© Journal of the Indian Academy of Applied Psychology
2015, Vol. 41, No.3 (Special Issue), 20-27.

Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods

Shweta
National University of Study and Research in Law,
Ranchi

R.C. Bajpai
Army College of Medical Sciences,
New Delhi

H. K. Chaturvedi
National Institute of Medical Statistics,
New Delhi

Evaluation of inter-rater agreement (IRA) or inter-rater reliability (IRR), either as a primary or a secondary component of study is common in various disciplines such as medicine, psychology, education, anthropology and marketing where the use of raters or observers as a method of measurement is prevalent. The concept of IRA/IRR is fundamental to the design and evaluation of research instruments. However, many methods for comparing variations and statistical tests exist, and as a result, there is often confusion about their appropriate use. This may lead to incomplete and inconsistent reporting of results. Consequently, a set of guidelines for reporting reliability and agreement studies has recently been developed to improve the scientific rigor in which IRA/IRR studies are conducted and reported (Gisev, Bell & Chen, 2013; Kottner, Audige, & Brorson, 2011). The objective of this technical note is to present the key concepts in relation to IRA/IRR and to describe commonly used approaches for its evaluation. The emphasis will be more on the practical aspects about their use in behavioral and social research rather than the mathematical derivation of the indices.

Keywords: Inter-Rater Agreement, Inter-Rater Reliability, Indices.

Although practitioners, researchers and policymakers often used the two terms IRA and IRR interchangeably, but there is a technical distinction between the terms agreement and reliability (LeBreton & Senter, 2008; de Vat, Terwee, Tinsely & Weiss, 2000). In general, IRR is defined as a generic term for rater consistency, and it relates to the extent to which raters can consistently distinguish different items on a measurement scale. However, some measurement experts defined it as the measurement of consistency between evaluators regardless of the absolute value of each evaluator's rating. In contrast, IRA measures the extent to which different raters assign the same precise value for each item being observed. In other words, IRA is the degree to which two or more evaluators using the same scale assigns the same rating to an identical observable situation. Thus, unlike IRR, IRA is a measurement of the consistency between the absolute value of evaluator's ratings. The distinction between IRR and IRA is further illustrated in the hypothetical example in Table 1 (Tinsley & Weiss, 2000).

Table 1. A hypothetical example of differences between reliability and agreement

	Low Agreement, High Reliability		High Agreement, High Reliability	
	Rater 1	Rater 2	Rater 3	Rater 4
Teacher A	1	2	1	1
Teacher B	2	3	2	2
Teacher C	3	4	3	3
Teacher D	4	5	4	4
Agreement	0.0		1.0	
Reliability	1.0		1.0	

In Table 1, the agreement measure shows how frequently two or more evaluators assign exactly the same rating (e.g., if both give a rating of "4" they are in agreement), and reliability measures the relative similarity between two or more sets of ratings. Therefore, two evaluators who have little to no agreement could still have high IRR. In this scenario, Raters 1 and 2 agree on the relative performance of the four teachers because both assigned ratings increased monotonically, with Teacher A receiving the lowest score and Teacher D receiving the highest score. However, though they agreed on the relative ranking of the four teachers, they never agreed on the absolute level of performance. As a consequence, the level of IRR between Raters 1 and 2 is perfect (1.0), but there is no agreement (0.0). By contrast, Raters 3 and 4 agree on both the absolute level and relative order of teacher performance. Thus, they have both perfect IRR (1.0) and IRA (1.0).

Another way to think about the distinction that IRA is based on a "criterion-referenced" interpretation of the rating scale: there is some level or standard of performance that counts as good or poor. On the other hand, IRR is based on a norm-referenced view: the order of the ratings with respect to the mean or median defines good or poor rather than the rating itself. Typically, IRA is more important in high-stake decisions about performance and planning whereas IRR is more frequently used in the research studies where only interest is the consistency of rater's judgments about the relative levels of performance (Gwet, 2012).

