

A Semidefinite Programming Formulation to Quantify Rater Ability When Performing Peer Evaluations

P. K. Imbrie, Alexander Malyscheff, and Junqiu Wang

School of Engineering Education, Purdue University, West Lafayette, Indiana, USA

Phone: (765) 496-7225

Fax: (765) 494-5819

imbrie@purdue.edu, amalysch@purdue.edu, and jqwang@purdue.edu

Context:

Over the past ten years, education in engineering has seen a significant increase in the emphasis on design and on the wide range of teamwork skills that engineering students will need when they enter the workplace (ASEE, 1994, 1995; Dahir, 1993; Hissey, 2000; Valenti, 1996). In the program outcomes at the heart of Engineering Criteria 2000 accreditation guidelines, students are mandated to be able to function on multidisciplinary teams in addition to acquiring traditional engineering knowledge of mathematics, science, and engineering and gaining experience in engineering problem solving and system design (ABET, 2000, 2002). Their ability to contribute to a multidisciplinary team will hinge on developing skills to communicate effectively and to understand a wide range of issues, including professional and ethical responsibility, the impact of engineering solutions in a global and societal context, and knowledge of contemporary issues. As our notion of an engineer's core competencies broadens to include these skills, we may find that a teaming emphasis in course learning provides additional leverage to achieve a gender- and ethnically diverse population of engineering professionals.

The use of peer assessments to monitor individual and team progress in engineering education is a widespread practice. However, rater variability is a factor that directly limits the inferences of rating scale scores (Engelhard, 1992, 1994, 1996; Sherrard, Raafat, & Weaver, 1994). According to Saal, Downey, and Lahey (1980), there are four types of rater variability: severity or leniency, halo effect, central tendency, and restriction of range. While empirical evidence suggests training raters to grade compositions promotes rater consistency (Weigle, 1994), limited research has addressed whether training engineering students to assess other students improves their rating judgments. Specifically, it is not known whether training engineering students to rate team members results in comparable ratings with other raters for a target student? As such, there is much to gain through research leveled at the use of peer assessments in engineering classrooms.

Therefore, this paper examines rater variability, or bias, in data based on a 9-item Likert type self-assessment peer-evaluation instrument measuring three constructs of effective teaming along (1) interdependence, (2) goal-setting, and (3) potency (Imbrie, Maller, and Immekus, 2005) A variety of approaches exist to address peer evaluation bias in the form of rater variability (Saal, Downey, & Lahey, 1980). Providing constructive feedback in the form of a self-report peer evaluation represents one possible method to improve rating accuracy (Smith, 1986). We focus on providing constructive feedback by quantifying observed discrepancies between student scores and a "true" reference score provided by a trained expert. Vignettes are introduced as part of the peer evaluation process in the form of an additional hypothetical student team member requesting actual student team members to evaluate each other and the hypothetical team member. Since the vignette has also been evaluated by an unbiased reference, we are able to relate the student evaluations to an unbiased reference score. A long-term objective in this

context addresses the possibility to provide faculty a means to adjust biased student peer evaluation data as well as help students understand what constitutes effective teaming.

Research Questions:

The difficulty in quantifying rating ability centers on the question:

How can one model a variable that could roughly be described by the notion of “general agreement” between student scores and reference score? Does this variable truly represent a bias in the students’ ability to evaluate his/her peers in teaming activities?

To begin answering this question we make the assumption that a student who evaluates along the same “direction” displayed in the reference score (the score provided by an unbiased trained expert who has evaluated the same hypothetical student team member vignette) is considered a much better rater than a student whose evaluation pattern generally appears to be inconsistent, when compared to the provided reference.

