A Generalized Thurstonian Paired Comparison Multicriteria Heuristic Model for Peer Evaluation of Individual Performance on IS Team Projects

Julian M. Scher
Scher@adm.njit.edu
Department of Information Systems
College of Computing Sciences
New Jersey Institute of Technology
Newark, New Jersey 07102 USA

Abstract

Information Systems instructors are generally encouraged to introduce team projects into their pedagogy, with a consequential issue of objectively evaluating the performance of each individual team member. The concept of "freeloading" is well-known for team projects, and for this, and other reasons, a peer review process of team members, by team members, is often advocated. We propose an objective heuristic model for obtaining a scale of individual performance, based upon a generalization of Thurstone's Law of Comparative Judgment, where pair wise comparisons of team member's performance are elicited with regard to various criteria, and we demonstrate how a scale may be obtained to objectively rate the individual members of each team. A numerical example is provided to illustrate our Generalized Thurstone model's heuristic methodologies.

Keywords: Thurstone, paired comparison, team projects, peer evaluation, multi-criteria, individual performance

1. INTRODUCTION

Team projects should be an inherent goal in the pedagogy of the IS instructor. The justification for incorporating team projects has been succinctly stated by (Steenkamp, 2002):

"The rationale is that once students enter the work environment they will be required to work in teams. Working in a team context challenges team members in a number of ways, such as:

- Teams are composed of individuals with different technical skills, cultural backgrounds, behavioral characteristics, cognitive styles and learning abilities.
- Performance of team members is influenced by the level of teamwork and in-field experience, knowledge of the application domain, pressures of schedule, geographical dispersion, full-time or part-time study."

The ABET-CAC accreditation criteria clearly specify that an accredited IS program must, as part of its objectives, outcomes and assessment, enable all its graduating IS majors to achieve, by the time of their graduation "an ability to function effectively on teams to accomplish a common goal" (ABET, 2008). The issue of proper evaluation of individual effort in a team project is a concurrent dilemma for the IS instructor, who, for instance, needs to be cognizant of the often-encountered "freeloading" which occurs in team projects (Fox, 2002). Being able to distinguish, identify and measure individual contributions and excellence on a team effort is vital to the success of the pedagogical exercise. The traditional practice of having the instructor

review the results of the team project, and then award the identical grade to all members of the team, is problematic, and, as pointed out in (Tu, Lu, 2004), actually encourages and provides incentives for some of the weaker students to "freeload." While this paper only addresses a peer evaluation methodology to measure relative individual performance in a group project, much has been written in terms of the pedagogical issues relating to optimizing the group project experience for students, and the reader is encouraged to peruse the research recommendations by Tu and Tu(2004b).

One strategy for evaluation of individual performance on project and client teams is that of peer evaluation, where individual students anonymously rate each other (Lewis, 2006). Several methodologies have been suggested for the measurement of individual student performance on teams (Ruble, Hernandez and Amadio, 2004), though there is no universal agreed-upon standard. (Tu, Lu, 2004) discuss the issue of truthfulness in peer evaluation rankings of team members, and offer a methodology in peer evaluation so that truth-telling becomes the dominant strategy of individual team members. (Kelley, Sadowski, 2005) found that teams using a peer evaluation instrument in an engineering design graphics course team project functioned better than teams not using a peer evaluation instrument. (Lewis, 2006) suggests that the peer evaluations be done on a weekly basis. On the other hand, there are researchers who have raised doubts as to the usefulness of peer assessments in group projects, such as (Kennedy, 2005), who questions the underlying value of peer assessments, based on his experience with group projects in university computing courses.

In this paper, we shall present a model for a peer evaluation of individual team member performance, based upon some generalizations and extensions of the classic Thurstone's Law of Comparative Judgment.

The "law" Thurstone created is essentially a measurement model, which requires subjects to make a preference comparison between each of a number of pairs of stimuli with regard to the magnitude of a property, attribute, or attitude. (Thurstone, 1927, 1929, 1959).

