

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

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

Mr. J. Tartell,
Air Force Occupational Measurement Squadron
AFOMS

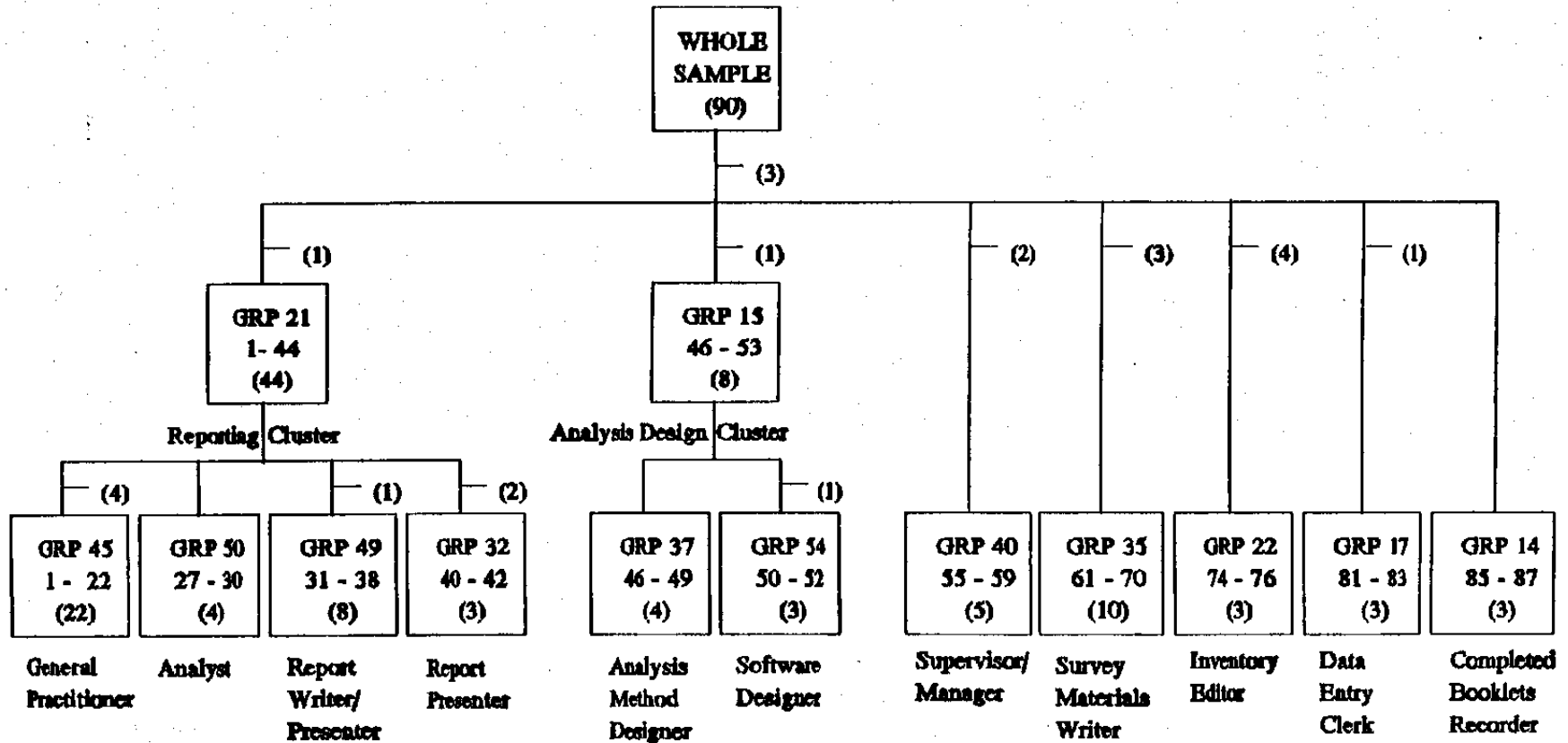


Figure 5. Job-type diagram for CODAP practitioners.

