ASSESSING INTERRATER AGREEMENT IN JOB ANALYSIS RATINGS

LEVEL II

L. A. JOHNSON, A. P. JONES, M. C. BUTLER & D. MAIN

REPORT NO. 81-17

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CALIFORNIA 92138

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

Assessing Interrater Agreement in Job Analysis Ratings

Lee A. Johnson, Allan P. Jones,

Mark C. Butler and Debbi Main

Naval Health Research Center

P. O. Box 85122

San Diego, California 92138

Report Number 81-17 was supported under Office of Naval Research Contract RR042-08-01 NR 170-915. Opinions expressed are those of the authors. No endorsement by the Department of the Navy has been given nor should be inferred. The authors would like to acknowledge the assistance of W. M. Pugh, Paul Biner, and Tony Vogel in different phases of the study.

Summary

The present study reviewed some of the commonly used indices of interrater agreement and determined advantages and disadvantages of each when used to assess the reliability of quantitative job analysis scores derived from detailed narrative job descriptions. A listing was obtained of all civilian employee positions and job titles from a medium-sized military hospital employing approximately 1,100 persons. This listing included 89 separate job categories, each uniquely identified by civil service commission job code, pay grade range, and a designator noting whether or not the position was supervisory in nature. Extensive narrative descriptions of each of the 89 job categories were obtained from the U.S. Civil Service Commission Qualifications Standards (1978). These descriptions contained detailed information about the scope of job duties, experience and training requirements, supervisory controls, and general work conditions typically encountered by incumbents in each job category.

Four graduate level psychology students trained in the use of the Position Analysis Questionnaire (PAQ) were asked to rate the above job descriptions. The PAQ is an extensive questionnaire divided into six major sections (Information Input, Mental Processes, Work Output, Relationships with Other People, Job Context, and Other Job Characteristics). Job analysts are required to evaluate different aspects of jobs in each section and to record their evaluations of various job attributes on either five-point Likert type or dichotomous (0-1) scales. Twenty-five jobs were randomly selected for rating by all four raters; the remaining 64 job descriptions were rated by two raters. Each of the indices of interrater agreement reviewed (i.e., percent agreement, Kappa, Weighted Kappa, Pearson product-moment correlation, intraclass correlation, and the Spearman-Brown formula) was computed for comparative purposes. The results indicated general agreement on the ratings obtained, but differences were noted in the estimates produced by the various indices. Reasons for such differences were explored and recommendations made for avoiding potential difficulties in assessing interrater agreement on job analysis ratings.

Assessing Interrater Agreement in Job Analysis Ratings

Research interest in job analysis and job classification is, once again, on the rise (Cornelius, Carron, & Collins, 1979; Tornow & Pinto, 1976). As detailed by Jones, Main, Butler, and Johnson (Note 1), this resurgence of interest is stimulated in part by the need to obtain precise, objective data for use in job development, performance appraisal, and other such situations in an efficient and cost-effective manner. Traditionally, job analysis information has been acquired by having expert observers rate the behaviors of job incumbents or by having incumbents complete lengthy questionnaires such as the Position Analysis Questionnaire (McCormick, Jeanneret, & Mecham, 1972). Issues related to time, cost, intrusiveness, and reliability of the information obtained by either procedure, however, have led to the search for alternative methods for obtaining job analysis data.

One alternative method capitalizes on the fact that many organizations have conducted extensive job analyses, but have reduced such data to detailed narrative descriptions about either individual job requirements or requirements for job families. Although such narrative descriptions are rich with information, they lack the numerical precision necessary for most research and applied uses. While it is possible to again conduct the original job analyses and thus produce more quantitatively oriented information, a more efficient technique would be the development of numerical ratings based on existing job descriptions (cf. Jones, et al., Note 1).

The utility of such a method requires at a minimum that trained raters agree on the ratings to be assigned to each position (i.e., one must address interrater reliability). However, a critical issue in determining agreement is the selection of a statistical method appropriate to the data collected (Jones & James, 1979). Because of the many indices of rater agreement in common use today, the problem of selecting an appropriate index seems particularly difficult. The purpose of the present study is to review some of the commonly used indices of interrater agreement and to determine advantages and disadvantages of each when used to assess agreement in job analysis ratings derived from narrative descriptions.

Indices of Interrater Agreement

Percent agreement is a commonly used index but has a number of problems associated with it (Cohen, 1960; Cohen, 1968; Jenkins, Nadler, Lawler, & Cammann, 1975; Mitchell, 1979; Repp, Deitz, Boles, Deitz, & Repp, 1976). First, this statistic includes chance agreement and may overestimate true agreement among observers. Paradoxically, percent agreement may also underestimate interrater reliability on rating scales requiring complex decisions because it fails to consider partial agreement.