2. A GENERALIZED THURSTONE MODEL

Let us assume we have n+1 students on a team, which the instructor has divided the class into project teams or client teams for a particular assignment, or set of assignments. As part of the evaluation process to measure individual student's performance on his/her team, each individual student will be asked to pairwise compare each of the other n students on his/her team according to a specific set of criteria chosen by the instructor, the number of such criteria to be denoted by m.

Then, since a given student will not be asked to vote in any pair wise comparison involving himself/herself, there will be n*(n-1)/2 paired comparisons between different students, for a specific criteria r.

Thurstone (1927) presented a conceptual model for paired comparisons based upon several assumptions:

1. When a stimuli pair is presented to a subject, it will elicit a continuous preference (referred to as a "discriminal process") for each stimulus.

2. The one stimulus whose value is greater at the moment of the comparison will be the one that is preferred by the subject.

3. The aforementioned preferences are normally distributed in the population.

4. It is assumed that each individual will respond to all of the possible paired comparisons.

Classic Thurstone analysis requires that after the individual students do their pairwise comparisons of other students for each of the m criteria, the results be presented in terms of a frequency matrix F, where f(i,j) denotes the count of those who prefer alternative i to alternative j, where in our case i and j are individual students.

We will extend the original Thurstone model of paired comparisons by assuming that the students will be evaluating each other based upon a set of criteria established by the instructor. The establishment of evaluation criteria for class projects by the instructor is fairly typical in universities – for instance,

Chandra (http://tinyurl.com/nuw6ns) utilizes the following 5 criteria in evaluating projects:

Depth and breadth of research

- Subject Knowledge
- Project Presentation Quality
- Final Project Report
- Original or New Contributions

We will denote by F^r the frequency matrix for criteria r, where r = 1,…,m. Thus, if we have m different criteria which we wish to use to evaluate members of the team, we will then have a total of m frequency matrices.

The individual elements $f^r(i, j)$ of the frequency matrix F^r will be the count of those who "prefer" student i over student j according to the r'th criteria.

Quantifying this, we define $x_{i,j}(k)$, the "rating" by student k in terms of his/her preference of student i to student j, for each of the r criteria, where k <> (i,j) and k = 1,…,n+1:

$$x_{i,j}(k) = \begin{cases} 1 \text{ if i is preferred to j by student k} \\ 0 \text{ if j is preferred to i by student k} \end{cases}$$

where we have the constraints that

$x_{i,j}(k) + x_{j,i}(k) = 1$ for i <> j and

$x_{i,i}(k) = 0$ for k = 1,…n+1 .

We also seek transitivity in student k's rating, i.e., if $x_{a,b}(k) = 1$ and $x_{b,c}(k) = 1$, then $x_{a,c}(k) = 1$. Also, we insist that each student be required to make every comparison, without having any 'indifferent' votes.

The elements $f^r(i, j)$ of the frequency matrix F^r are then computed as follows:

$$f^r(i, j) = \sum_{k=1}^{n+1} x_{i,j}(k) \quad \text{for i,j = 1,…, n+1}$$

To illustrate these preliminary concepts with some numerical data, suppose that we have a team of n+1 or 5 students, who we will denote by S1. S2. S3. S4 and S5.

Let us assume that we desire the frequency matrix F^r for a particular criteria (for instance, r=1), and have queried the students to obtain the following pairwise comparisons :

Let student 1's ratings be as follows:

$x_{2,3}(1) = 1,\ x_{2,4}(1) = 1,\ x_{2,5}(1) = 0$

$x_{3,4}(1) = 1,\ x_{3,5}(1) = 0,$

$x_{4,5}(1) = 0$

Let student 2's ratings be as follows:

$x_{1,3}(2) = 0,\ x_{1,4}(2) = 0,\ x_{1,5}(2) = 0$

$x_{3,4}(2) = 1,\ x_{3,5}(2) = 1$

$x_{4,5}(2) = 1$

Let student 3's ratings be as follows:

$x_{1,2}(3) = 0,\ x_{1,4}(3) = 0,\ x_{1,5}(3) = 0$

$x_{2,4}(3) = 0,\ x_{2,5}(3) = 1$

$x_{4,5}(3) = 0$

Let student 4's ratings be as follows:

