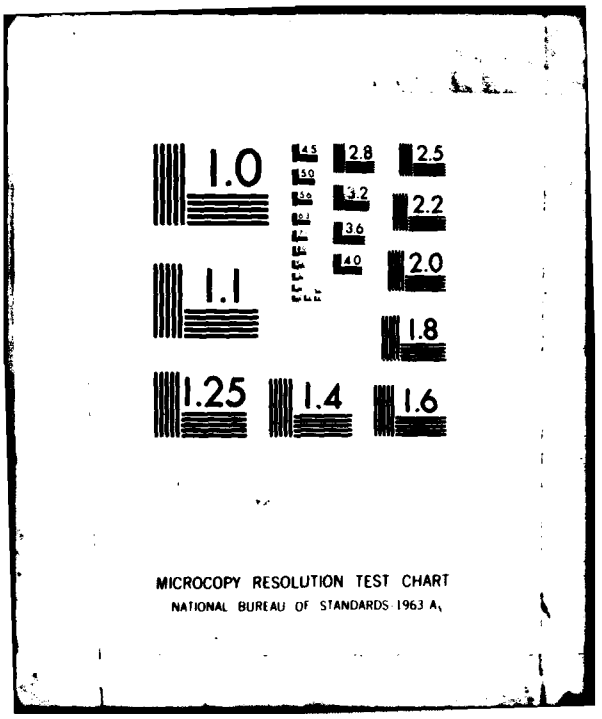
AD-A108 400 AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX F/G 5/9
INTERRATER RELIABILITY: THE DEVELOPMENT OF AN AUTOMATED ANALYSI--ETC(U)
NOV 81 M R STALEY, J J WEISSMULLER
UNCLASSIFIED AFHRL-TP-81-42 NL

END
DATE FILMED
1-82
DTIC


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

AFHRL-TP-81-12

AIR FORCE HUMAN RESOURCES LABORATORY

LEVEL II

INTERRATER RELIABILITY: THE DEVELOPMENT OF AN AUTOMATED ANALYSIS TOOL

By

Michael R. Staley
Johnny J. Weissmuller

TECHNICAL SERVICES DIVISION
Brooks Air Force Base, Texas 78235

November 1981

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

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

This paper has been reviewed and is approved for publication.

NANCY A. PERRIGO
Chief, STINFO Office

WENDELL L. ANDERSON, Lt Col, USAF
Chief, Technical Services Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TP-81-42	2. GOVT ACCESSION NO. AD-A108 400
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): INTERRATER RELIABILITY: THE DEVELOPMENT OF AN AUTOMATED ANALYSIS TOOL	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Michael R. Staley, Johnny J. Weissmuller	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Technical Services Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62703F, 63230423
11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235	12. REPORT DATE: November 1981
	13. NUMBER OF PAGES: 12
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Paper presented at the 23rd Annual Conference of the Military Testing Association, 26—30 October 1981.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

CODAP
complex emphasis
Comprehensive Occupational Data Analysis Programs
deviant raters
interrater
rater
reliability
REXALL
REXSPC
RXXNDX
training emphasis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Comprehensive Occupational Data Analysis Programs (CODAP) system is a basic tool in both operational job analysis and in occupational research. This system of programs is augmented and enhanced to meet changing requirements from the research community. In particular, this paper discusses the development and evolution of interrater reliability procedures. The history of these procedures is traced from techniques in use prior to its inclusion in the CODAP system through to the present time. This paper explains the program capabilities in terms of the research requirements which necessitated the enhancements and may be used as an analyst's guide in the future. The paper closes by examining current research streams, potential applications, and the operational use of the newest version of this program aimed at fulfilling these anticipated needs.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AFHRL Technical Paper 81-42 November 1981

INTERRATER RELIABILITY: THE DEVELOPMENT OF AN AUTOMATED ANALYSIS TOOL

By

Michael R. Staley
Johnny J. Weissmuller

Reviewed By

Charles R. Rogers
Chief, Software Development Section
Computer Programming Branch

Submitted for Publication By

Jimmy D. Souter
Chief, Computer Programming Branch

TECHNICAL SERVICES DIVISION
Brooks Air Force Base, Texas 78235

Paper presented at the 23rd Annual Conference of
the Military Testing Association,
26—30 October 1981.

