 95 task groups, for an average of just under eight tasks per group. Only 35 tasks were not classified. The largest task cluster contained 47 tasks and the smallest again consisted of a pair of tasks.

Table 2 reports the F&M agreement statistics between the task groups formed by SME card sorting and the task clusters identified from the coperformance cluster diagrams in each occupation. All comparisons are statistically significant at the .01 level, indicating that the agreement between the

statistically defined and the SME generated clusters was statistically greater than expected by chance. Additionally, the magnitude of the F&M statistics for these comparisons is roughly equivalent to that found for comparisons between different SME card sorts, although the statistics for the military police occupation are consistently greater than those for the independent card sorts in the same occupation.

Table 2
Comparisons of SME card sorts to the task clusters identified
from the task coperformance cluster diagrams

SME Card Sorts	Analyst's Interpretation
Military Police	
Coperf. reconcil.	0.445
STS reconcil.	0.463
Coperf. school	0.441
STS school	0.442
Coperf. field	0.305
STS field	0.472
Navigation Systems Maintenance	
Reconciliation	0.315
STS school	0.273
Coperformance school	0.153

Discussion

The descriptive statistics reported in this and the previous section suggest that there is considerable agreement as to the appropriate level of specificity for task groups. Across both occupations, the reconciliation card sorts and analysts' interpretations of the coperformance cluster diagrams yielded groups with about eight to 11 tasks. While some of the independent card sorts resulted in both broader and more narrowly defined groups, these differences tended to be minimal when final groups were formed.

The type of starter pile used in the card sorting exercise, appears to have had no appreciable effect on the results. Card sorts which began with coperformance starter piles tended to be no more similar to the analyst-identified statistical clusters, which was based on coperformance, than the card sorts produced by SMEs using STS starter piles. This result is perhaps to be expected, as the card sorting directions encouraged SMEs to use their expertise to restructure the initial piles as necessary.

In the military police occupation, where both training and field personnel participated in the card sorts, the background of the SMEs may have had minor effects on the results. Field SMEs' card sorts tended to more closely match the coperformance clusters than did the school SMEs' sorts, although the difference is small. Presumably, field SMEs are more sensitive to performance criteria, and so, produced groupings similar to task coperformance. Additionally, it should also be noted that the results from the two field teams showed the highest agreement and that the reconciliation sorts were more similar to the field SMEs' results than to the school SMEs' task groups (as reported in Table 1).

An additional interesting finding in the comparisons between the card sorts and the statistical clusters for the military police involves the pattern of results. The F&Ms between the statistical clusters and each of the card sorts were higher than those between any independent card sort in the occupation. The closest agreement between two independent SME sorts was that for the two field teams, which yielded an F&M statistic of 0.304 (Table 1). All comparisons between the card sorts and the coperformance clusters yielded values of the F&M statistic greater than this level. While this result suggests substantial common agreement between these techniques, the data do not necessarily support this interpretation. Agreement statistics such as the F&M have been shown to be substantially affected by decisions about the exclusion of cases from the analysis (Edelbrock, 1979). With over 17 percent of the tasks omitted from the statistical clusters (115 of 666 tasks omitted), the relatively higher level of the F&Ms found may be an artifact of the exclusion criteria used by the analyst.

CONCLUSIONS

In a number of ways, card sorting by SMEs and task coperformance clustering represent complementary ways of forming task groups. On the one hand, statistical clustering can replicate, to a great extent, the clusters SMEs produce when asked to sort tasks into groups that should be trained together, doing so efficiently and using existing information. The resulting statistical clusters can be expected generally to agree with SME-produced groups as well as solutions from independent teams of SMEs agree with each other. This result is in contrast to previous research on factor analysis, which found that statistically derived factors could not faithfully reproduce SME judgments of job dimensions (Cranny & Doherty, 1988).

On the other hand, statistical clustering has certain limitations in its flexibility, regardless of it ability to replicate SME-produced task groups. It is not always possible to assign a given task to a cluster solely on the basis of a cluster diagram. In identifying jobs from a case cluster diagram, for example, it is common to have from 5 to 10 percent of the cases unclassified. This problem was even more pronounced for task clustering in the military police occupation, where 115 of 666 tasks were omitted from any group. One possible solution would be to leave these tasks as single task groups; however, this solution does not compare favorably with the card sorting results. Another solution would be to relax the restrictions on the coperformance clusters, so that more of the tasks would be grouped. It is not clear, however, that the hierarchical procedure would produce acceptable groupings using relaxed inclusion criteria. It has been suggested that non-hierarchical methods for refining task and job clusters around analyst-identified "seed" groups might be a more realistic approach (Phalen, Staley, & Mitchell, 1987), but the work needed to demonstrate such an approach has not yet been reported.