THE CODAP GROUP RELIABILITY (GRPREL) PROGRAM

FUNCTION

GRPREL IS A PROGRAM WHICH COMPUTES A MEASURE OF THE EXTENT TO WHICH A REPRESENTATIVE SAMPLE OF RATERS (WITH RESPECT TO WHAT IS BEING RATED) PRODUCES A COMPARABLE SET OF RATINGS WHEN RATING A COMMON SET OF ITEMS (TASKS)

- COMPUTES AND EVALUATES THE RELIABILITY OF A SINGLE RATER (R_{11}) OR A COMPOSITE OF "K" RATERS (R_{1k}) BASED ON RAW OR ADJUSTED RATINGS
- IDENTIFIES AND REMOVES DIVERGENT RATERS AND/OR TASKS
 - .. RATERS WHOSE CORRELATION WITH THE MEANS OF ALL RATERS ON THE RATED ITEMS IS NOT GREATER THAN ZERO AT $P = .05$.

COMPARISON OF INTER-RATER RELIABILITY (GRPREL)
WITH TEST RELIABILITY IN CLASSICAL MENTAL
TEST THEORY

TEST-RETEST RELIABILITY

HAVE SET OF RATER'S RATE THE SAME SET OF TASKS TWICE.

PARALLEL FORMS RELIABILITY

HAVE DIFFERENT BUT EQUIVALENT SETS OF RATER'S RATE THE SAME SET OF TASKS ONCE.

SPLIT-HALF RELIABILITY (INTERNAL CONSISTENCY)

RANDOMLY SPLIT RATER SAMPLE INTO TWO EQUAL SUBSAMPLES, CORRELATE THE MEAN TASK RATINGS FOR THE TWO SUBSAMPLES, AND USE SPEARMAN-BROWN FORMULA TO ESTIMATE CORRELATION FOR THE FULL SAMPLE SIZE.

KR-20 OR COEFFICIENT ALPHA RELIABILITY

RESULTANT RELIABILITY COEFFICIENT IS EQUIVALENT TO COMPUTING THE AVERAGE OF ALL POSSIBLE SPLIT-HALF RELIABILITIES. THE SAME RESULT IS ACHIEVED IN GRPREL BY COMPUTING R_{KR} WITH ONLY THE RATER MEANS ADJUSTED.

NOTE:

A RATER = A TEST ITEM

ALL RATER'S = A TEST

A TASK = AN EXAMINEE

ALL TASKS WITH NONZERO RATINGS = ALL EXAMINEE

THE CODAP GROUP RELIABILITY (GROREL) PROGRAM

SOME DEFINITIONS

- INTRACLASS CORRELATION: A MEASURE OF THE RELATIVE HOMOGENEITY OF RATINGS WITHIN CLASSES (TASKS) IN RELATION TO THE TOTAL VARIATION AMONG THE INDIVIDUAL RATINGS OF TWO OR MORE RATER'S.
- INTER RATER AGREEMENT: AN INTRACLASS CORRELATION BASED ON RAW (UNADJUSTED) RATINGS
- INTER RATER RELIABILITY: AN INTRACLASS CORRELATION BASED ON ADJUSTED RATINGS
 - EQUALIZING RATER MEANS ONLY
 - EQUALIZING RATER MEANS AND S.D.'S
 - STANDARDIZATION OF EACH RATER'S RATINGS
 - CONVERSION OF EACH RATER'S RATINGS TO RANKS (EQUIVALENT TO KENDALL'S COEFFICIENT OF CONCORDANCE)

