A RANK-BASED METHOD FOR THE USABILITY COMPARISON OF COMPETING PRODUCTS

James R. Lewis
International Business Machines, Corp.
Boca Raton, FL

This paper describes a method for a rank-based analysis of data for competitive usability evaluation. The technique works with data collected during scenario-based usability studies. In a scenario-based study, participants are asked to perform realistic tasks (scenarios) with products. Dependent measures commonly include such variables as time-on-task, successful task completion rates, and subjective ratings, reported at the scenario level. Scenario-based studies are sometimes used to set benchmarks or testable behavioral objectives for products. Multivariate statistics such as discriminant analysis can be used with raw data to determine if one product differs from another on the basis of patterns of dependent variables; however, multivariate statistics cannot be used to demonstrate that one product is more usable than another if the designs are based on different usability tradeoffs. Converting raw data to ranks allows the establishment of rank-weighting schemes that combine different dependent measures and allows the assessment of relative product usability. The data that are generated can be analyzed with rank statistical methods. The elimination of various types of biases associated with missing data is also presented. This method of analyzing competitive usability data is a mixture of the subjective and the objective. To use the method, several subjective decisions concerning rank weighting and control of biases must be made before applying the objective part of the method. Application of the method allows a single composite number representing relative usability to be assigned to a product, simplifying product usability comparison.

INTRODUCTION

This purpose of this paper is to describe a method for rank-based analysis of competitive usability data. The technique works with data collected during scenario-based usability studies. In a scenario-based study, participants are asked to perform realistic tasks (scenarios) with products. Dependent measures commonly include such variables as time-on-task, successful task completion rates, and subjective ratings, reported at the scenario level. Scenario-based studies are sometimes used to set benchmarks or testable behavioral objectives for products (Gould, 1988; Lewis, Henry, and Mack, 1990; Whiteside, Bennett, and Holtzblatt, 1988).

THESIS

Multivariate statistics such as discriminant analysis (Cliff, 1987) use raw data to determine if one product differs from another on the basis of patterns of dependent variables. However, multivariate statistics cannot be used to demonstrate that one product is more usable than another if their designs are based on different usability tradeoffs. Multivariate analysis of usability data can show significant differences among the products, but will probably not provide an interpretable measure of usability. Converting raw data to ranks, however, allows the establishment of rank-weighting schemes that combine different dependent measures into a single composite measure, allowing easy comparison of relative product usability.

Ranking addresses the question, "Which is the best?" Often, a specific numerical value is not interpretable in this way.

> For example, in professional baseball, . . . relative -- rather than absolute performance is what counts. A baseball team's share of post-season monies does not depend directly on the percentage of games the team won. Rather, it depends on how the team performed relative to the other teams in the same division. . . . If we are told that the Phillies finished first in the National League East, we know that they won more games than any of the other five teams in that division; we would not know this if we had been told, instead, that the Phillies won 75 percent of their games. On the other hand, of course, knowing that a team finished first does not tell us what percent of games it won. (Hildebrand, Laing, and Rosenthal, 1977)

In this sense, competitive usability analysis is like a baseball league. The essential task is to identify the most usable product after considering all of the available data. Also, because the data gathered in usability studies are usually human performance or attitude data, little may be known about the distributions from which the scores come, making conversion to ranks desirable. Finally, converting raw scores to ranks makes it possible to combine different types of data without using multivariate procedures

based on least-squares minimization to create centroids, which are often difficult to interpret.

DETAILED CONCEPT DESCRIPTION

Assume that a group of products has been studied under similar conditions using similar participants performing the same scenarios while observers record the same dependent variables. The measurements can be collapsed across participants to create a set of matrices, one matrix per dependent variable, with products for columns and scenarios for rows. The raw data in each matrix can be converted to ranks in a way similar to the Friedman rank-sum test (Bradley, 1976; Friedman, 1937). If raw values are tied, the appropriate ranks can be split (as described in Mendenhall, 1971). At this point, if desired and the matrices are complete (no missing data), Friedman tests can be run for each dependent variable matrix. Multiple comparisons tests based on Friedman ranks can also be used (Hollander and Wolfe, 1973). The next step is to collapse the data again over dependent variables, obtaining a matrix of composite rank scores arranged by product and scenario. This matrix can also be analyzed using Friedman tests. Finally, this matrix can be collapsed across scenarios (tasks), resulting in a single composite rank score representing the relative usability of products.