To compensate for the shortcomings of percent agreement, Cohen (1968) introduced Kappa ($\underline{K}$) and Weighted Kappa ($\underline{K\omega}$). These statistics assess agreement after correcting for chance, while $\underline{K\omega}$ allows partial credit to ratings which are similar but not identical. On the surface, $\underline{K}$ seems to resolve the problems faced by percent agreement. However, $\underline{K}$ assumes that the selection of a category in one observation is independent of the selection of a category on successive observations. Further, observations should not be concentrated in a few cells. Such assumptions may be easily violated if one is conducting job analyses on similar jobs that are relatively few in number. Assume that two individuals are rating a homogenous set of jobs on a single five-point scale. Further, assume that the attribute is obvious and easy to rate. The homogeneity of the rated group increases the probability that one may find a high proportion of scores occurring in one or two cells while other cells have small or nonexistent proportions. This example violates the assumptions upon which $\underline{K}$ is based and will tend to reduce the power of the test to a point of providing little useful information (Overall, 1980).

The Pearson product-moment correlation coefficient (r) has also been used by a number of authors as an index of interrater agreement (Bernardin, Alvares, & Cranny, 1976; Cornelius, et al., 1979; Jenkins, et al., 1975; Taylor & Colbert, 1978). However, Ebel (1951) noted that only pairwise comparisons can be made, the table of ratings must be complete, and the between-rater variance is always removed in calculating the product-moment formula. Unfortunately, Ebel pointed out situations where between-rater variance should be included as part of the error term. Namely, when comparisons are made among single raw scores assigned to different subjects by different raters, the between-rater variance should be included in the error term. For these reasons, Ebel (1951) and others (Bartko, 1966, 1976; Selvage, 1976; Winer, 1971) advocated the use of the intraclass correlation as an index of interrater agreement because it permits the researcher to choose whether or not to include the between-rater variance.

Despite this advantage, the intraclass correlation (because of its basis in the ANOVA paradigm) suffers many of the same criticisms mentioned earlier for K. For example, Selvage (1976) noted that ANOVA fails to produce coefficients which reflect the consistency of ratings when assumptions of normality are grossly violated. Further, ANOVA requires substantial between-item variance to produce a significant indication of agreement (Finn, 1970; Selvage, 1976). Thus an example similar to the one given earlier for K would also violate the underlying assumptions of ANOVA and distort the intraclass correlation as an index of interrater agreement.

Lastly, Jones and James (1979) noted that indices of reliability of mean scores among raters (e.g., the Spearman-Brown formula) have also been used in assessing interrater agreement. They argued, however, that "the Spearman-Brown formula is not well suited to large numbers of raters. ... large sample sizes tend to yield high estimates of mean score reliability even when relatively heterogeneous individual scores are used to compute such means. Thus, the reliability of the mean scores appears to provide an overly optimistic estimate of agreement." (1979, p. 207). Jones and James, after reviewing many of the same indices discussed above, concluded that no single index was adequate for all situations. They suggested that more than one index of interrater agreement may be necessary to assess reliability among raters.

On the basis of such recommendations and because the weaknesses of some of the indices appear to be offset in other indices of agreement (and vice versa), we computed several of the available indices to determine whether the combination of trained raters and structured job analysis questionnaires could produce a viable method for translating the detailed narrative descriptions of jobs present in many organizations to quantified data useful in performance assessment, job classification, and so forth.

Method

A listing was obtained of all civilian employee positions and job titles from a medium-sized military hospital employing approximately 1,100 persons. This listing included 89 separate job categories, each uniquely identified by civil service commission job code, pay grade range, and a designator noting whether or not the position was supervisory in nature. Extensive narrative descriptions of each of the 89 job categories were obtained from the U.S. Civil Service Commission Qualifications Standards (1978). These descriptions contained detailed information about the scope of job duties, experience and training requirements, supervisory controls, and general work conditions typically encountered by incumbents in each job category.

Four graduate level psychology students trained in the use of the Position Analysis Questionnaire (PAQ) were asked to rate the above job descriptions. The PAQ is an extensive questionnaire divided into six major sections (Information Input, Mental Processes, Work Output, Relationships with Other People, Job Context, and Other Job Characteristics). Job analysts are required to evaluate different aspects of jobs in each section and to record their evaluations of various job attributes on either five-point Likert type or dichotomous (0-1) scales. Twenty-five jobs were randomly selected for rating by all four raters; the remaining 64 job descriptions were rated by two raters. Each of the indices of interrater agreement discussed earlier was computed for comparative purposes.