Measurement of key indices

The following methods are commonly used to calculate IRR/IRA indices:

Reliability

If a measurement procedure consistently assigns the same score to individuals or objects with equal values, the instrument is considered reliable. In other words, the reliability of a measure indicates the extent to which it is without bias and hence insures consistent measurement across time and across the various items in the instrument. It is an indication of the stability (or repeatability) and consistency (or homogeneity) with which the instrument measures the concept and helps to assess the "goodness" of a measure (Shekharan & Bougie, 2010; Zikmund, 2003).

$$\text{Reliability}=\frac{(\text{Subject variability})}{(\text{Subject variability}+\text{Measurement error})} \tag{1}$$

Percent Agreement

The percentage of absolute agreement is the simplest to understand (Altman, 1991). One simply calculates the number of times raters agree on a rating, then divides by the total number of ratings. Thus, this measure can vary between 0 and 100%. Other names for this measure include percentage of exact agreement and percentage of specific agreement. It may also be useful to calculate the percentage of times the ratings fall within one performance level of one another (e.g., count as agreement cases in which rater one gives Teacher-A 4 points and rater two gives Teacher-A 5 points). This measure has been called the percentage of exact and adjacent agreement. When there are more than 4 or 5 rating levels, exact and adjacent agreement may be a more realistic measure to use. Also, there is no limit to the number of raters that can be assessed (Gisev et al, 2013).

$$\text{Percent agreement}=\frac{(\text{Number of concordant responses})}{(\text{Total number of responses})}\times 100 \tag{2}$$

The Kappa Index

Kappa measurements are one of the original and most commonly used IRA indices. In 1960, Cohen proposed and described the Kappa Index for nominal categorical variables assessed by two raters. Since then, a number of modifications have been proposed, and the term Kappa now refers to a group of indices. These indices provide a chance-corrected index of IRA and are based on the ratio of the proportion of times the agreement is observed to the maximum proportion of times that the raters could agree (both corrected for chance agreement) (Siegel & Castellen, 1988).

$$\text{Kappa (k)}=\frac{(\text{Proportion observed agreement-Proportion expected chance agreement})}{(\text{1- Proportion expected chance agreement})} \tag{3}$$

Kappa can take any value between -1 and +1, where +1 indicates perfect agreement.

However, mathematically, a value of -1 is difficult to achieve and is only observed in extreme circumstances. Furthermore, the lower limit of Kappa varies and is dependent on the number of categories. Negative values indicate that the observed agreement is less than that expected from chance alone; a value of 0 indicates exact chance agreement, and positive values indicate that the observed agreement is greater than that expected from chance (Cohen, 1960).

There is a wide distinction in the elucidation of Kappa values, and several efforts have been made to assign practical meaning to calculated Kappa values. The most comprehensive and widely accepted interpretation was proposed by Landis and Koch in 1977. This classification is often simplified into three categories such that a Kappa value of 0.75 or greater is considered to represent an excellent level of agreement, a value of 0.40 or less is indicative of poor agreement, and values between 0.40 and 0.75 represent fair to good agreement (Fleiss, Levin, & Paik, 2003). However, because of the inherent properties of the Kappa formula, it has been suggested that this upper limit is unnecessarily high and realistically may not be achievable in the context of some research studies (LeBreton & Senter, 2008). Hence, a low Kappa value may not always be indicative of low agreement.

