Testing The Accuracy Enhancement of Pairwise Comparisons by a Monte Carlo Experiment [1]

Waldemar W. Koczkodaj
Laurentian University, Department of Mathematics and Computer Science
Ramsey Lake Rd., Sudbury, Ontario P3E 2C6
icci@nickel.laurentian.ca

Abstract

A statistical experiment was designed to check if the pairwise comparisons method, which was introduced by Fechner in 1860 and developed by Thurstone in 1927, improves the accuracy of estimation of stimuli. The experiment has been designed and implemented to minimize statistical bias. The accuracy improvement by the pairwise comparisons method (when compared with the direct rating method) is decisive: the mean value of the improvement exceeds 500% and a 95% confidence interval is (4.657, 5.389).

Keywords: pairwise comparisons, rating, accuracy, Monte Carlo, hypothesis testing

1.The Method of Pairwise Comparisons

The method of pairwise comparisons introduced by Fechner in 1860 [Fechner, 1860] and worked out by Thurstone in 1927 [Thurstone, 1927] was a milestone in inferencing. It may be compared to the introduction of differentation in calculus or eigenvalues in linear algebra. The pairwise comparisons method can always be used to draw the final conclusions in a brilliant and yet comparatively straightforward way as shown by numerous examples (see, for example, [Hwang and Yoon, 1981, Nijkamp, 1991]. The ingenuity of the pairwise comparisons method can be reduced to a common sense rule: take two (stimuli, criteria, alternatives, etc.) at a time whenever handling everything at once is more difficult. Its practical significance is even greater since there are situations where direct measurements are impossible. No one questions the practicality of measures of length (such as meter or foot), or weight (kg or pound) since they are in common use. We have become so accustomed to having standards that sometimes we find it difficult to imagine anything without a standard measure. In real life, however, there are many objects lacking standard measures. The environment or environmental pollution are good examples of situations where a standard yardstick seems to be missing. For example, it would be hard to use a cubic meter of environment as a standard measure since in one cubic meter of environment there could be millions of ants but only a fraction of an elephant. How could one decide if a fraction of an elephant is less significant than a colony of ants?

---

[1]This project was partially supported by the Natural Science and Engineering Council of Canada under grant OGP 0036838 and by the Ministry of Northern Development and Mines through the Northern Ontario Heritage Fund Corporation

It is intriguing, however, why such a natural and powerful tool has never become widely accepted by decision makers despite its extreme practicality. One of the reasons may be the lack of convincing scientific evidence that better accuracy can be achieved with the pairwise comparisons method than without it.

2. Is it really so hard to observe the accuracy improvement?

Intuitively it is obvious that considering two at a time is a better approach than considering all at once for any kind of comparison. To show that the pairwise comparisons method is superior to the common sense by an expert's eye approach, however, is not a trivial task. Even Saaty (who is probably the greatest single contributor to the popularization of the pairwise comparisons method) failed to provide convincing scientific evidence in [Saaty, 1980] (considered by some researchers as one of the most comprehensive sources for theoretical aspects of pairwise comparisons). Saaty's attempt to show the superiority of the pairwise comparisons method by estimating areas of different shapes, however, falls short from a statistical point of view as it introduces a bias towards the direct rating method. His statistical experiment with five different shapes (a square, two rectangles, a triangle, and a circle) for area estimation has several shortcomings which are common to other similar studies conducted in the past. First of all, most respondents suspect trickery in the shapes (some kind of optical illusion) and in trying to outsmart the system, provide incorrect answers. A more important problem is in the unfortunate quantitative selection of the actual areas. The true values of the areas are: 47%, 5%, 23.4%, 14.9% and 9.6%. These are rather hard for estimation, while entering 2, 1.5, 5, 9, etc. as the approximations for pairwise comparisons gives practically 0% error. From a mathematical point of view, all numbers are equally hard (or easy) to guess; however, rounded numbers (such as integers or integers plus 0.5) are more frequent random selections than numbers such as 23.4% or 9.6%. Therefore it is practically impossible to achieve the same level of accuracy without pairwise comparisons. From a statistical point of view this kind of bias is unacceptable and so the result is just a meaningless confirmation of the superiority of the pairwise comparisons method over direct estimations when, as this paper shows, a more convincing alternative exists.

