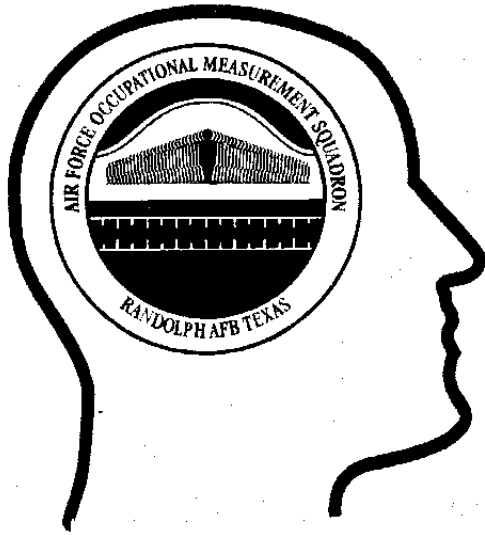


AGENDA



10th International Occupational Analyst Workshop

10-12 June 1997
San Antonio, TX

Hosted by: THE OCCUPATIONAL ANALYSIS PROGRAM
AIR FORCE OCCUPATIONAL MEASUREMENT SQUADRON
RANDOLPH AFB TX 78150-4449

1345 - 1430 LECTURE - Defining and Characterizing Subgroups Using Interrater Reliability and Cluster Analysis Techniques
Mr. Johnny J. Weissmuller (IJOA)
Mr. William J. Phalen (IJOA)
Mr. Joseph S. Tartell (AFOMS)

1430 - 1500 BREAK

1500 - 1530 LECTURE - Navy Training Continuum
CDR Robin Gandolfo (NETPDTC)
Ms. Diane Kalivoda (NETPDTC)
Ms. Carol Huggins (NETPDTC)
Mr. Al Reynolds (NETPDTC)

1530 - 1600 LECTURE - The Automated Test Outline (ATO): An Occupational Analysis Success Story
Mr. Paul Stanley (AFOMS/OMD)

1600 ADJOURNMENT

1800 SOCIAL HOUR - SALON I

1830 DINNER - SALON 1

THURSDAY, 12 JUNE 1997

MODERATOR:??

0800 - 1000 SYMPOSIUM - Innovation and Advanced Technology Research and Applications in Military Occupational Analysis

Co-Chairs:

**Maj Archie M. Smith, II (HQ Air Force
Director of Manpower and Personnel)**

**Lt Col Martin L. Fracker (Armstrong
Laboratory Human Resources Directorate)**

Presenters:

**Dr. Walter G. Albert (Armstrong Lab Human
Resources Directorate)**

**Dr. Winston Bennet, Jr. (Armstrong Lab Human
Resources Directorate)**

Mr. Johnny Weissmuller (Metrica, Inc.)

Mr. Gary Grimes (Metrica, Inc.)

Defining and Characterizing Subgroups
using
Interrater Reliability and Cluster Analysis Techniques

by

Johnny J. Weissmuller
William J. Phalen
Institute for Job and Occupational Analysis (IJOA)
and
Mr. J. Tartell
Air Force Occupational Measurement Squadron (AFOMS)

Abstract

Within the past forty years in the task-anchored occupational analysis community, standard procedures have evolved for conducting routine studies within job families/career fields. These procedures employ a two-pronged approach. The first approach is to discover "what is" from self-reports provided by actual job incumbents. The second approach is to discover "how important" from subject-matter-experts (SMEs). SMEs are selected for their broad knowledge of the target career field and their ability to rate "importance" from the specified perspective. "Importance" is a multi-dimensional notion which means *significant* along any one or more of a number of operationally viable factors such as *Task Learning Difficulty, Training Emphasis, Testing Importance* or *Consequences of Inadequate Performance*.