$$R_{11} = \frac{KW - 1}{K - 1}$$

$$W = \frac{(K-1)R_{11} + 1}{K}$$

INTRACLASS CORRELATION

$$R_{nn} = \frac{BTMS - WMS}{BTMS + \left(\frac{n}{k} - 1\right) WMS}$$

$$R_{..} = \frac{BTMS - WMS}{BTMS + (n-1) WMS}$$

$$R_{kk} = \frac{BTMS - WMS}{BTMS}$$

INTER-RATER AGREEMENT / RELIABILITY

(ONE-WAY ANOVA WITHOUT REPLICATIONS)

$$R_{nn} = \frac{BTMS - WMS}{BTMS + \left(\frac{\bar{k}}{n} - 1\right) WMS}$$

(GENERAL FORMULA
with built-in Spearman-Brown
Brown extension, i.e., $\frac{\bar{k}}{n}$ -

$$R_{11} = \frac{BTMS - WMS}{BTMS + (\bar{k} - 1) WMS}$$

↑
number of raters
for which reliability
is estimated by
Spearman-Brown
prophecy formula

$$R_{\bar{k}\bar{k}} = \frac{BTMS - WMS}{BTMS}$$

$$R_{kik} = \frac{k R_{11}}{1 + (k-1) R_{11}}$$

$$\bar{k} = \frac{1}{t-1} \left(\sum k_i - \frac{\sum k_i^2}{2k_i} \right)$$

WHERE

t = number of tasks
k_i = number of ratings on task i

$$R_{11} = \frac{F - 1}{F + (\bar{k} - 1)}$$

WHERE

"F" has its usual meaning in ANOVA
df for numerator = df for BTMS
df for denominator = df for WMS

$$F = \frac{1 + (\bar{k} - 1) R_{11}}{1 - R_{11}}$$

where :

R_{nn} = interrater agreement for "n" raters

BTMS = between tasks mean square

WMS = within (tasks) mean square

\bar{k} = average number of raters per task

t = number of tasks

k_i = number of raters per task

EXAMPLES OF SETS OF TASK RATINGS
PRODUCING VARIOUS LEVELS OF R_{11}

I.

TASK	RATINGS			EX
A	4	4	4	12
B	3	3	3	9
C	2	2	2	6
D	1	1	1	3

$R_{11} = 1.00$

II.

TASK	RATINGS			EX
A	3	3	3	9
B	3	3	3	9
C	3	3	3	9
D	3	3	3	9

$R_{11} = .00$

III.

TASK	RATINGS			EX
A	1	4	7	12
B	4	7	1	12
C	7	1	4	12
D	1	7	4	12

$R_{11} = -.50$

IV.

RAW

TASK	RATINGS			EX
A	1	2	3	6
B	2	4	6	12
C	3	6	9	18
D	4	8	12	24

MEAN 2.5 5.0 7.5 5.0

S.D. 0.97 5.92 3.35 3.16

$R_{11} = .36$

(maximum possible disagreement)
maximum negative $R_{11} = \frac{-1}{k-1}$

V.

(RAW RATINGS FROM III. ADJUSTED ON MEANS MEAN = 5.0) (EX UNCHANGED)

TASK	RATINGS			EX	\bar{X}
A	3.5	2.0	0.5	6	2.0
B	4.5	4.0	3.5	12	4.0
C	5.5	6.0	6.5	18	6.0
D	6.5	8.0	9.5	24	8.0

MEAN 5.0 5.0 5.0 ← MEANS CHANGED
S.D. 0.97 5.92 3.35 3.16 ← S.D.'S UNCHANGED

$R_{11} = .79$

VI.

(EX CHANGED)

TASK	RATINGS			EX	\bar{X}
A	3.66	3.66	3.66	10.98	3.66
B	4.55	4.55	4.55	13.65	4.55
C	5.45	5.45	5.45	16.35	5.45
D	6.34	6.34	6.34	19.02	6.34

MEAN 5.00 5.00 5.00 5.00 ← MEANS SAME
S.D. 1.00 1.00 1.00 1.00 ← S.D. CHANGE

$R_{11} = 1.00$