Table 2. Interpretation of Kappa values proposed by Landis and Koch

Kappa	Interpretation
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Furthermore, Kappa is sensitive to the bias between raters and the overall prevalence of responses (Byrt, Bishop & Carlin, 1993). In some situations, a relatively high proportion of observed agreement can result in a low Kappa value and an unbalanced or biased distribution of responses can result in a higher kappa value than a more balanced distribution of responses (Feinstein & Cicchetti, 1990). To assist in the interpretation of Kappa values and identify potential bias, the reporting of average proportions of agreement for positive and negative responses is recommended in addition to the overall Kappa value (Cicchetti & Feinstein, 1990). These proportions, referred to as ppos and pneg, respectively, are calculated by dividing the number of positive (or negative) ratings observed by the mean number of positive (or negative) ratings. For example, ppos and pneg were reported in a study evaluating the IRA of physician's responses to a tool identifying potentially inappropriate prescriptions given to older people (Gallagher, Baeyens & Topinkova, 2009). In this study, raters agreed in most cases, the particular criterion in question was not fulfilled, producing a heavily skewed distribution of responses. Reporting of ppos and pneg was therefore necessary to accurately interpret the results of the analyses. Another index, the prevalence-adjusted bias-adjusted Kappa (denoted as PABAK), has also been proposed to correct for any potential bias in the Kappa value for dichotomous variables assessed by two raters (Byrt et al, 1993). As an example, the PABAK was often used by researchers comparing ratings on the presence or absence of specific psychiatric disorders related problems in behavioral sciences. In addition, a version called generalized Kappa can compare groups of more than two raters. Three assumptions must be met when using Kappa:

1. The items to be rated are independent.
2. The categories are independent, mutually exclusive and exhaustive.
3. The raters are independent.

Similar to other statistical tests, Kappa values should be reported with the corresponding standard error and hypothesis testing undertaken to determine statistical significance (Siegel & Castellen, 1988). The null hypothesis that Kappa would equal to 0 against the alternative hypothesis that Kappa is greater than 0 is tested. Rejection of the null hypothesis therefore indicates that any agreement observed is statistically significant. Information on hypothesis testing and calculation of the standard error and confidence interval for Kappa are detailed elsewhere and commonly included in the output provided by statistical software such as SPSS, SAS, and R etc (Sheskin, 2007).

Kendall coefficient of concordance (W)

The Kendall coefficient of concordance is suitable for ordinal variables assessed by multiple raters (Siegel & Castellen, 1988). The 'W' score provides an indication of the strength of agreement and is interpreted with its corresponding P-value. It scores between 0 and 1, where 0 indicates no agreement and 1 signifies complete agreement. Negative W values are impossible because complete disagreement cannot be achieved with more than two raters. It becomes increasingly harder to achieve high W scores when the number of raters increases, and consequently, low W scores can become significant (Schmidt, 1997). When testing the significance of the W score, the null hypothesis that the ratings of the different judges are independent of one another is tested. Rejection of the null hypothesis (using a 1-tailed test) therefore enables one to conclude that any agreement observed between the judges is statistically significant.

Various interpretations and reporting of the results of tests using the Kendall coefficient of concordance have been reported in the literature. The Landis and Koch (1977) interpretation of Kappa categories has been extended to the interpretation of W scores. Furthermore, an interpretation linking the W score and confidence in rankings has been proposed (Schmidt, 1997). Additionally, the W score has also been interpreted analogously to the correlation coefficient.

In practice, the Kendall coefficient of concordance has been used in various disciplines such as social and behavioral science studies employing a panel of experts to make judgments on rankings of an item. For example, in one study, a four-member multidisciplinary expert panel was employed to assess the expected outcomes of comprehensive medication reviews for clients of community mental health teams. Using a 5-point Likert type scale, each panelist independently assessed review findings, review recommendations, likelihood of recommendation implementation, and the overall expected clinical impact. Agreement among panelists was established with W scores for each of the scales. Similarly, another useful application of the Kendall coefficient of concordance is in the conduct of multiple-round Delphi surveys (Schmidt, 1997). The W score can be used to determine whether consensus has been reached, whether consensus is increasing between rounds, and also the relative strength of the consensus.

Bland-Altman plots

Calculating a Pearson's product-moment correlation coefficient (r) seems a logical choice to assess the level of agreement of two raters for interval or ratio data. However, use of the Pearson's correlation coefficient is inappropriate as an IRA index because it indicates the strength of the relationship. To overcome these limitations and accurately evaluate IRA, Bland and Altman (1999, 1986) proposed an alternative approach that relies on graphically plotting scores.