If, prior to data collection, decisions have been made regarding the relative importance of scenarios and dependent measures, this importance can be quantified by assigning weights to scenarios and measures. When the matrices are collapsed, the appropriate weights can be applied to emphasize the more important scenarios and measures. For example, if successful scenario completion rate is deemed twice as important as task time, then the ranked scenario completion rates would be multiplied by twice the weight applied against the ranked task times.

Sometimes, due to the variety in their functional capabilities, not all tasks can be performed on all products. This results in matrices with missing data, and a potential for a number of biases. These biases can be corrected by assigning ranks to complete the matrices, according to the following rules:

- If a task cannot be performed with a product because the product has been designed in an innovative way that relieves the user of the burden of performing the task, then that product should receive the lowest (best) rank. For example, if the task is to change the paper-handling device on a printer, but the printer has been designed to accept different types of paper automatically, then it should be credited for this innovation.
- If a task cannot be performed with a product because the product does not have the functional capability to perform the task, then that product should receive the highest (worst) rank. For example, if the task is to change paper-handling devices, but the printer can only work with pinfeed paper, then it should be penalized for this lack of function.
- If a task cannot be performed with a product, but it is not clear whether to credit or to penalize the product, then that product should be assigned the average rank across products for that task. For example, if the task is to change paper-handling devices, but the experimenter was unable to acquire the necessary equipment to study this task for a particular printer, then the printer should be neither penalized nor credited for the missing data. The average value is determined by adding one to the number of products and dividing this total by two.
- If a task cannot be performed with more than one product for the same reason, then the appropriate ranks should be split among those products.

The method that has been described is a mixture of the subjective and the objective. Several subjective decisions must be made before the objective part of the procedure can be applied, such as:

- Do the tasks or dependent variables differ in importance? If so, how should they be weighted?
- For each task with missing data, is it more reasonable to assess a penalty for a product's failure to perform a task, to credit a product for an innovative function, or simply to remove the product from comparison? These situations may exist simultaneously in a single set of usability data.

HYPOTHETICAL EXAMPLE 1

For this example, a number of assumptions have been made that simplify the situation.

- All tasks can be performed on all products.
- All tasks are considered to be equally important.
- All dependent measurements are considered to be equally important.

Table 1 shows the results of a hypothetical usability study conducted with three products. Assume a between-subjects design was used, with ten participants per product performing five tasks. Assume that the values in Table 1 were averaged across participants. For task times, a faster time is better. For errors, a lower number is better. For the satisfaction ratings, a lower number is better.

Table 1. Results for the First Hypothetical Example

Dependent Measure: Task Time (in minutes)

	Product A		Product B		Product C	
Task	Raw	Rank	Raw	Rank	Raw	Rank
1	8.0	3	7.2	1	7.3	2
2	7.6	2	5.3	1	8.5	3
3	6.3	2	5.2	1	7.5	3
4	4.2	3	4.0	2	3.0	1
5	6.1	2	4.2	1	6.5	3
Rank Ave:		2.4		1.2		2.4

Dependent Measure: Number of Errors

	Product A		Product B		Product C	
Task	Raw	Rank	Raw	Rank	Raw	Rank
1	0	1	3	3	0.5	2
2	1	2	0	1	2	3
3	1	2.5	0	1	1	2.5
4	1	1.5	1	1.5	2	3
5	0	2	0	2	0	2
Rank Ave:		1.8		1.7		2.5

Dependent Measure: User Satisfaction Rating

	Product A		Product B		Product C	
Task	Raw	Rank	Raw	Rank	Raw	Rank
1	3.5	3	2.0	1.5	2.0	1.5
2	2.2	2	2.0	1	3.0	3
3	1.3	1.5	1.3	1.5	2.0	3
4	1.0	1.5	1.0	1.5	1.7	3
5	1.7	2	1.5	1	1.9	3
Rank Ave:		2.0		1.3		2.7
Grand Rank Average:		2.1		1.4		2.5

Because these matrices are complete, Friedman tests can be conducted on the ranks. Given the small samples in this example, I have used $\alpha = .10$ as the criterion for significance. The Task Time matrix shows a significant difference ($X^2(2) = 4.8$, $p < .10$), the Number of Errors matrix is not significant ($X^2(2) = 1.9$, $p = .39$), and the User Satisfaction Rating matrix shows a significant difference ($X^2(2) = 4.9$, $p < .10$).