$x_{1,2}(4) = 0,\ x_{1,3}(4) = 0,\ x_{1,5}(4) = 0$

$x_{2,3}(4) = 1,\ x_{2,5}(4) = 1$

$x_{3,5}(4) = 1$

Let student 5's ratings be as follows:

$x_{1,2}(5) = 0,\ x_{1,3}(5) = 0,\ x_{1,4}(5) = 0$

$x_{2,3}(5) = 1,\ x_{2,4}(5) = 0$

$x_{3,4}(5) = 0$

We then generate the frequency matrix F^r as presented in Figure 1 of the Appendix, noting that we obtain the remaining

elements in the frequency matrix F^r by using the fact that $x_{i,j}(k) + x_{j,i}(k) = 1$.

The elements in this frequency matrix are computed by the relationship between $f^r(i, j)$ and $x_{i,j}(k)$, namely

$$f^r(i, j) = \sum_{k=1}^{n+1} x_{i,j}(k) \quad \text{for } i,j = 1,\ldots, n+1$$

Thus, for instance,

$$f^r(3,4) = x_{3,4}(1) + x_{3,4}(2) + x_{3,4}(5)$$

$$= 1 + 1 + 0 = 2$$

Also, we need to satisfy f(i,j)+f(j,i) = n-1, that is, the number of pair-wise comparisons done by the group for any two students will be n-1.

With a team of n+1 students, the number of paired comparisons we ask of each student, for each criteria, as previously stated, is n*(n-1)/2. For our illustrative example, with n+1 or 5-student teams, this involves 4*3/2 or 6 paired comparisons of fellow student teammates for each of the m criteria. The total number of paired comparisons for each of the k students will therefore be m*n*(n-1)/2. Typical student team sizes, such as client teams, are often between 4 and 5, and if the instructor seeks to keep the number of different criteria to a small number, such as 3 or 4, then the total number of paired comparisons required of each student in the peer evaluation will be 24 or less.

The second phase of the Thurstone model will be the transformation of the frequency matrices F^r into Probability matrices P^r, where P^r denotes the Probability matrix for the r'th criteria, where r = 1,….m.

If there are n+1 students, each individual student being asked to make paired comparisons involving each pair of the other n students, then there will be n*(n-1)/2 paired comparisons, and the number of students making a paired comparison between i and j, i.e., f(i,j), will be (n+1)-2, or (n-1) since we omit the two specific students i and j who do not make paired comparisons involving themselves. The elements of the probability matrix, denoted by $p^r(i,j)$, are then computed as follows:

$$p^r(i,j) = f^r(i, j) / (n-1)$$

We also compute, for each row k in P^r, the sum of the probabilities in row k, which we denote by V_k (k = 1,…,n+1)

The resulting Probability matrix P^r is given in Figure 2 of the Appendix.

Following the computation of the Probability matrix, a new matrix is then computed, traditionally called X in the psychometric literature, but for our nomenclature we will refer to it as the Z matrix. The cell values of matrix Z are the standardized normal deviates corresponding to the probabilities given in matrix P^r. Thurstone's Law of Comparative Judgment prescribes that the scale value difference between any two stimuli in a paired comparison assessment is a random variable following a Normal (Gaussian) probability density function. The mean value of this Normal distribution represents the scale value difference between the two stimuli in question.

We next transform the Probability Matrix P^r into the standardized Normal Matrix Z^r, where the Z(i,j) values are computed from the N(0,1) tables. We will approximate the presumed underlying theoretical Gaussian distribution with a doubly truncated standardized normal distribution having truncation endpoints at -3 and +3, corresponding to CDF(0) and CDF(1), where CDF(x) represents the cumulative distribution function at point x. While one may obtain precise values for the doubly truncated normal distribution (Johnson and Thermopolous, 2002), there will be no harm in approximating these values by the more widely accessible standardized Normal tables. Invoking the standardized normal tables will yield the Z^r matrix of Figure 3 in the Appendix.

The last column, $T_k(r)$, of Figure 3 represents the sum of the k'th row's standardized normal values for the r'th criteria, and so, for each of the m criteria, we have a vector T with k components.