INTERRATER RELIABILITY:
The Development of
An Automated Analysis Tool

Michael R. Staley
Johnny J. Weissmuller

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

INTRODUCTION

Within the United States Air Force, as well as many other military and civilian agencies, there is a fundamental technology which supports occupational analysis for both operational and research programs. The core of this technology is called CODAP, an acronym for the Comprehensive Occupational Data Analysis Programs. The CODAP "system" is a set of analysis tools and procedures which use, as raw material, information provided by members of the occupational field being studied (Thew & Weissmuller, 1979). This system may be used to improve classification structures, assess job-related skills, verify the relevance of training courses and support a host of other applications for which an accurate knowledge of job content at the task level is desirable (Weissmuller, Moore, & Thew, 1980).

Prior to the 1970's, the CODAP system primarily analyzed relative time spent ratings from job incumbents. About 1970, it was decided that the value of CODAP reports would be significantly improved if supervisors' ratings of task importance could be reported alongside of incumbent performance data. Because the concept of "importance" is multifaceted, a technology was required to handle any number of factors which might be considered "important" for tasks in any specific application. One of the more promising "task factors" was Task Learning Difficulty, which allowed supervisors to rate the relative difficulty of training new personnel to perform various tasks within their specialties. A program called TSKDIF was added to the CODAP system to compute the mean task difficulty ratings of supervisors across all tasks in a job inventory. These ratings were used to resequence the tasks in standard job descriptions into descending order based on average task learning difficulty. Hence, within any specified job, Air Force managers could easily see the tasks hardest to learn and the degree to which journeyman level personnel were performing and spending time on those tasks.

About the same time, an interrater reliability program was being developed within the Laboratory. Although reliability theory was developed to measure the trustworthiness of test instruments, the applications within the Laboratory were concerned with assessing the level of agreement within a group of raters. This interrater reliability program allowed researchers to make some statements about the stability of the ratings they were collecting and reporting. In order to increase the credibility of the CODAP reports, it was decided that this procedure should be incorporated into the system and routinely applied to the task learning difficulty ratings. This paper traces the development and evolution of the interrater reliability technique within the CODAP system.

PRE-CODAP VERSIONS: EARLY 1972

One of the earliest interrater reliability programs within the Laboratory was developed by Mr. Manuel Pina. The program was later modified and applied to aptitude rankings of 207 task statements from eight career ladders. This work was accomplished by Mr. Charles A. Greenway in a non-CODAP applications section. Although the use of ratings from 10 research scientists was a modest beginning, the results were promising enough to warrant further development. Subsequent uses in "real-world" settings uncovered some operational shortcomings. The program required that all raters rate ALL tasks. Some analyses had to drop as many as 80 percent of the raters because they failed to meet this condition. As this was deemed unacceptable, it was clear that further changes were required before inclusion in CODAP.

THE FIRST CODAP VERSION: AUGUST - NOVEMBER 1972

Dr. Raymond E. Christal, then Chief of the Occupation and Career and Development Branch of the Personnel Research Division of the Air Force Human Resources Laboratory, assigned to Mr. William J. Phalen the responsibility of integrating an interrater reliability procedure in the CODAP system. Mr. Phalen worked with Master Sergeant (MSgt) William D. Stacey, head of the CODAP Programming Unit, to coordinate the development of the required program. MSgt Stacey assigned the programming task to Airman First Class (A1C) Johnny J. Weissmuller. Many of the details of the first CODAP version were worked out between Mr. Phalen and A1C Weissmuller (Stacey, Weissmuller, Barton & Rogers, 1974).

The new CODAP program (RXXNDX) followed the example of previous versions and reported the interrater reliability coefficients (R_{11} and R_{kk}), computed by Lindquist's intraclass correlation technique (Lindquist, 1953, p.361) and, optionally, the mean rating for each task. In addition, this program provided some new capabilities. It compensated for raters who did not rate every task (Haggard, 1958, p.14) and, optionally, adjusted the ratings of raters who tended to rate either high or low as compared to all other raters. The program also reported the number of tasks rated, the rater's correlation with the grand mean vector, the mean of the rater's ratings and the mean of task means for the tasks rated by the rater.