Each point on the line is derived by plotting the difference in scores of the two raters (x) against the average of the two scores (y). The magnitude of disagreement and any outliers and trends in scores can then be determined from the graph. Additionally, the 95% limits of agreement can be estimated by calculating the mean difference ±1.96 multiplied by the standard deviation of the differences, providing an interval in which 95% of the differences in ratings are expected to lie, provided that the differences are normally distributed. A nonparametric alternative has also been described. For example, in the context of medical research, Bland-Altman plots have been used to determine agreement between blood pressure measurements taken by community pharmacists and values obtained through ambulatory and home blood pressure recordings.

Intra-class correlation coefficient (ICC)

The ICC is widely reported in literature and is used to measure agreement when there are many rating categories (5 or more) or when ratings are made along a continuous scale (e.g., one that allows ratings of rational numbers such as 2.3, 2.4, 2.5, etc) or when there are missing ratings. Based on analysis of variance (ANOVA) models, the ICC was originally applied to the evaluation of differences between interval or ratio variables (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Similar to the other tests described, ICC values should be reported with corresponding P-values or confidence intervals. When measuring rater agreement, the ICC represents the proportion of the variation in the ratings that is due to the performance of the person being evaluated rather than factors such as how the rater interprets the rubric. Subtracting the ICC from 1 gives the proportion of variation between raters that occurs due to rater disagreement. ICC scores generally range from 0 to 1, where a 1 indicates perfect agreement, and a 0 indicates no agreement. There are several versions of the ICC exist, so it is important to choose the appropriate one depending on whether:

1. To treat the data as a 1-way or 2-way ANOVA model; and
2. The absolute value or consistency of ratings is important; and
3. The unit of analysis is an individual rating or the mean of several ratings.

Based on the above criteria, a comprehensive flowchart has been developed by McGraw and Wong (1996) to assist in the selection of an appropriate ICC, each with their own specific formula for calculation. Each type of ICC can be explained by 1 of 3 underlying models that stem from the typical IRA/IRR scenario of a number of raters independently assessing a random sample of items:

1. ***One-way random effects model***- Each item is assessed by a different set of randomly selected raters.
2. ***Two-way random effects model***- Each item is assessed by all raters who have been randomly selected from a larger population of raters.
3. ***Two-way mixed model***- Each item is assessed by all raters in the population of interest.

Under certain conditions, the ICC has shown to be equivalent to Cohen's Kappa, weighted Kappa, and the Kendall coefficient of concordance (Sheskin, 2007; Fleiss & Cohen, 1973). It has also been argued that the ICC should replace Cohen's Kappa and weighted Kappa because it offers greater flexibility in data analysis (Streiner, 1995). However, in doing so, the fundamental rules regarding levels of measurement need to be disregarded, although the outcomes and interpretation of the results may not differ significantly. Furthermore, the application of the ICC to nonparametric contexts is a developing field of research and new indices continue to be developed. An example of the use of the ICC to social and administrative pharmacy research includes establishing the IRR of a newly developed Medication-Based Disease Burden Index for the quantification of disease burden using chronic drug therapy data.

Selection of an appropriate index

There is debate in the statistical literature about the applications and appropriateness of the different IRA/IRR indices and their derivatives. Although, some are strictly only valid as IRA measures (e.g., Bland-Altman plots), others have been used in the literature as measures of both IRA and IRR (e.g., Cohen's Kappa). The main questions to consider when selecting an IRA/IRR index are:

1. What is the purpose of the analysis?
2. Is the absolute value or trend in ratings important?
3. What type of variable is being analyzed?
4. How many raters are involved?

General considerations

IRA/IRR values can be interpreted in a number of ways and ranges, indicating that degrees of agreement and reliability are arbitrary. Therefore, it may not be possible to predefine an acceptable level of agreement or reliability. Rather, a judgment should be made regarding the interpretation of IRA/IRR values, considering the nature of the study and possible implications of the results. Another point to be acknowledged is that the high IRA indicates the