SME card sorting represents a very flexible method for grouping tasks to portray training requirements, using established methods and providing descriptive frameworks for the resulting groupings. The method has some limitations as well. Our results lead us to two conclusions. The first is that multiple teams of SMEs should be involved. This conclusion follows from the somewhat

9

inconsistent results produced by one of the military police SME teams. Given the current state of the requirements for card sorting, a single SME team may produce task groups which cannot be replicated and probably would not adequately fulfill the requirements for training outcome estimation. At a minimum, a second team of SMEs should independently sort the tasks, so that the appropriate comparisons can be made. If the results are not consistent, further SME input would be warranted. Second, our results suggest that having an SME team that includes operational personnel is also highly desirable. Where data were available, the field SME teams were more consistent with each other, were more consistent with the reconciliation card sorts, and were more consistent with the coperformance clusters than were the SME teams composed solely of training personnel.

Given the requirement for multiple SME teams composed of both training and field personnel and the fact that card sorting required a minimum of 2 days (3 if a reconciliation sort was produced), arranging for card sorting in any particular occupation can be difficult. These difficulties will likely be compounded for small, highly specialized career fields, when personnel are dispersed geographically, or when workload is high.

Because the strengths of one method tend to compensate for the weaknesses of the other, we formed a composite procedure for grouping tasks -- coperformance clustering interpreted by an analyst, followed by SME refinement. This procedure capitalizes on the strengths of both clustering and card sorting, while avoiding many of the pitfalls. Coperformance clusters are identified to capture the core of agreement in task groupings with moderate costs. SMEs then refine these groups, classify any tasks that the analyst could not place, and supply descriptive labels, giving the procedure additional flexibility. This composite method was applied to both the navigation systems maintenance and military police occupations with the following results:

1) All refinement of task groups, including placement of all tasks not clustered, was completed in one day. Additionally, the number and magnitude of changes the SMEs made to refine the analyst-identified task groups were rather small. The comparison between the task groups initially identified by the analysts and the SME-refined groups yielded an F&M agreement statistic of 0.919 for the military police and 0.877 for navigation systems maintenance.

2) The descriptive statistics on the resulting task groupings were comparable to previous efforts in the size of the task groups (between 9 & 11 tasks); use of duplicate tasks (0 and 6 duplicate tasks); and number of single task groups (4 and 5 single task groups).

3) The F&M agreement statistics for the comparisons between the card sorts and the task groups formed by this composite method were all statistically significant, tended to exceed the statistics for comparisons between independent sorts, and tended to be greatest for the reconciliation card sorts. Overall, this method appears to have tapped a common core of agreement among SMEs as to which tasks should be trained at the same time.

Thus, the combined procedure has the strengths of both computerized statistical clustering and expert human judgment. It uses statistical clustering to quickly and economically capture general patterns in task groupings about which experts can agree. It then uses expert judgment to refine and succinctly characterize these task groups.

As noted previously, identifying similarities in tasks, which may result in economies in training materials, content, equipment, and the like, is obligatory in the Training Impact Decision System (TIDES), which seeks to estimate training costs. Quite apart from this use, however, these groups of

tasks provide part of a common perspective for viewing the instructional, personnel use, and financial tradeoffs in training planning in TIDES. More generally, task clusters have been widely recognized as an important source of information for describing jobs and for assessing training needs (Goldstein, 1993). The training implications of personnel policies may also be more apparent when common task groupings are used, rather than task-based descriptions.

As a final point, some of these additional benefits of using task groups derived from the method described in this report are beginning to be realized. For example, research performed for the U.S. Coast Guard found empirically defined task coperformance clusters to be preferable to SME defined "duty areas" for obtaining a broad perspective on a career field appropriate to basic job skill training or selection issues (Weissmuller & Driskill, 1991). And, in work with the Internal Revenue Service, task coperformance clusters were found to provide linkage between detailed, behavioral job descriptions and broad knowledge, skill, and ability categories that would provide the basis for personnel actions (Weissmuller, Driskill & Moon, 1991).

REFERENCES

Alliger, G.M., & Janek, E.A. (1989). Kirkpatrick's levels of training criteria: Thirty years later.Personnel Psychology, 42, 331 - 342.

Archer, W.B. (1966). Computation of group job descriptions from occupational survey data (PRL-TR-66-12, AD-653543). Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division.