As the program was nearing completion and several tests were being run, it was noted that the mean of the task means was not 4.0 as would be expected from a seven-point relative scale. The standard deviations also varied widely from one data set to another. It was decided that if two different factors were ever evaluated within a single analysis, it would be important to standardize the task means to some common value with a specific standard deviation. This would allow one to compare different ratings on a given task and determine which factor was "more critical." The program was modified so that if any adjustments were made to individual ratings, the task means would be standardized to a mean of 5.0 and 1.0 standard deviation. This version of the RXXNDX program was released in November 1972.

THE FIRST YEAR REVISIONS: NOVEMBER 1972 - NOVEMBER 1973

During the first year of use, several revisions were required to handle ongoing research projects. One of the most prominent research streams was the investigation of multiple rating scales conducted by Lt. Col. James B. Carpenter (Carpenter, Giorgia, & McFarland, 1975). Interrater reliability measures were desired for ratings on 7-, 9-, 25-, 99- and 9999-point scales. It should be pointed out that the 9999-point scale was actually a direct percent time spent estimate in which two decimals of accuracy (e.g. 45.67) were permitted. The original version of RXXNDX only permitted single digit ratings. These ratings were assumed to be in the range of 1 to 7, with the SETCHK program doing all error-checking and resetting out-of-range values to zero. The RXXNDX program was modified to handle ratings of up to six digits.

Later in the first year, it became apparent that the report of rater correlations with the grand mean vector was effectively identifying deviant raters. As the reliability of a test instrument can be improved by the removal or replacement of poor items, so can the reliability of the mean ratings be improved by the removal of deviant raters. Since the primary measure of stability is the R_{kk}, removing too many raters may actually reduce stability (Haggard, 1958, p.89), and hence the removals should not be done indiscriminately. For this reason, a capability was added to the RXXNDX program to bypass any analyst-specified raters. A t-value column was added to the rater correlation report in order to help analysts evaluate the significance of correlations based on the differing number of tasks rated by each rater. Although the rules-of-thumb varied from analyst to analyst, deviant raters were usually defined to be raters showing a very low (or negative) correlation with the task mean vector across all raters, or those having a t-value of 1.65 or less.

During this period the program name was changed because "RXXNDX" was deemed unpronounceable. As the program was more commonly called "REXALL," the name change was made official to avoid confusion and was justified as "RXXNDX with ALL options." As time went on, however, it became clear that ALL options had not yet been added.

THE EXPLOITATION OF TASK FACTORS: NOVEMBER 1972 - OCTOBER 1976

During this time frame, REXALL became a standard part of the Air Force's operational occupational analysis procedures. By mid-1975 other task factors were being considered and programs were developed to exploit this newly acquired methodology (Christal & Weissmuller, 1976). Some of the motivation for pursuing task factors at this time was the high-level discussion in the Australian Department of Defence aimed at adopting a single occupational analysis system as their DoD standard. Because there were two competing systems being used in the Australian armed forces, their basic choice was between the USAF version of CODAP adopted by the Australian Air Force and an independently developed Australian Army system. A major advantage of the Army system was its ability to effectively handle various task factors. As these capabilities seemed highly desirable, the USAF developed and integrated a generalized task factor capability into the CODAP system. This action may have contributed to the Australians' final decision to adopt the CODAP approach as their standard. The set of programs which comprise this capability has become known as the Task Factor extension or the Task Factor Applications Package.

With the increased emphasis on new factors and the intention of making more use of these data, the practice of removing so-called "deviant" raters came into question. Some researchers felt that the raters who were removed might represent a consistent, valid, albeit minority viewpoint. Squadron Leader (SQN LDR) Kenneth Goody, an Australian exchange officer working at the Laboratory, investigated this hypothesis. He concluded that, in general, raters removed using the REXALL process were simply uncooperative (Goody, 1976).

During the course of this analysis, however, he wondered if the deviant raters were members of a smaller, easily identified subgroup with a different policy. He hoped that these groups might be identified by some background characteristic such as grade level, major command, etc. SQN LDR Goody obtained a copy of the REXALL program and modified it to permit the screening of raters based on predetermined categories. He had trouble in testing these ideas because background data were not routinely collected from the supervisory raters. Even when he obtained the necessary information, his technique required the use of special programming to code each rater according to the category of interest prior to running REXALL. Although this was inconvenient, the potential use of that option was considered important enough to warrant making his version of the program the CODAP standard. Also based on his recommendation, data collection procedures were modified to solicit more background information from raters.