Table 3. Strengths and weaknesses of commonly used method of measuring IRA/IRR

Index	Concept	Advantages	Limitations
Percent absolute agreement	How often do raters agree on the exact rating?	**1.** Easy to calculate when number of raters and rating levels is small. **2.** Easy to interpret. **3.** Best measure to use when many ratees receive the same rating.	**1.** Hard to calculate and interpret if there are very many categories. **2.** Does not take chance agreement into account, so may overestimate the agreement that can be expected in the future. **3.** Does not distinguish between a 1-level disagreement and a 2- or more level disagreement.
Cohen's Kappa	How well do raters agree, corrected for chance agreement?	Kappa is a better estimate of the agreement that might be expected from raters rating a different group of ratees.	**1.** Hard to calculate and interpret if there are many rating levels. **2.** Can be misleadingly low if a large majority of ratings are at the highest or lowest level.
Intra-class correlation	What proportion of the variation in rating is due to ratee performance rather than rater error?	**1.** Easier to calculate than other measures when there are a lot of raters and has 5 or more levels. **2.** The only measure that works well when ratings are on a continuous scale.	**1.** Requires some understanding of statistics to calculate. **2.** Can be misleading if there is low variation in ratings across ratees.

rater's concordance on a particular response. Thus, all the raters may be applying the same (incorrect) reasoning when scoring items. Furthermore, the results of an IRA/IRR analysis are unique to the individual study. They are a function of the population of interest and are dependent on the raters, the responses, and the rating scale used. IRA/IRR values are therefore not generalized to other studies (Tinsley & Weiss, 1975).

Many times one method may not be considered the best under all circumstances therefore, it is often appropriate to get calculations of more than one measure. For example, if the ICC is lower than expected, calculating the percentage of absolute agreement can show whether the problem is low in agreement or is limited in variation in the performance ratings. Typically, if there are four or fewer discrete rating levels, Kappa and the percentage of absolute agreement should both be calculated. If there is a moderate number of performance levels (e.g., 5-9), one could use the ICC as well as the percentage of absolute agreement. If scores are on a continuous scale, then one should always use the ICC to calculate inter-rater agreement. After inter-rater agreement is calculated using the ICC, one can group the scores into categories based on expected thresholds for consequences (e.g., the scores required for rewards, tenure, or triggering remediation). Based on the groupings, one can calculate the percentage of absolute agreement by dividing the number of times raters placed individual teachers in the same performance category by the total number of teachers observed.

Level of Acceptable Agreement

There are no rigid rules regarding the level of agreement that is needed to use a set of ratings to make high-stakes decisions or to consider the assessment process reliable. There are two types of benchmarks that one can use

Table 4. Rule of thumb for determining whether IRA is sufficient for consequential use of ratings

Agreement Summary Statistic	Minimum (%)	High (%)	Comment
Absolute agreement	75	90	There should be no ratings more than 1 level apart. If there are more than 5-7 rating levels, an absolute agreement level closer to 75% would be acceptable, but the exact and adjacent agreement should be close to 90%.
Cohen's kappa	61	81	Since the value of Kappa depends in part on how ratings are distributed across levels, high values should not be expected if most of the ratings are at one level.
Intra-class correlation	80	90	As the value of the ICC depends in part on the variation of ratings across ratees, high values should not be expected if many ratees get the same rating.

to judge how much agreement is sufficient. One rule of thumb suggested by various experts is: while using percentage of absolute agreement remember the values from 75-90% demonstrate an acceptable level of agreement (Stemler, 2004). For Kappa, popular benchmarks for high agreement are 0.80 (Altman, 1991; Landis & Koch, 1977). There are fewer consensuses among the researchers on a sufficient ICC score. A score of 0.70 would be sufficient for a measure used for research purposes, but some researchers advocate a value of 0.8 or 0.9 as a minimum while using scores for making important decisions (Hays & Revicki, 2005). A second benchmark is to compare the levels of agreement researchers have reported in the literature on assessing practice. Table 4 summarizes the thresholds for each of the methods of calculating inter-rater agreement.