The main goal of the presented experiment is to compare the accuracy of judgments based on the pairwise comparisons method with the direct method which is also referred to as by eye estimation for an obvious reason. At the current stage of pairwise comparisons theory it is impossible to prove analytically the superiority of the pairwise comparisons method, since we are coping with *subjective judgments* and no general metric (or measure) exists for subjective judgments or tastes. In fact existence of such a measure of taste would not solve the problem entirely since we need to compare two different approaches to measurements and humans have a habit (known as learning from experience) of memorizing and using former results in the next step.

It is not easy to design a statistical experiment for showing that the pairwise comparisons method gives better results than the direct rating method without a bias toward one of them. Not only do we need to avoid the optical illusion trap (leading to the shape simplicity assumption), but we cannot allow the solution to be easier to guess for one of the approaches. The former remarks strongly

imply that randomly generated simple shapes should be used in the experiment. Moreover, in each trial, the respondent is asked to answer a number of questions.

The pairwise comparisons method allows the respondent to concentrate his/her attention on only two items (two bars) at a time. Therefore it would be necessary to print each pair on a separate sheet of paper, which would result in an experiment 50 pages long. Presenting the experiment by a traditional hardcopy questionnaire is thus impractical. Yet generating new pairs for each respondent is desirable from the statistical point of view. Thus under the above constraints, the only reasonable solution is to implement the experiment as a computer program. This program should have a proper graphical interface and be installed on portable computers (or on a network) accessible by as many respondents as necessary for the statistical analysis. Each respondent may run the program as many times as he or she wishes and the results are appended to databank. The necessity of a computer implementation may explain why the problem of accuracy had not been properly addressed in the 1950's or 1960's when most of the theoretical work on the pairwise comparisons method took place.

3. Pairwise Comparisons and Weights

In the pairwise comparisons method, stimuli (for example, criteria or alternatives) are presented in pairs to one or more referees (e.g., experts or decision makers). It is necessary to evaluate individual alternatives, derive weights for the criteria, construct the overall rating of the alternatives and identify the best alternative. Let us denote the stimuli by $A_1, A_2,..., A_n$ (n is the number of compared stimuli), their actual weights by $_1, _2,..., _n$ and the matrix of the ratios of all weigths by $=[_i/ _j]$. The matrix of pairwise comparisons $\mathbf{A}=[a_{ij}]$ represents the intensities of the expert's preference between individual pairs of alternatives (A_i versus A_j, for all i,j=1,2,...n) chosen usually from a given scale. The elements a_{ij} are considered to be estimates of the ratios $_i/ _j$ where is the vector of actual weights of the stimuli, which is what we want to find. All the ratios are positive and satisfy the reciprocity property: $a_{ij}=1/a_{ji}$ i,j=1,2,...,n. Saaty's eigenvector solution of $\mathbf{A} =$ always exists if the consistency (or transitivity) condition $a_{ij}*a_{jk} = a_{ik}$ (i,j,k=1,...,n) is satisfied. More details about the problem of inconsistent judgments and definitions of inconsistency can be found, for example in: [Saaty, 1977], [Koczkodaj, 1993], and the convergence analysis of the new inconsistency in [Holsztynki and Koczkodaj, 1996] .

A number of different methods have been recommended for the translation of inconsistent judgments arranged in the pairwise comparisons matrix **A** into a numerical scale. The common feature of all the methods is that for a positive reciprocal matrix $\mathbf{A} = [a_{ij}]$ a vector $= (_1, _2,..., _n)$ is determined such that the matrix of ratios $[_i/ _j]$ is a close approximation to **A** according to some metric. The following methods can be used for finding the vector of weights :