THE IMPACT OF TRAINING EMPHASIS: 1978 - 1979

Training Emphasis is a task factor designed to capture a field supervisor's recommendations about which tasks ought to be included in entry-level training and how much relative emphasis should be placed on each task (Ruck, Thompson, & Thomson, 1978). This task factor departed from the

CODAP standard 1 to 9 rating scale. For this factor, raters were asked to use a scale of 0 to 9, in which zero meant "Do not train." Although this is quite logical, there were some problems in using these data in the REXALL program. REXALL considered ratings of zero to be nonresponses and substantial program changes had to be made. In addition, the SETCHK program which range-checks the data was also modified to permit zero as a valid rating.

Training emphasis continued to impact on REXALL. Although many career fields produced stable vectors of task means, some Air Force specialties failed to achieve a minimally acceptable value for interrater agreement. These specialties were labelled "complex specialties" and further analyses were conducted. It was thought that there might be different rater policies for various subsets of tasks. Various policy-capturing techniques were attempted without much success. One subsequent attempt involved using REXALL to identify not only a reduced set of raters, but also a reduced set of tasks on which there would be general agreement. The REXALL program was modified to make multiple passes of the data and compute interrater agreement on sets of tasks that were rated by at least a specified percentage of raters (e.g. 10%, 20%, 40%). Of course, for these purposes, a zero had to be considered a nonresponse.

Unfortunately, this modification also failed to achieve the desired results. It was next hypothesized that the 0 to 9 scale was inadequate to fully represent the range of emphasis intended by the raters. Because zeroes were averaged in with the typical 1 to 9 ratings, the overall means for tasks only reached a maximum value of about 3 for the most important tasks. The proposed solution was to reweight the responses so that 9's became 512, 8's became 126, and so on. This line of thought was explored by another Australian exchange officer, SQN LDR Michael J. Cassidy, and several such weight substitutions were suggested. The REXALL program was modified to permit this capability, but none of the substitution schemes seemed to solve the problem.

The next approach to complex specialties was to try clustering raters using the standard CODAP procedures. Using these procedures, the research analyst was provided with a cluster-merger diagram from which to identify rater groups of interest. It was hoped that these empirically defined groups would represent the various rating policies. Using this approach, a problem was identified in that the REXALL program was designed to use all raters on the input file with the exception of those it was specifically told to ignore. The clustering programs, on the other hand, produced short lists of raters to be used instead of the longer lists of raters to be ignored. For example, in this analysis groups typically consisted of 20 to 30 raters, while the rest of sample usually contained 300 raters. Rather than require researchers to specify all raters NOT included in each group, a new version of REXALL called "REXNON" was made which used only the raters requested. The new program caused "REXALL" to be reinterpreted as "interrater reliability on ALL raters" and "REXNON" to mean "interrater reliability on NONE but the specified raters."

THE IMPACT OF STRENGTH & STAMINA: 1979 - 1980

The Strength and Stamina research project was designed to help identify career ladders and tasks within career ladders that warranted a further in-depth study of various physical demands (Gott & Alley, 1980). As in the Training Emphasis area, a 0 to 9 scale was used. Because of the lessons learned, however, raters were required to enter an actual zero digit to signify "no lifting requirement" and an "X" to signify "unknown requirement." Again, REXALL had to be modified -- this time to interpret an "X" as a nonresponse. The SETCHK program edits the data file that is later used by the REXALL program. This program also had to be changed and the required changes were not straightforward. For this reason a new version of SETCHK called "SETCKR" was developed.

As this project had short suspenses and spanned nearly all career fields within the Air Force, a large level of effort was required within a short time frame. For this reason a major portion of the logistics was completed under contract. In order to minimize training time, many of the computer runs were preprogrammed to require only clerical support. In the REXALL program, an option was added to record which raters were deemed "deviant" and on subsequent passes, automatically remove those raters.

TRAINING EMPHASIS, CONTINUED: 1980

In the intervening time period, several approaches outside of the REXALL methodology were tried with the training emphasis ratings in the "complex specialties." In 1980 another REXALL-based approach was suggested by William J. Phalen in which only those tasks with low standard deviations would be considered for inclusion in a stable, majority viewpoint. Instead of adding another option like the "percentage of raters" mentioned above, a more generalized approach was taken that reduced the task set based on the values of any specified task factor. For example, with this new option one could investigate the training emphasis for only those tasks that are performed by a high percentage of members. This and other approaches are still being considered but to date, no operational method has been adopted.