For each of the m criteria, the instructor will assign a weight given by w_j, where j = 1,…,m and the w_j are non-negative, and

$\sum_j w_j = 1$ (i.e, the weights constitute a convex combination).

Once we compute the $T_k(r)$ values for all criteria r (r = 1,…,m) for each of the k students (k = 1,…,n+1), we may then compute the scale values A_k as follows:

$$A_k = \sum_{J=1}^{m} w_j T_k(r) \qquad \text{for } k = 1,\ldots,n+1$$

To illustrate the computations of the scale values A_k , let us assume that we have 3 criteria (i.e., m =3) assigned by the instructor:

Criteria(1) = overall quality of work contributed

Criteria(2) = availability and willingness to work with other team members and support the work of the team

Criteria(3) = perceived amount of effort

The instructor believes that criteria(1), the overall quality of work contributed, is twice as important as either of the other two criteria, and that criteria(2) and criteria(3) are equal in importance, which leads us to the following weights:

$W_1 = .5$ $W_2 = .25$ $w_3 = .25$

For simplicity of presentation, we will assume that the previously computed Z matrix was the one generated for criteria(1), and so

$T_1(1) = -12.0$

$T_2(1) = 6.0$

$T_3(1) = 0.861$

$T_4(1) = 2.5692$

$T_5(1) = 2.5692$

We will provide data values for $T_k(2)$ and $T_k(3)$, (for k=1,…5) and not bother the reader with the background details/computations for the associated Z, P and F matrices.

So, let us assume we have $T_1(2) = -5.4$, $T_2(2) = 4.31$, $T_3(2) = -1.8$, $T_4(2) = 1.69$ and $T_5(2) = 1.2$ for criteria(2)'s $T_k(r)$ values.

For criteria(3)'s values, we have $T_1(3) = -2.7$, $T_2(3) = 1.2$, $T_3(3) = -.9$, $T_4(3) = 1.0$ and $T_5(3) = 1.4$.

The "T" matrix will then be:

-12.0	6.000	.8616	2.5692	2.5692
-5.4	4.31	-1.8	1.69	1.2
-2.7	1.2	-0.9	1.0	1.4

and

$A_1 = .5*(-12) + .25*(-5.4) + .25*(-2.7)$

$A_1 = -8.025$

$A_2 = .5*(6.0) + .25*(4.31) +.25*(1.2)$

$A_2 = 4.3775$

$A_3 = .5*(.8616) +.25*(-1.8) +.25*(-.9)$

$A_3 = -.2442$

$A_4 =.5*(2.5692) +.25*(1.69) +.25*(1)$

$A_4 = 1.9571$

$A_5 =.5*(2.5692) +.25*(1.2) +.25*(1.4)$

$A_5 = 1.9346$

Plotting these points on a scale, we obtain Figure 4 in the Appendix.

Clearly, Student #1 is the *least* preferred

student on the team, as judged by the peer evaluation of the team, and by a significant degree. Student #3 is second in the least preferred category, as evaluated by his/her peers. Student #2's performance was recognized as being the best on the team. After Student #2, we have Student #4 and Student #5 coming relatively close in the peer evaluation, with Student #4 barely edging out Student #5.

The objective measures we have thus obtained will guide and support the instructor in the difficult task of assigning grades to individual members of this team of five students. While there will be some subjectivity on the part of the instructor in his/her interpretation of these results, our inclination would be to reward Student 2 with the highest grade, an "A," on his/her individual performance on the team. Student #1 would be awarded the lowest grade, an "F," for a performance that was clearly recognized as deficient by his/her teammates. Since student #4 and student #5 were viewed positively and had a near-equivalent performance, they each should be awarded identical grades, very likely a grade of "B." For student #3, whose individual performance was viewed as slightly negative in the Thurstonian comparative evaluation, we would be generous and award him/her with a ("gentlemen's") "C."

3. CONCLUSIONS

We have presented a generalized heuristic Thurstonian model for use in peer evaluation of individual performance on team projects. It is based on paired comparisons of individual students, whereby each student will compare pairs of other students according to instructor-selected criteria, and the instructor will also select the relative importance of each criteria. Classical Thurstonian concepts are utilized and extended to produce an apropos scale where instructors may review the relative performance of individual students on teams, and observe the resultant scales.