In the eigenvector method (introduced in [Saaty, 1977]), the vector of weights is a eigenvector corresponding to the eigenvalue $_{max}$ of maximum modulus of the matrix **A.** According to the Perron-Frobenius Theorem (see, [Strang, 1988]), the eigenvalue $_{max}$ is positive and real. Furthermore, the vector can be chosen with all positive coordinates. It is a normalized solution of the following equation:

$$A \quad \max$$

The least squares method (computationally challenging) which minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \left(a_{i,j} \quad \frac{i}{j} \right)^2$$

subject to

$$\sum_{i=1}^{n} {}_i \quad 1, \qquad {}_i > 0, \qquad i \ 1,2,\ldots,n.$$

The geometric means method (introduced in [Rabinowitz, 1976] and also known as the logarithmic least squares method) where the approximating vector has elements of the form

$${}_i \quad \left(\prod_{j=1}^{n} a_{i,j} \right)^{1/n} \qquad\qquad i \ 1,\ldots,n$$

The vector is usually normalized so that the sum of the elements is one. It can be proved (e.g. [Jong, 1984]), that the vector , with elements defined by the above equations, is the solution of the problem of minimizing the sum of squares

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \left(\log a_{i,j} \quad \log {}_i \quad \log {}_j \right)^2$$

4. Designing a Statistical Test

The statistical experiment (briefly described in [Koczkodaj, 1996]) compares two methods of estimating the weights with a minimum of statistical bias:

the direct rating method where a respondent is asked to estimate a stimulus directly,

the pairwise comparisons method based on the result of pairwise comparisons of all combinations of the stimuli.

The direct rating method distributes a constant number of points (let us assume 100) among the objects to be distinguished in such a way that the number of points allocated to an object reflects its relative importance. For the pairwise comparisons method, the input for the calculation of weights is a pairwise comparisons matrix **A**. Saaty's eigenvector method was chosen for estimating the lengths. The choice of a particular method is not critical as the study by Krovak [Krovak, 1987] has established that for pairwise comparisons matrices with a high degree of consistency, the weights derived by different methods are close to each other. The inconsistency according to Saaty's definition, ($_{max}$-n)/(n-1), was used to verify the reliability of our computations. The acceptable

threshold for Saaty's inconsistency is considered to be 0.1 (see [Saaty, 1977] for details). In our case it was far below 0.01.

Randomly generated bars have been used since it is assumed that everyone is capable of estimating lengths. The program (implemented in C on personal computers) was installed in one of the computer laboratories at Laurentian University for students to access as part of their homework assignment. It generates combinations of three, four, five, and six bars of random lengths which are stored in internal tables. All combinations of bars with a ratio of lengths greater than five are discarded since some bars would have been undistinguishable on the screen. This is of no practical significance since the same sets of bars are estimated by both methods. In the direct rating method, the respondent is repeatedly asked to estimate the lengths of the presented bars by answering the following question: *What percent (out of 100%) would you assign to the bar shown in red (with the remaining part of the bar displayed for the point of reference) ?* The program normalizes the estimates to 100%. This procedure is repeated for each number of bars.

After the completion of the direct rating phase, the same bars are presented again but this time one pair at a time on the screen. The program displays all *n*(n-1)/2* combinations of pairs with the respondent being asked to answer *How many times bar A is longer than bar B?* The program builds a pairwise comparisons matrix and computes the lengths of the n bars as the eigenvector corresponding to the dominant eigenvalue of the matrix following Saaty's algorithm. The entire procedure is repeated for all the combinations of three, four, five and six bars.

The exact lengths of the randomly selected bars which were stored in a table are used to compute the error for both methods as follows. Let us assume, that a real vector $\mathbf{v} = [v_1,v_2,...,v_n]$ is approximated by the vector $\mathbf{w} = [w_1,w_2,...,w_n]$. Then the error of this approximation can be defined by the formula

$$d([w_1,w_2,...w_n],[v_1,v_2,...v_n]) \quad \sum_{i=1}^{n} \frac{w_i - v_i}{v_i}$$

In the context of the experiment, the right hand side can be interpreted as the sum of the absolute values of the relative differences between the generated lengths and the estimated lengths for each bar. The statistical difference between mean errors for each method , the relationship between the number of estimated bars and the group mean error, and the hypothesis that the estimation error of the pairwise comparisons method is substantially smaller than of the direct rating method are examined.