Results

Table 1 presents a summary of the reliability estimates obtained on the PAQ ratings using each technique. These estimates are presented in terms of the number of coefficients lying within a given range. In general $\underline{K}$ and $\underline{K\omega}$ provided the most conservative estimates of interrater agreement, producing median values of .19 and .22, respectively. Conversely, percent agreement and Spearman-Brown approaches provided the most optimistic indications of interrater agreement, with median values of .63 and .73, respectively. Finally, the average pairwise correlation and intraclass techniques occupied an intermediate position as indicated by median values of .51 and .39.

Table 1

FREQUENCY DISTRIBUTION OF ESTIMATES OF INTERRATER AGREEMENT ON THE PAQ

	Kappa		Weighted Kappa		Percent Agreement		Average Correlation		Intraclass Correlation		Spearman Brown	
Range	Freq.	Cum %	Freq.	Cum %	Freq.	Cum %	Freq.	Cum %	Freq.	Cum %	Freq.	Cum %
.91 - 1.00	1	0.7	1	0.7	8	5.9	2	1.5	1	0.7	12	8.8
.81 - .90	-	0.7	-	0.7	18	19.1	5	5.1	5	4.4	27	28.7
.71 - .80	-	0.7	-	0.7	26	38.2	17	17.6	6	8.8	31	51.5
.61 - .70	-	0.7	4	3.7	25	56.6	18	30.9	12	17.6	18	64.7
.51 - .60	5	4.4	5	7.3	13	66.2	27	50.7	16	29.4	3	66.9
.41 - .50	10	11.8	17	19.8	21	81.6	20	65.4	24	47.1	8	72.8
.31 - .40	14	22.0	22	36.0	12	90.4	9	72.1	22	63.2	7	77.9
.21 - .30	35	47.8	23	53.0	10	97.8	11	80.1	6	67.6	11	86.0
.11 - .20	35	73.5	29	74.3	3	100 0	13	89.7	15	78.7	2	87.5
≤ .10	36	100.0	35	100.0	-	100.0	14	100.0	29	100.0	17	100.0
Median Value		.19		.22		.63		.50		.39		.73

While the above results describe general trends, an examination of estimates obtained for each item suggested additional trends of interest. As shown in Table 2, some of the indices appeared to be more sensitive to characteristics of the underlying distribution than others. When there was high agreement among raters and limited variance on the characteristic being rated (Condition A), all of the indices except percent agreement tended to produce unrealistically low estimates of agreement. With greater variance on the rated item, there was considerably more consistency in the estimates obtained in either favorable (high agreement among raters; Condition B) or unfavorable (low agreement among raters; Condition C) directions.

Table 2

Item Level Comparisons Between Estimates of Interrater Agreement and Underlying Distributional Characteristics

Distributional Characteristic Conditions	INDEX: Kappa	Weighted Kappa	Percent Agreement	Average Correlation	Intraclass Correlation	Spearman-Brown
A. Homogeneity across jobs and high agreement among raters.[1]	.04	.04	.77	-.01	-.03	-.14
B. Heterogeneity across jobs and high agreement among raters.[2]	.43	.45	.88	.79	.74	.92
C. Heterogeneity across jobs and low agreement among raters.[3]	.01	.04	.16	.13	.05	-.23

[1]Calculations based on ratings of PAQ item number 27, "Body balance (sensing the position and balance of the body when body balance is critical to job performance, as when walking on beams, climbing high poles, working on steep roofs, walking on slippery floors, etc.), rated in terms of "importance to this job."

[2]Calculations based on ratings of PAQ item number 21, "Far visual differentiation (seeing differences in the details of objects, events, or features beyond arm's reach, for example, operating a vehicle, landscaping, sports officiating, etc.), rated in terms of "importance to this job."

[3]Calculations based on ratings of PAQ item number 153, "Non-job required social contact (the opportunity to engage in informal, non-job required conversation, social interaction, etc. with other while on the job, for example, barber, taxi driver, receptionist, journeyman and apprentice, etc.; do not include personal contacts required by the job), rated in terms of "frequency of opportunity."