4.ACKNOWLEDGEMENT

We acknowledge a colleague, Distinguished Professor Emeritus Murray Turoff, who well over 35 years ago brought Thurstone's Law of Comparative Judgment to our attention, and which lied dormant with us for many years, until the need arose for a methodology to assess the individual contributions of students on team projects.

5. REFERENCES

ABET/Computing Accreditation Commission, (2008), "CRITERIA FOR ACCREDITING COMPUTING PROGRAMS," p. 5. (http://tinyurl.com/l4zu5x)

Fox, Terry, (2002), "A Case Analysis of Real-World Systems Development Experiences of CIS Students," Journal of Information Systems Education, Vol. 13(4), p. 345.

Johnson, Arvid and Nick Thomopolous, (2002), "Characteristics and Tables of the Doubly Truncated Normal Distribution," Proceedings of POM High Tech, Production and Operations Management Society, 2002.

Kelley, David and Mary Sadowski, (2005), "Peer Evaluation Within a Team Design Project," Proceedings of the Annual Meeting of the American Society for Engineering Education (EDGD), Ft, Lauderdale, FL.

Kennedy, Geoffrey (2005), "Peer-Assessment in Group Projects: Is It Worth It?," Australasian Computing Education Conference 2005, Newcastle, Australia. Proceedings of the Conferences in Research and Practice in Information Technology, Vol. 42. Alison Young and Denise Tolhurst, Eds.

Lewis, K., (2006), "Evaluation of Online Group Activities: Intra-Group Member Peer Evaluation," Proceedings of the 22nd Annual Conference on Distance Learning and Teaching, University of Wisconsin-Madison.

Ruble, Thomas L. and S. Hernandez and W. Amadio (2004), "A Comparison of Peer Evaluation Systems in Team-Based Learning," Proceedings of the Academy of Business Education Annual Conference

Steenkamp, Annette Lerine (2002), "A Standards-Based Approach To Team-Based Student Projects In An Information Technology Curriculum," Proceedings of the 17th Annual Conference of the International Academy for Information Management, pp. 54-62.

Thurstone, Louis Leon (1927), "A Law of Comparative Judgment," Psychological Review, 34, 278-286.

Thurstone, Louis Leon (1929), "The Measurement of Psychological Value," in T. Smith and W.K. Wright (Eds.), "Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago." Chicago: Open Court.

Thurstone, Louis Leon. (1959), "The Measurement of Values." Chicago: The University of Chicago Press.

Tu, Yanbin and Min Lu (2004), "Mechanism Design for Peer and Self Assessment of a Group Project," The Proceedings of the Information Systems Education Conference 2004, v 21 (Newport): §3262. ISSN: 1542-7382.

Tu, Yanbin and Yanlin Tu (2004), "Achieving an Effective and Successful IS Group Project." The Proceedings of ISECON 2004, v 21 (Newport): §3263. ISSN: 1542-7382

APPENDIX

Figure 1: The Frequency Matrix F^r

	S1	S2	S3	S4	S5
S1	-	0	0	0	0
S2	3	-	3	1	2
S3	3	0	-	2	2
S4	3	2	1	-	1
S5	3	1	1	2	-

Figure 2: The Probability Matrix P^r

	S1	S2	S3	S4	S5	v_k
S1	-	0	0	0	0	0
S2	1.0	-	1.0	.3333	.6667	3.0
S3	1.0	0	-	.6667	.6667	2.333
S4	1.0	.6667	.3333	-	.3333	2.333
S5	1.0	.3333	.3333	.6667	-	2.333

Figure 3: The Standardized Normal Matrix Z^r (for r=1)

	S1	S2	S3	S4	S5	v_k
S1	-	-3	-3	-3	-3	-12.0
S2	3.0	-	3	-.4308	.4308	6.0
S3	3.0	-3	-	.4308	.4308	0.8616
S4	3.0	.4308	-.4308	-	-.4308	2.5692
S5	3.0	-.4308	-.4308	.4308	-	2.5692

Figure 4: The 5-Student Peer Evaluation Thurstonian Scale of Individual Performance On A Team