5. Interpretation of Results

The following numbers of records were collected and analysed: 139 for three bars , 138 for four bars, 129 for five bars, and 127 for six bars. The decreasing number of examples is due to some students giving up when the number of replies became too large. After some initial processing (mostly with a spreadsheet program) to locate and eliminate occasional outliers, the results were transferred to a statistical system (SPSS-X , see [Noru is, 1988]) for further analysis. Each observation consists of three variables: the estimated lengths by the pairwise comparisons and the

direct rating methods, and the number of bars. Fig. 1 shows the histogram for the two methods. The results of each group were merged together (for illustrative purposes and to shorten the length of the paper) since the histograms for each group of bars are very similar as far as shape is concerned. The statistical analysis was, however, conducted separately for each group of bars (that is 3, 4, 5, and 6). The histogram shows that the mean error for the direct rating method is greater than the mean error for the pairwise comparisons method.

Figure 1. The histogram of the combined data

The histogram for the pairwise comparisons method (black bars) has a high degree of regularity. The majority of observations are concentrated in an narrow interval (3.758156, 4.207322). No explanation for the sizable irregularities in the histogram plot for the direct rating method (hashed bars) can be given other than that estimating lengths by eye is not as easy as we tend to believe.

Table 1 below summarizes the descriptive statistics for the experiment. A graph of the means for both methods is displayed in Fig. 2. The K-S test value is the Kolmogorov-Smirnov test statistic for normality. The last row of the table shows the critical values of K-S test for a significance level of =0.05. According to [Siegel and Castellan, 1988] it is computed as 1.36/ n for the number of observations n 36.

Number of bars	3		4		5		6	
Number of observations	139		138		129		127	
Mean error	PC	DR	PC	DR	PC	DR	PC	DR
	4.150	11.583	4.092	13.166	3.902	15.219	3.763	16.587
Standard deviation	2.866	6.195	2.671	7.572	2.507	7.918	2.485	8.905
K-S *z* value	1.185	1.412	1.412	1.419	1.560	1.587	1.584	1.760
critical value for =0.05	0.115		0.116		0.120		0.121	

Table 1. The descriptive statistics (PC stands for the pairwise comparisons, and DR stands for the direct rating estimation)

On the basis of the results shown in Table 1, the hypothesis about normality of the distribution should be rejected, for each group of bars (3, 4, 5, and 6), for the assumed level of significance =0.05. The Kolmogorov-Smirnov test was selected because of its popularity. The more powerful D'Agostino test (see [D'Agostino, 1990] for details) gave the same conclusions. These results are necessary for selecting the proper procedures for further statistical analysis of means.

A population's mean value describes the population's centre or location. In this experiment, the mean of a sample is an estimator of the error for each method. A hypothesis about the means of two samples of these populations must be formulated and tested to show a statistical difference between the two different methods. Let μ_1 and μ_2 be the expected values of the error for the pairwise comparisons and for the direct rating method respectively and $_1$ and $_2$ the corresponding standard deviations of the errors. The null hypothesis, H_0: $\mu_1=\mu_2$, that there is no significant difference between these two methods is tested against the alternative hypothesis H_1: $\mu_1>\mu_2$. The Wilcoxon matched-pairs test for analysing two dependent samples was selected since it is considered to be one of the most powerful of the nonparametric procedures for a set of paired observations. It should be noted that this test performs better for normal distributions. However, the differences between the averages are quite decisive and according to an Internet discussion with experienced statisticians, the applicability of the Wilcoxon matched-pairs test could only be in question for smaller differences of means. A more precise method would simply discard the equality hypothesis with a higher level of significance. The results of this test are presented in Table 2.

Number of bars	number of examples	value of z	rank analysis			
			shown by	PC < Dir	PC > Dir	PC = Dir
3	139	-9.9521	cases	129	10	0
			mean rank	74.41	13.15	-
4	138	-9.7736	cases	130	7	1
			mean rank	71.35	25.36	-
5	129	-9.8551	cases	129	0	0
			mean rank	65.00	0	-
6	128	-9.7643	cases	125	3	0
			mean rank	64.98	3.00	-

Table 2. Results of the Wilcoxon matched-pairs test

No assumption has been made about the type of distribution. The critical value for z for a significance level of 0.0001 is 3.72, so the hypothesis H_0: $\mu_1=\mu_2$ should be rejected (with a 2-Tailed p=0.000 for every group of bars). It is also worthwhile noting that whenever the Wilcoxon matched-pairs test leads to rejection of the null hypothesis about equality then another, possibly more powerful test, would also reject this hypothesis.