Bennett, W.R. & Perrin, B.M. (August, 1989). The task characteristics subsystem: Allocating task modules to training settings. Paper presented at the 97th Annual American Psychological Association Convention, New Orleans, LA.

Cascio, W.F. (1989). Using utility analysis to assess training outcomes. In I.L. Goldstein, (ed.), Training and Development in Organizations. San Francisco: Jossey-Bass.

Christal, R.E. (1974). The United States Air Force occupational research project (AFHRL-TR-73-75; AD-774 574). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.

Christal, R.E., & Weissmuller, J.J. (1988). Job-task inventory analysis. In S. Gael (Ed), Job analysis handbook for business, industry, and government. New York: John Wiley and Sons, (Chapter 9.3)

Cranny, C.J., & Doherty, M.E. (1988). Importance ratings in job analysis: Notes on the misinterpretation of factor analysis. Journal of Applied Psychology, 73, 320 - 322.

Edelbrock, C. (1979). Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. Multivariate Behavioral Research, 14, 367 - 384.

Ford, J.K., Quinones, M.A., Sego, D.J. & Sorra, J.S. (1992). Factors affecting the opportunity to perform trained tasks on the job. Personnel Psychology, 45, 511 - 527.

Fowlkes, E. B. & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings, (with comments and rejoinder). Journal of the American Statistical Association, 78, 553-584.

Goldstein, I.L. (1993). Training in organizations (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing.

Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 16, 379-407.

Mitchell, J.L., Vaughan, D.S., Knight, J.R., Rueter, F.H., Fast, J., Haynes, W.R., & Bennett, W.R. (1992). Training decisions technology analysis. (AL-TP-1992-0026). Brooks AFB, TX: Technical Training Research Division, Armstrong Laboratory, Human Resources Directorate.

Mitchell, J.L., Yadrick, R.M., & Bennett, W.R. (1993). Estimating training requirements from job and training models. Military Psychology, 5(1), 1 - 20.

Perrin, B.M., Knight, J.R., Mitchell, J.L., Vaughan, D.S., & Yadrick, R.M. (1988, September). Training decisions system: Development of he task characteristics subsystem (AFHRL-TR-88-15, AD-A199 094). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Peters, L.H. & O'Connor, E.J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. Academy of Management Review, 5, 391 - 397.

Phalen, W.J., Mitchell, J.L., & Hand, D.L. (1990, November). ASCII CODAP: Progress report on applications of advanced occupational analysis software. Proceedings of the 32nd Annual Conference of the Military Testing Association. Orange Beach, AL: Naval Education & Training Program Management Support Activity.

Phalen, W.J., Staley, M.R., & Mitchell, J.L. (1988). ASCII CODAP programs for developing job and task clusters. In the symposium, New ASCII CODAP technology: Manpower, Personnel, & Training Applications. Proceedings of the 30th Annual Conference of the Military Testing Association. Arlington, VA: Army Research Institute.

Rueter, F.H. & Feldsott, S. I. (August, 1989). The resource/cost subsystem: Estimating training capacities and costs. Paper presented at the 97th Annual American Psychological Association Convention, New Orleans, LA.

Rueter, F.H., Feldsott, S. I., & Vaughan, D.S. (1989, December). Training decisions system: Development of the resource cost subsystem (AFHRL-TR-88-52). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Schmitt, N. (1987). Principles III: Research issues. Symposium conducted at the meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Schoenfeld, A.H. & Herrmann, D.J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 484 - 494.

Vaughan, D.S., Mitchell, J.L., Yadrick, R.M., Perrin, B.M., Knight, J.R., Eschenbrenner, A.J., Rueter, F.H., & Feldsott, S. I. (1989, June). Research and development of the training decisions system: Final report (AFHRL-TR-88-50). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244.

Weissmuller, J.J., & Driskill, W.E. (1991, May). Improving job type definition for skill training by clustering on task module values. Proceedings of the Seventh International Occupational Analysts' Workshop. San Antonio, TX: USAF Occupational Measurement Center.

Weissmuller, J.J., Driskill, W.E., & Moon, R.A. (1991, May). Economical method for establishing knowledges, skills, and abilities for job families. Proceedings of the Seventh International Occupational Analysts' Workshop. San Antonio, TX: USAF Occupational Measurement Center.

Yadrick, R.M., Knight, J.R., Mitchell, J.L., Vaughan, D.S., & Perrin, B.M. (1988, July). Training decisions system: Development of the field utilization subsystem (AFHRL-TR-88-7). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.