Table 3

Frequency Distribution for Estimates of Interrater Agreement for PAQ Dimension Scores

	Average Correlation		Intraclass Correlation		Spearman-Brown	
	Freq.	Cum. %	Freq.	Cum. %	Freq.	Cum. %
.91 - 1.00	1	3.4	-	0.0	13	44.8
.81 - .90	8	23.1	3	10.3	6	65.5
.71 - .80	5	48.3	10	44.8	5	82.8
.61 - .70	7	72.4	4	58.6	4	96.5
.51 - .60	4	86.2	2	65.5	1	100.0
.41 - .50	3	96.5	5	82.8	-	100.0
.31 - .40	1	100.0	2	89.6	-	100.0
.21 - .30	-	100.0	3	100.0	-	100.0
.11 - .20	-	100.0	-	100.0	-	100.0
.0 - .10	-	100.0	-	100.0	-	100.0
Median		.70		.63		.89

The foregoing results indicate that at an item level, only under one condition (Condition B in Table 2) did interrater agreement approach acceptable levels. Thus, reliability estimates were also calculated for dimension scores (see McCormick, et al., 1972). These estimates, shown in Table 3 in terms of the number of coefficients lying within a given range, were generally higher than those obtained for items and appeared to represent acceptable levels of agreement.

Discussion

The present study explored the reliability of job analysis scores based on narrative job descriptions. The computation of several of the commonly used indices, however, posed difficulties for drawing straightforward conclusions about interrater agreement. First, some indices tended to provide generally higher or lower estimates than others. For example, the median estimates provided by percent agreement and the Spearman-Brown correction were more than three times the magnitude of those provided $\underline{K}$ and $\underline{K\omega}$. Second the relative differences among the indices varied according to the nature of the distribution.

All indices appeared sensitive to situations where there was little or no agreement among the raters, yielding reliability estimates that were low and similar in magnitude. Likewise, when agreement was high and there was heterogeneity on the characteristics being rated, all indices yielded relatively high reliability coefficients. Even in these cases, however, $\underline{K}$ and weighted $\underline{K\omega}$ provided generally lower estimates than the others. The greatest difficulty came in situations where there was substantial homogeneity on the characteristic and the raters were able to reach high agreement. In this condition, percent agreement tended to reflect the actual level of interrater agreement whereas the other estimates tended to be quite low.

These findings suggest that the user must exercise caution in selecting and interpreting indices of interrater agreement. When there is heterogeneity among jobs on the characteristics being rated, the user need be sensitive only to the degree to which a particular index is likely to be systematically higher or lower than another. However, when there is reason to suspect that the jobs being rated will be quite similar, some of the estimates may lead to erroneous conclusions that rater agreement is low. Thus, the user must consider the underlying distribution of the ratings in reaching conclusions about rater agreement.

Bearing such points in mind, it appears that the raters were able to reach at least moderate levels of agreement in developing scores from narrative descriptions of complex jobs. For the dimension scores at least, the estimates appear consistent with indices reported in the literature. For example, Smith and Hakel (1979) reported mean reliabilities for the PAQ (based on pairwise correlations) ranging from .49 to .63. Taylor and Colbert (1978) used a similar technique and reported an average correlation of .68 among pairs of raters while McCormick, et al., (1972) reported mean reliabilities of .80. Based on these values, the reliabilities of the present effort appeared at least acceptable.

It should be noted, however, that many different and often interrelated factors exist which can affect or distort any index of reliability. In the present context, for example, differences in understanding of the rating format and instructions provided by the PAQ or difference in the education, experience, and other such background characteristics between raters are plausible sources of confounding in obtaining identical information from narrative job descriptions. To avoid, or at least minimize, problems of this type Kaye (1980) noted the importance of establishing the reliability of data (regardless of type) at the same level that it will be used to address research questions. Our results suggest that trained raters are able to use narrative job descriptions to develop quantitative job profiles, but only at the job dimension or job characteristic composite level. Thus, such ratings likely to differentiate primarily among types of jobs. The poor reliability associated with item level analysis, coupled with the fact that the PAQ is based on generic terminology and lacks job specific language (Levine, Ash, & Bennett, 1980) would tend to obscure the subtle differences

that are necessary to distinguish among jobs of the same type.

The degree to which the loss of such distinctions is problemmatic depends upon the use to which the job analysis is put. Many of the common uses of such data are, in fact, more consistent with a job family approach than with an ipsative job approach. For example, Levine, et al., (1980) compared the utility of different job analysis techniques for the development of selection test batteries. They concluded that "... no matter how rich in detail a job analysis report may be, resultant exam plans will not vary enough to produce significantly different evaluations of their quality." (p. 534). Similar conclusions are likely for situations where job analysis is used for setting salary levels, or designing training programs.