This paper uses international civilian and military data sets (Occupational Analysts and Mental Health Technicians) to accentuate the differential use of interrater reliability techniques and cluster analysis techniques as available in the Comprehensive Occupational Data Analysis Programs (CODAP) system. The goal is to explore traditional and innovative applications of these tools to identify and characterize subgroups within a larger population. Topics covered include an overview of when it is appropriate to apply interrater reliability techniques to job incumbent data and several approaches to handle SME factor data when more than one mainstream policy is at work.

Defining and Characterizing Sub-Groups
using
Interrater Reliability
and
Cluster Analysis
Techniques

by

William J. Phalen,
Johnny J. Weissmuller
Institute for Job and Occupational Analysis (IJOA)

Mr. J. Tartell
Air Force Occupational Measurement Squadron (AFOMS)

BACKGROUND

Within the past forty years in the task-anchored occupational analysis community, standard procedures have evolved for conducting routine studies within job families/career fields. These standards employ a two-prong approach. The first approach is to discover "what is" from self-reports provided by actual job incumbents. The second approach is to discover "how important" from subject-matter-experts (SMEs). SMEs are selected for their broad knowledge of the target career field and the ability to rate "importance" from the specified perspective. "Importance" is a multi-dimensional notion which means "significant" along any one or more of a number of more concrete factors such as "Task Learning Difficulty", "Training Emphasis", or "Consequences of Inadequate Performance". Once collected, these occupational data are processed by the Comprehensive Occupational Data Analysis Programs (CODAP) system (Christal, 1974; Christal & Weissmuller, 1988).

INTRODUCTION

This paper highlights this traditional use of the hierarchical clustering and interrater agreement tools in the CODAP system and suggests other additional uses which may be helpful in given situations. The interplay between hierarchical clustering (job typing) and interrater reliability (i.e., agreement) can provide a fresh perspective in characterizing the stability of targeted subgroups within the survey sample.

HIERARCHICAL CLUSTERING

To identify occupational subtypes, each job incumbent's data record is reduced to only the time ratings on task statements and processed through the cluster analysis subsystem (Ward, 1961; Bottenberg & Christal, 1968). The cluster analysis subsystem expects data in which each person may, legitimately, have a unique position and perspective. In practice, this tool typically yields several main job clusters with perhaps 20 or more job types per job family.

INTERRATER AGREEMENT

On the other hand, each factor of the SME data is processed through the interrater reliability subsystem to produce a single data vector representing the mainstream policy of that SME rater pool. (Kuder & Richardson, 1937; Goody, 1976). The interrater reliability analysis tool assumes that all judges are rating a common factor against a common external context. In this model, there is assumed to be a real "true score" for each item and all rater deviation is typically assumed to be the result of lack of attention to detail, insufficient topic knowledge, misunderstanding of the target factor, misuse (inversion) of the rating scale, or non-cooperative behavior such as capricious marking or non-response.

INTERPLAY: HIERARCHICAL AGREEMENT

The Agreement within Cluster-defined Job Types

Job incumbent time spent data are collected on an ipsitive scale, i.e., each relative time spent on a task rating is anchored to his or her own "time on an average task". In addition, time ratings are for their own unique job situation, not for any external or standard job classification. For these reasons, the incumbent jobs in the total sample don't fit the model for interrater agreement. Assumptions can be made which bring these data in line with the assumptions of an interrater agreement model.

By limiting the focus of analysis to incumbents within a specific job type from the hierarchical clustering, one might declare the job type to be common external entity which all members of that job type were rating as experts. The fact that all ratings were anchored to a personally chosen, typical task, may also be addressed. By assuming the average task selected from within the job type task set may vary a little in level, but that all other rating characteristics would be unaffected, one can interject corrective calculation. Adjusting relative time ratings for all raters to a common mean (usually 5.0) would remove the unwanted level effects. Using the new "absolute time" or "actual time" factor circumvents this level adjustment problem at the outset. Computing interrater agreement on the targeted job type members with the adjusted relative or the actual-time ratings will yield a valid, useful statistic.