Table 2 shows the results of collapsing the matrices along the dimension of task and re-ranking the data. (It is possible to collapse along the dimension of dependent variable, but the resulting matrix would have order of 3 x 3, providing a smaller set of data and, therefore, a less powerful analysis.) A Friedman test conducted on the matrix given in Table 2 is significant ($X^2(2) = 6.7$, $p < .05$).

Table 2. Matrices Collapsed Across Tasks

	Product A		Product B		Product C	
Task	Rank Ave.	New Rank	Rank Ave.	New Rank	Rank Ave.	New Rank
1	2.3	3	1.8	1.5	1.8	1.5
2	2.0	2	1.0	1	3.0	3
3	2.0	2	1.2	1	2.8	3
4	2.0	2	1.7	1	2.3	3
5	2.0	2	1.3	1	2.7	3
New Rank Averages		2.2		1.1		2.7

The new rank averages from Table 2 are the composite rank scores that represent the relative usability of the products. The conclusion drawn from these results depends on which product is under development. If the product under development is Product B, then multiple comparisons can be conducted to help the developers determine the extent to which the product is significantly more usable than its competitors (see Hypothetical Example 2). If the product under development is Product C, then the developers can examine the data from Table 1 to prioritize their efforts to redesign the product.

HYPOTHETICAL EXAMPLE 2

A number of simplifying assumptions were made for Hypothetical Example 1, but an effective procedure needs to be able to deal with the following real-world problems:

- Not all tasks can be performed on all products.
- Different tasks may not be considered equally important.
- Different dependent measurements may not be considered equally important.

The problems of differentially important tasks and measures can be handled by creating weighting vectors ω_t for tasks and ω_m for measures. Following convention, the elements of the weighting vectors should sum to one. The way in which the weights are

established is not important for purposes of this discussion. The weights can be arrived at by consensus among the parties interested in the outcome of the analysis, or by more sophisticated methods such as the Analytical Hierarchy Process (Saaty, 1988). When the matrices are collapsed, the weighting vectors are used to produce the appropriate weighted average. For this example, assume that user rating is twice as important as number of errors, and the number of errors is twice as important as task time. Thus, ω_m = (.14, .29, .57) for time, errors, and rating respectively. Also assume that Tasks 2 and 3 are three times as important as Tasks 1, 4 and 5. Thus, ω_t = (.113, .33,.33, .113, .113) for the five tasks.

Table 3 shows the results of a hypothetical usability study conducted with three products. Assume a between-subjects design was used, with ten participants per product performing five tasks. Assume that the values in Table 3 were averaged across participants. For task times, a faster time is better. For errors, a lower number is better. For the satisfaction ratings, a lower number is better. A "-" indicates that the task could not be done with the product. Assume that data for Product A, Task 5 could not be collected due to a clerical error in ordering equipment, so it receives the average rank value (across the row) of 2. Assume that Product B has been designed to do Task 2 automatically, and receives the lowest (best) rank of 1. Finally, assume that Product C cannot do Task 2 at all, and is assigned the worst (highest) rank of 3. The resulting rank matrices are identical to those in Example 1, although the weighted averages are slightly different from the unweighted averages.

As was done in Hypothetical Example 1, Friedman tests can be applied against these matrices, but will provide only approximate results because missing data points have been estimated. Also, the Friedman procedure assumes equal weighting of the elements in the rows of the matrices (tasks in this case).

The effect of applying weights is to expand or compress the distance between the rank averages relative to the unweighted averages. An approximate statistical procedure that takes into account the effect of applying weights is to use multiple comparisons based on Friedman rank averages (Hollander and Wolfe, 1973), but to use the weighted averages rather than the unweighted averages. Table 4 shows the result of multiplying the matrices in Table 3 by ω_m. The matrix in Table 4 was re-ranked and multiplied by ω_t to get the weighted rank averages to analyze statistically. The results are shown in Table 5.