Figure 2. Mean errors for both methods for each group of 3, 4, 5, and 6 bars

A second population parameter that is often considered is ^{2}, the population variance, a measure of the population dispersion. Two distributions can have the same value for measures of central tendency (mean value) and yet be very dissimilar in other respects. For example, if the estimated error for both methods is the same, say 10%, but the variance of first method is twice as large as that of the second method, then the second method is considered better since it generates more reliable results (less dispersed). For the experiment with bars, the null hypothesis H_0: $_1^2 = _2^2$, stating that there is no significant difference between two variances, is tested against the alternative hypothesis H_1: $_1^2 \quad _2^2$. Most widely available tests about homogeneity of variances (like F_{max}) are sensitive to departures from normality. Since we have rejected the hypothesis about normality of the distribution in our case, Scheffé's test (e.g., [Winer, 1971]) has been employed. Scheffé's test is conservative for pairwise comparisons of means. It requires larger differences between means for significance than most of the other methods. The Scheffé multiple comparison procedure in SPSS-X shows that *no two groups are significantly different at the 0.0001 level* for the pairwise comparisons method while the estimates in groups of three and six bars are *significantly different at the 0.0001 level* for the direct rating method.

We can also investigate if there is a relation between the error of the method and the number of bars compared in the experiment. This can be rephrased to: *Is the number of bars an essential factor in estimating the lengths of the bars?* Usually an ANOVA answers this type of conjecture, but under the assumption of normality of both distributions. Since this assumption does not apply to our case (see the results shown in Table 1), the Kruskal-Wallis test, a nonparametric analog of ANOVA, was used (see Table 3).

Method	type of analysis	value H	prob(H >Chi-square)	degrees of freedom
pairwise comparisons	regular	1.1106	0.7745	3
	corrected for ties	1.1110	0.7744	
direct ranking	regular	33.1798	0.0000	
	corrected for ties	33.1814	0.0000	

Table 3. The results of Kruskal-Wallis test

The Kruskal-Wallis test indicates that the hypothesis H_0 that *there is no difference in the medians of errors for the groups of 3, 4, 5, and 6 bars* cannot be rejected (probability 0.77) for the pairwise comparisons method. However, H_0 is decisively rejected for the direct method (probability 0.0000).

The results of this test are consistent with reality since the number of objects compared at a time by the pairwise comparisons method is constant (two) while the number of bars in the direct rating method increases from three to six.

Finally, we create a new variable *improvement* (*impr*) computed as the error for the pairwise comparisons method divided by the error for the direct rating method for each observation. It passes a normality test with p=0.0001 and therefore a confidence interval for the population mean can be computed (see Tab. 4).

Variable	observations	mean	std error	95% confidence interval
improvement	533	5.023	0.1863	(4.657 , 5.389)

Table 4. A 95% confidence interval for the *improvement* variable computed for all observations

6. Conclusions

The statistical results of this experiment decisively favour the pairwise comparisons method. The mean error for the pairwise comparisons method was about 3.98% versus 14.07% for the direct rating method. Furthermore, the mean error of the pairwise comparisons method decreases with the an increasing number of bars while the mean error of the direct rating method increases with the number of bars. The standard error for the pairwise comparisons method is 0.114 versus 0.342 for the direct method. The mean value of the accuracy improvement is computed as 5.02 (that is over 500%) and the 95% confidence interval for the accuracy improvement is (4.657, 5.389). As a consequence, the argument about the superiority of the direct rating method over the pairwise comparisons method, when the number of factors to be compared increases, is powerless. While the direct rating method may be straightforward to use, there are trade offs in terms of accuracy and reliability. Further inference may be useless (if not dangerous), if the accuracy at this stage is low. The pairwise comparisons method seems to be more difficult for a novice. For example, some respondents (for no apparent reason) answered 0 instead of 1 for equal bars. This problem can be easily overcome with additional training.