In addition to the BETWEEN and WITHIN homogeneity values from the clustering, this second perspective describes the stability of job descriptions in a manner consistent with the treatment of the SME ratings. This standard treatment approach might simplify reporting these stability statistics in the Occupational Survey Report (OSR).

The Agreement within Cluster-Defined Policy Groups

During the validation of the Training Emphasis (TE) factor, much effort was directed at explaining "complex specialties" in which low interrater rater agreement was found (Jansen, 198?). Various methods were tried and rejected for handling this disconcerting problem. One conclusion specifically stated that hierarchical clustering in combination with interrater agreement did not resolve the problem. "The Problem" was defined as identifying which raters deviated from the main policy and explain (using demographic data) why their ratings did not agree with the global policy for the career field. Today, the "complex specialty" problem is recognized as a multiple policy situation in which there is more than one "true score" at work (Weissmuller, 1990). One cannot eliminate or explain-away the multiple policies, one can only distill them and report the various viewpoints that are uncovered.

The Agreement within Residual-Defined Policy Groups

Over the years, eager young analysts pop-up after a conference and suggest their new an innovative approach: Don't use hierarchical clustering (because it is so machine-intensive), just use interrater agreement to define job types by declaring the main rater set to be a job type and iteratively analyze the residual pool of raters declared to be divergent. As noted above, interrater agreement assumes that there is ONLY ONE policy being targeted and is very liberal about including cases submitted for processing. Since all raters and ratings are considered simultaneously (whereas clustering aggregates cases in a stepwise fashion), the residual-defined job type method tends to find very few but very large main groups

While the interrater agreement may not be appropriate to replace clustering for analysis of time spent in job types, the residual-defined policy group does make sense. In the Automated

Test Outline (ATO) project, weighted outlines for structuring promotion tests were generated using on-the-shelf survey data augmented by freshly collected testing importance ratings from the field. A key to the rapid deployment and data collection of testing importance ratings was the ability to select a subtest of around 200 tasks from the target job inventory. These tasks were selected using a Predicted Testing Importance (PTI) algorithm (Albert, et al).

Once the mini-surveys were produced, distributed and returned, these field ratings of testing importance were subjected to an interrater agreement program (RATINGS) developed for this project. Because of the sensitive nature of testing importance ratings (i.e., impacting on promotions), the RATINGS program went to great lengths to detect and characterize subpolicies within the rating pool.

The logic of the residual-defined policy group in the RATINGS program was to start with a typical interrater agreement analysis. If divergent raters were found, subsequent iterations were performed automatically with the divergent raters removed. Because of the lessons learned in the complex specialty study, the standard interrater agreement process was extended to note the impact of removing raters. One might assume that if divergent raters were removed (i.e., eliminating a source of noise), that the interrater agreement among the remaining rating pool would go up. During development, it was noted that there were raters in the main pool whose correlation with the grand mean vector went DOWN after the removal of supposedly divergent raters. The RATINGS program provided a "SUB-POLICY" warning if the number of dropped correlations exceeded 20% of the original rating pool. Across approximately 40 career fields, 6 career fields demonstrated a multiple-policy behavior in rating testing importance. As these were promotion tests for advancement into middle enlisted ranks (E-5 and E-6/7), it was not surprising that the most common policy issue was whether or not the target grade level should

heavily emphasize their **TECHNICAL** job or their **SUPERVISORY** responsibility. For example case, in the Mental Health Clinic specialists, one policy rater group rated administrative (forms, scheduling, records management) as the most important area while other raters identified patient contact and doctor-support issues as a primary concern.

CONCLUSIONS

Within the **CODAP** system, the interplay between hierarchical clustering and interrater agreement can provide a fresh perspective in characterizing the stability of targeted subgroups within the survey sample. With emerging new types of data such as actual time spent, one needs to be mindful of the appropriate and inappropriate of analyzing data.