Table 3. Results for the Second Hypothetical Example

Dependent Measure: Task Time (in minutes)

	Product A		Product B		Product C	
Task	Raw	Rank	Raw	Rank	Raw	Rank
1	8.0	3	7.2	1	7.3	2
2	7.6	2	-	1	-	3
3	6.3	2	5.2	1	7.5	3
4	4.2	3	4.0	2	3.0	1
5	-	2	4.2	1	6.5	3
Rank Ave: (Weighted)		2.2		1.1		2.7

Dependent Measure: Number of Errors

	Product A		Product B		Product C	
Task	Raw	Rank	Raw	Rank	Raw	Rank
1	0	1	3	3	0.5	2
2	1	2	-	1	-	3
3	1	2.5	0	1	1	2.5
4	1	1.5	1	1.5	2	3
5	-	2	0	2	0	2
Rank Ave: (Weighted)		2.0		1.4		2.6

Dependent Measure: User Satisfaction Rating

	Product A		Product B		Product C	
Task	Raw	Rank	Raw	Rank	Raw	Rank
1	3.5	3	2.0	1.5	2.0	1.5
2	2.2	2	-	1	-	3
3	1.3	1.5	1.3	1.5	2.0	3
4	1.0	1.5	1.0	1.5	1.7	3
5	-	2	1.5	1	1.9	3
Rank Ave: (Weighted)		1.9		1.3		2.8
Grand Ave: (Weighted)		2.0		1.3		2.7

Table 4. Weighted Rank Averages

Task	Product A Rank Ave.	Product A New Rank	Product B Rank Ave.	Product B New Rank	Product C Rank Ave.	Product C New Rank
1	2.42	3	1.86	2	1.71	1
2	2.00	2	1.00	1	3.00	3
3	1.86	2	1.28	1	2.85	3
4	1.71	2	1.57	1	2.72	3
5	2.00	2	1.29	1	2.71	3
New Rank Averages: (Weighted)		2.1		1.1		2.8

Table 5. Results of Multiple Comparisons Tests

B (1.1) A (2.1) C (2.8)

Note: Products connected by a line are not significantly different at the .10 level. The products' final weighted rank averages are given in parentheses.

The conclusion drawn from these results would be the same as that for the first hypothetical example. If the product under development is Product B, then the developers can determine that the product is usable relative to its competitors. To further improve Product B relative to Product A, Product B should be redesigned to reduce the number of errors that occur when users perform Task 1. If the product under development is Product C, then the developers can examine the data from Table 3 to prioritize their efforts to redesign the product.

DISCUSSION

After usability data have been collected on a product, the method described in this paper can be used to determine how the product compares to its competitors, assuming an appropriate data base exists. The goal for a product under development is to receive the best composite rank score, corrected for various biases if necessary. At any time, the data base can be examined to determine how a product under development has received its highest (worst) ranks. A product developer can use this information to determine where an effort to modify the product would be best made. This is important because summary measures for multidimensional concepts (such as these composite usability rank-averages) always lose information relative to the entire data base from which they are calculated. After modifications have been made, the product could be tested again and, if necessary, iteratively modified and tested until the goal has been achieved. Application of the method allows a single composite rank-average to be assigned to a product to represent its relative usability, allowing easy comparison of products. Rank-based statistical procedures can be applied at various stages of analysis to determine if the products are significantly different.

ACKNOWLEDGEMENT

I want to express my appreciation to the anonymous reviewers who took the time to read and comment on my proposal, and to Alan J. Happ, for reviewing the final paper. Thank you.

REFERENCES

Bradley, J. V. (1976). *Probability; decisions; statistics.* Englewood Cliffs, NJ: Prentice-Hall.

Cliff, N. (1987). *Analyzing multivariate data.* San Diego, CA: Harcourt-Brace-Jovanovich.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32,* 675-701.

Gould, J. (1988). How to design usable systems. In M. Helander (Ed.), *Handbook of Human-Computer Interaction.* New York: North-Holland.

Hildebrand, D. K., Laing, J. D., and Rosenthal, H. (1977). *Analysis of ordinal data.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-008. Beverly Hills, CA: Sage.

Hollander, M. and Wolfe, D. A. (1973). *Nonparametric statistical methods.* New York, NY: John Wiley.

Lewis, J. R., Henry, S. C., and Mack, R. L. (1990). Integrated office software benchmarks: A case study. In D. Diaper et al. (Eds.), *Human-Computer Interaction – INTERACT '90* (337-343). New York: North-Holland.

Mendenhall, W. (1971). *Introduction to probability and statistics.* Belmont, CA: Duxberry Press.

Saaty, T. (1988). *Decision making for leaders: The analytical hierarchy process for decisions in a complex world.* Pittsburgh, PA: University of Pittsburgh.

Whiteside, J.; Bennett, J.; and Holtzblatt, K. (1988). Usability engineering: our experience and evolution. In M. Helander (Ed.), *Handbook of Human-Computer Interaction.* New York: North-Holland.