CURRENT RESEARCH & FUTURE REQUIREMENTS — REXSPC: 1981 - 1982?

There are two projects currently underway which would be greatly facilitated by program changes. The first is a continuation of the Training Emphasis research project and is being done by the most recent Australian exchange officer, SQN LDR Hans P. Jansen. This analysis is exploring factor analysis techniques and addresses both the question of sample sizes and the reliability of task means. A large number of raters is being randomly subdivided into various groups and the stability of the interrater agreement cofficient is being monitored. Under the present system, approximately six computer runs are required to prepare the data, identify samples, select subsets and compute interrater reliability for each sample. A program called "REXSPC" is currently being developed which will reduce this to two runs. This may not seem significant, but there are over 300 random subsamples to be processed and this could save over 1200 computer runs.

The second project, being done by Captain James H. Gilbert at the USAF Occupational Measurement Center, is designed to evaluate the difference in training emphasis policies between supervisory personnel for maintainance of different aircraft systems. In this analysis, over 20 different aircraft systems are represented, and the intent is to produce a task mean vector for each system. Again, under the current system, this requires three computer runs, while the new REXSPC program would handle this in one run. If Capt. Gilbert's research finds significant differences between these groups of supervisors, this approach may become the operational standard for all future analyses. If this happens, REXSPC will result in a greater savings impact than is expected for the 20 systems currently being studied.

Not all questions have been answered -- indeed, not all questions have yet been asked. New questions usually require new tools; hence, all development has not yet been completed. The interrater reliability capability began simply as a method to produce a single number representing the confidence in mean ratings. It has expanded and is helping to form new guidelines for the types of raters and ratings that are necessary to provide Air Force managers with sound information on which to base their policy decisions. As long as Air Force managers face new challenges, the development and evolution of automated analysis tools will be required to meet their changing needs.

References

Carpenter, J. B., Giorgia, M. J., & McFarland, B. P. **Comparative analysis of the Relative Validity for Subjective Time Rating Scales.** AFHRL-TR-75-63, AD-A017-842. Lackland AFB, TX: Occupational and Manpower Research Division, December 1975.

Christal, R. E., & Weissmuller, J. J. **New CODAP programs for Analyzing Task Factor Information.** AFHRL-TR-76-3, AD-A026-121. Lackland AFB, TX: Occupational and Manpower Research Division, May 1976.

Goody, K. **Comprehensive Occupational Data Analysis Programs (CODAP): Use of REXALL to Identify Divergent Raters.** AFHRL-TR-76-82, AD-A034-327. Lackland AFB, TX: Occupation and Manpower Research Division, October 1976.

Gott, S. P., & Alley, W. E. "Physical Demands of Air Force Occupations: A Task Analysis Approach." Proceedings of the Twenty-Second Annual Conference of the Military Testing Association. Toronto, Canada: Canadian Forces Personnel Applied Research Unit, October 1980.

Haggard, E. A. Intraclass Correlation and Analysis of Variance. New York: Dryden Press, Inc, 1958.

Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin, 1953.

Ruck, H. W., Thompson, N. A., and Thomson, D. C. "The Collection and Prediction of Training Emphasis Ratings for Curriculum Development." Proceedings of the Twentieth Annual Conference of the Military Testing Association. Oklahoma City, OK: United States Coast Guard Institute, October-November 1978.

Stacey, W. D., Weissmuller J. J., Barton, B. B., & Rogers, C. R. **CODAP: Control Card Specifications for the Univac 1108.** AFHRL-TR-74-84, AD-A004-085. Lackland AFB, TX: Computational Sciences Division, October 1974.

Thew, M. C., & Weissmuller, J. J. "CODAP: A Current Overview." Proceedings of the Twenty-First Annual Conference of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center, October 1979.

Weissmuller, J. J., Moore, B. E., & Thew, M. C. "CODAP: Applications and Their Implications for Higher Level Design." Proceedings of the Twenty-Second Annual Conference of the Military Testing Association. Toronto, Canada: Canadian Forces Personnel Applied Research Unit, October 1980.

END

DATE
FILMED

1-82

DTIC