It has been tacitly assumed that the number of compared stimuli should not exceed a certain number (it is usually assumed to be seven) because of the increasing number of combinations involved (which is $O(n^2)$). A hierarchical structure is used if the number of criteria is larger, which is usually the case in most applications (see [Saaty, 1977]). This raises a question about the influence of the hierarchical structure itself on the accuracy. It would be useful to know if a hierarchical structure with more groups having fewer members generates a lower (or higher) error that a hierarchical structure with fewer groups with but with more members in each group. This is an area of research which needs further exploration .

The other possible approach, when dealing with a large number of criteria, would be to cluster them into two groups at each level. This is equivalent to the procedure for building a binary hierarchy

tree. There is no statistical evidence that this type of hierarchy tree would perform worse (or better) than the hierarchy with the number of criteria in each group equal to a maximum of seven. Comparing the algorithms based on these two different assumptions could be an interesting research problem.

The low value of inconsistency implies that the resulting error between the actual lengths of the randomly selected bars and the lengths estimated by the respondent (in the experiment) depends only on the method of assessing the weights: pairwise comparisons versus direct rating. With a substantially lower standard deviation, the pairwise comparison method gives the practitioner more confidence in the final outcomes of his/her judgments.

7. Acknowledgments

The author of this paper wishes to thank M. Herman for verifying some of the statistics and for his comments on the draft of this paper. Gratitude is also expressed to Z. Duszak for computing some of the statistics during his postdoctoral fellowship at the Elliot Lake Field Station of Laurentian University and to Laurentian University students for their participation in the experiment involving the estimation of randomly generated bars.

8. References

D'Agostino, R. (1990), *A Suggestion for Using Powerful and Informative Tests of Normality*, American Statistician, Vol. 44, No. 4, pp.316-323.

Duszak, Z., Koczkodaj, W.W. (1994), *The Generalization of a New Definition of Consistency for Pairwise Comparisons*, Information Processing Letters, Vol. 52, pp. 273-276.

Duszak, Z., Koczkodaj, W.W. (1994), *A Consistency-driven Approach to the CD-ROM Selection*, Library Software Review, Vol. 13, No. 4, pp. 262-268.

Fechner, G.T. (1860), Elements of Psychophysics, Vol. 1, New York: Holt, Rinehart & Winston, 1965, translation by H.E. Adler of Elemente der Psychophysik, Leipzig: Breitkopf und Hartel, 1860.

Hwang, C-L. and Yoon, K. (1981), *Multiple Attribute Decision Making*, Springer-Verlag, Berlin.

Hwang, G-J. (1992), *Knowledge Elicitation and Integration from Multiple Experts*, in Proc. ICCI'92, Computing and Information, eds. W. Koczkodaj at al., IEEE Computer Society Press, pp. 208-211

Jong, P. de, (1984), *A Statistical Approach to Saaty's Scaling Method for Priorities*, Journal of Mathematical Psychology 28, pp.467-478.

Koczkodaj, W.W. (1993), *A New Definition of Consistency of Pairwise Comparisons*. Mathematical and Computer Modelling, Vol. 18, 7, pp.79-84.

Koczkodaj, W.W., (1996), *Statistically Accurate Evidence of Improved Error Rate by Pairwise Comparisons,* Perceptual and Motor Skills, 82, pp.43-48.

Krovak, J. (1987), *Ranking alternatives - Comparisons of different methods based on binary comparison matrices*, European Journal of Operational Research, 32, pp.96-99.

Noru is, M.J., (1988), SPSS-X Introductory Statistics Guide for SPSS-X Release 3, SPSS Inc.

Rabinowitz, G., (1976), Some Aspects of World Influence. Journal of Peace Science, 2, 49-55.

Saaty, T.L. (1977), *A Scaling Methods for Priorities in Hierarchical Structure*, Journal of Mathematical Psychology, vol. 15, pp.234-281.