Factors Affecting IRA/IRR

As mentioned earlier, it is important to recognize that neither it is possible nor cost-effective to achieve the perfect agreement. Some degree of professional judgment is necessary if ratings are to represent different levels of complex behavior. However, the evaluation system administrators can take many concrete steps to improve the consistency of evaluation results. In general, there are three important factors that affect IRA/IRR indices significantly while evaluating agreement among ratees:

1. Rater training
2. Rater selection
3. Accountability for accurate rating

Conclusion

IRA and IRR relate to two different concepts. The absolute value is important in the assessment of IRA, whereas the consistency of ratings is important in the evaluation of IRR. Opinions on the appropriateness and suitable applications of IRA/IRR indices vary and are prompted by the fact that several indices produce similar results. Selection of an index therefore needs to be justified, bearing in mind the context and purpose of the study, as well as ease of calculation and interpretation of the results.

References

Altman, D. G. (1991). *Practical statistics for medical research* (reprint 1999). CRC Press: Boca Raton, Florida.

Bland, J. M., & Altman D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research. 8*, 135-160.

Bland, J. M. & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet. 1*, 307-310.

Byrt, T., Bishop J. & Carlin, J. B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology. 46,* 423-429.

Cicchetti, D. V. & Feinstein, A. R. (1990). High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology. 43*, 543-549.

Cohen, J. (1960). A coefficient for nominal scales. *Educational and Psychological Measurement. 20*, 37-46.

de Vet H. C. W., Terwee, C. B. & Knol D. L., & Lex, M. Boute (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology. 59*, 1033–1039.

Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low Kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology. 43*, 543-549.

Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted Kappa and the Intra-class correlation coefficient as measures of reliability. *Educational and Psychological Measurement. 33*, 613–619.

Fleiss, J. L., Levin, B. & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*, (3rd, ed.). John Wiley: New York.

Gallagher, P., Baeyens J. P. & Topinkova, E. (2009). Interrater reliability of STOPP (Screening Tool of Older Persons' Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment) criteria amongst physicians in six European countries. *Age Ageing, 38*, 603–606.

Gisev, N., Bell J. S. & Chen, T. F. (2013). Inter-rater agreement and inter-rater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy. 9*, 330–338.

Gwet, K. L. (2012). *Handbook of Inter-rater Reliability,* (3rd, ed.). Advanced Analytics, LLC: Gaithersburg, MD.

Hays, R. D. & Reviki, D. A. (2005). *Reliability and validity (including responsiveness)*. In P. M. Fayers, & R. D. Hays (ed.). Assessing quality of life in clinical trials: Methods and practice. Oxford University Press: New York.

Kottner, J., Audige, L. & Brorson, S. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology. 64*, 96–106.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics. 33*, 159-174.

LeBreton, J. M. & Senter, J. L. (2008). Answers to 20 questions about inter-rater reliability and inter-rater agreement. *Organization Research Methods. 11*, 815–852.

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intra-class correlation coefficients. *Psychological Methods. 1*, 30–46.

Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. *Decision Science Journal. 28*, 763-774.

Shekharan, U. & Bougie, R. (2010). *Research Methods for Business: A Skill Building Approach,* (5th ed.). John Wiley: New Delhi.

Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*, (4th ed.). Chapman & Hall: Florida.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin. 86*, 420–428.

Siegel, S. & Castellen, N. J. Jr. (1988). Nonparametric *Statistics for the Behavioral Sciences*, (2nd ed.). McGraw-Hill Book Co: New York.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*, 4.

Streiner, D. L. (1995). Learning how to differ: Agreement and reliability statistics in psychiatry. *Canadian Journal of Psychiatry. 40*, 60-66.

Tinsley, H. E. A. & Weiss, D. J. (1975). Inter-rater reliability and agreement of subjective judgments. *Journal of Counseling Psychology. 22,* 358-376.

Tinsley, H. E. A. & Weiss, D. J. (2000). *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Academic Press: San Diego. CA.

Zikmund, W.G. (2003). *Business Research Methods,* (7th ed). Thompson South-Western: Ohio.

Shweta , PhD, Natinal University of Study and Research in Law, Ranchi

Bajpai RC, Army College of Medical Sciences, New Delhi. Email: rambajpai@hotmail.com

Chaturvedi HK, National Institute of Medical Statistics, New Delhi