Insofar as many of the applied and research uses of the PAQ are concerned with the classification of jobs into similar families, the use of narrative job descriptions to derive quantitative scores about such families provides a relatively quick, inexpensive, and unobtrusive means of obtaining needed data. Thus, the technique appears to overcome many of the objections to the current reliance on job incumbents or trained observer for settings where detailed information about relevant jobs already exists.

Reference Notes

1. Jones, A.P., Main, D.S., Butler, M.C., & Johnson, L.A. Narrative job descriptions as potential sources of job analysis ratings (Report No. 81-13). San Diego: Naval Health Research Center, May, 1981.

References

Bartko, J.J. The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 1966, 19, 3-11.

Bartko, J.J. On various intraclass correlation reliability coefficients. Psychological Bulletin, 1976, 83, 762-765.

Bernardin, H.J., Alvares, K.M., & Cranny, C.J. A recomparison of behavioral expectations scales to summated scales. Journal of Applied Psychology, 1976, 61, 564-570.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70, 213-220.

Cornelius, E.T., Carron, T.J., & Collins, M.N. Job analysis models and job classification. Personnel Psychology, 1979, 32, 693-708.

Ebel, R.L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Finn, R.H. A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 1970, 30, 71-76.

Jenkins, G.D., Nadler, D.A., Lawler, E.E. & Cammann, C. Standardized observations: An approach to measuring the nature of jobs. Journal of Applied Psychology, 1975, 60, 171-181.

Jones, A.P., & James, L.R. Psychological climate: Dimensions and relationships of individual and aggregated work environment perceptions. Organizational Behavior and Human Performance, 1979, 23, 201-250.

Kaye, K. Estimating false alarms and missed events from interobserver agreement: A rationale. Psychological Bulletin, 1980, 88, 458-468.

Levine, E.L., Ash, R.A., & Bennett, N. Exploratory comparative study of four job analysis methods. Journal of Applied Psychology. 1980, 65, 524-535.

McCormick, E.J., Jeanneret, P.R., & Mecham, R.C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 1972, 56, 347-368.

Mitchell, S.K. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 1979, 86, 376-390.

Overall, J.E. Power of Chi-Square test for 2 x 2 contingency tables with small expected frequencies. Psychological Bulletin, 1980, 87, 132-135.

Repp, A.C., Deitz, D.E.D., Boles, S.M., Deitz, S.M., & Repp, C.F. Differences among common methods for calculating interobserver agreement. Journal of Applied Behavior Analysis, 1976, 9, 109-113.

Selvage, R. Comments on the analysis of variance strategy for the computation of intraclass reliability. Educational and Psychological Measurement, 1976, 36, 605-609.

Smith, J.E., & Hakel, M.D. Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. Personnel Psychology, 1979, 32, 677-692.

Taylor, L.R., & Colbert, G.A. Empricially derived job families as a foundation for the study of validity generalization: Study II. The construction of job families based on company-specific PAQ job dimensions. Personnel Psychology, 1978. 31, 341-353.

Tornow, W.W., & Pinto, P.R. The development of a managerial job taxonomy: A system for describing, classifying, and evaluating executive positions. Journal of Applied Psychology, 1976, 61, 410-418.

U.S. Civil Service Commission. Qualification Standards for Positions under the General Schedule. Washington, D.C.: U. S. Government Printing Office, 1978. (Handbook X-118).

Winer, B.J. Statistical Principles in Experimental Design (2nd Ed.) New York: McGraw-Hill, 1971.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 81-17	2. GOVT ACCESSION NO. AD-A108 794	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Assessing Interrater Agreement in Job Analysis Ratings		5. TYPE OF REPORT & PERIOD COVERED Technical Report 05-80 to 12-80
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Lee A. Johnson, Allan P. Jones, Mark C. Butler and Debbi Main		8. CONTRACT OR GRANT NUMBER(s) ONR
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Health Research Center P.O. Box 85122 San Diego, CA 92138		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Medical Research & Development Command Bethesda, MD 20014		12. REPORT DATE 5/22/81
		13. NUMBER OF PAGES 9
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Bureau of Medicine & Surgery Department of the Navy Washington, DC 20372		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

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

Job Analysis
Interrater Agreement
Reliability

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

The present study explored the reliability of quantitative job analysis scores derived from detailed narrative job descriptions. Descriptions of 25 different jobs in a medium-sized military hospital were scored by four trained raters using the Position Analysis Questionnaire. Interrater agreement was assessed using a number of common indices. These indices suggested general agreement on the ratings, but differences were noted in the estimates produced. Reasons for such differences were explored and recommendations were

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 68 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

made for avoiding potential difficulties in assessing interrater agreement on job analysis ratings.