Siegel, S., Castellan, J.N., (1988), Nonparametric Statistics for Behavioural Sciences, 2nd edition, McGraw-Hill.

Strang, G., (1988), Linear Algebra and its Applications, third ed., Harcourt Brace Jovanovich College Publishers, New York.

Thurstone, L.L. (1927), *Law of Comparative Judgements*, Psychological Review, 34, pp.273-286.

Winer, B.J. (1971), *Statistical Principles in Experimental design*. New York, John Wiley and Sons.

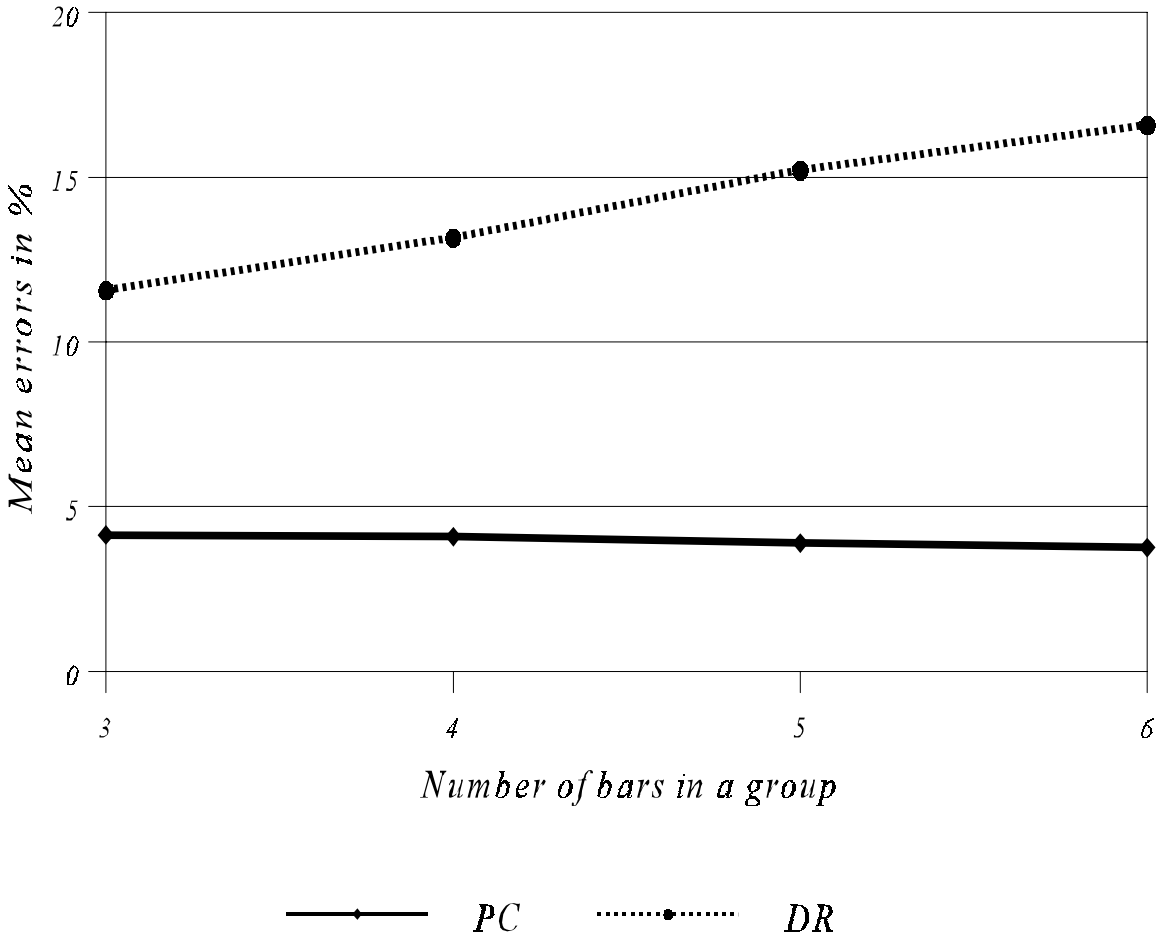

Figure 2. Mean errors for both methods for each group of 3, 4, 5, and 6 bars