We approach this question by providing a metric accounting for direction and distance when comparing student scores to the reference score. Consider Student 1 whose evaluation vector differs only in a constant bias, as for example displayed in an over- or under- evaluation of the reference score by a constant value through all nine items. In this case we will consider Student 1 to be a more-reliable rater, as he/she manages to show a correct general tendency with respect to the provided reference, albeit his/her score represents an over or under bias of the actual score . In contrast, consider Student 2 who over-estimates item 1, under-estimates item 2, and continues in the same fashion through the remaining items. We label Student 2 a significantly less reliable rater than student 1. Geometrically, the angle between the evaluation vector of Student 1 and the reference vector is rather small, the two vectors are close to collinear. In contrast, the angle between the evaluation vector of Student 2 and the reference vector is very large, in extreme cases the two vectors are close to perpendicular. We approach this problem by requiring an outside expert to label a randomly selected set of difference vectors x_j by assigning functional values ranging from 0 to 1 reflecting the expert’s opinion regarding the rater’s evaluation ability. In doing so, we adjust the notion of distance and direction in the space containing the rating data providing more “weight” in one direction and less “weight” in another.

Theoretical Framework and Methodology:

In this paper peer evaluation scores, as provided by students on a hypothetical team member (vignette), are compared to a reference score. Student scores and reference score are collected for a 9-item Likert-type peer evaluation assessment instrument and subsequently modeled for l students as a vector in a 9-dimensional space, $s_j, r \in \mathfrak{R}^d$, $\forall j = 1, \dots, l$, $d = 9$. The student scores range from 0 to 100 for each item, thus, the rating data is confined to a d -dimensional hypercube in the nonnegative orthant, $s_{ji} \in [0, 100]$, $\forall i = 1, \dots, d$. We focus in particular on the l differences between the student scores and a reference score, $x_j = s_j - r$. Students whose difference vectors x_j are highly collinear with respect to the reference vector r are considered superior raters as opposed to students for whom x_j is almost perpendicular to r . It is our objective to calculate a functional approximation of rating ability based on a minimum of outside information, which is provided here in the form of an expert evaluating the difference between student scores and reference score for a small subset. In other words we require for a small number of students a label $y_j \in \mathfrak{R}$ indicating a student’s rating ability and obtain based on that

information a function, which computes rating abilities for the remaining students. We choose to model student rating ability using a function of the type

$$f(x) = e^{-\frac{1}{2}x^T Ax} \quad (1)$$

with the matrix $A \in \mathfrak{R}^{d \times d}$ as unknown variable. By employing this type of function we ensure that $f(x_j) = 1$ for the perfect rater for whom we have $s_j = r$, or $x_j = 0$. We furthermore satisfy that for very bad raters, $\|x_j\| \rightarrow \infty$, we have $f(x_j) \rightarrow 0$ as long as the matrix A is positive definite. In order to determine the matrix $A \in \mathfrak{R}^{d \times d}$ we rewrite equation (1) introducing the function $\tilde{f}(x)$ defined as:

$$\tilde{f}(x) = -2 \ln(f(x)) = x^T Ax. \quad (2)$$

Similarly, for a subset of student scores k labels provided by an expert must be converted, therefore we have $\tilde{y}_j = -2 \ln(y_j)$. The problem of identifying matrix $A \in \mathfrak{R}^{d \times d}$ satisfying $\tilde{y}_j = x_j^T Ax_j$ for a set $T = \{(x_j, \tilde{y}_j)_{j=1}^k\} \subset \mathfrak{R}^d \times \mathfrak{R}$ can be formulated as a semidefinite programming problem:

$$\begin{aligned} \min \quad & \sum_{j=1}^k (x_j^T Ax_j - \tilde{y}_j)^2 \\ \text{subject to} \quad & \\ & A \succ 0 \end{aligned} \quad (3)$$

where $A \succ 0$ indicates that the matrix $A \in \mathfrak{R}^{d \times d}$ be positive definite, satisfying $z^T Az > 0$ for all $z \in \mathfrak{R}^d$, $z \neq 0$. We impose positive definiteness on A , in order to guarantee that $e^{-\frac{1}{2}x^T Ax} \rightarrow 0$ as $\|x\| \rightarrow \infty$. Introducing a set of scalars t_j for $j = 1, \dots, k$ equation (3) can be modified resulting in:

$$\begin{aligned} \min \quad & t_1 + \dots + t_k \\ \text{subject to} \quad & \\ & t_1 \geq (x_1^T Ax_1 - \tilde{y}_1)^2 \\ & \vdots \\ & t_k \geq (x_k^T Ax_k - \tilde{y}_k)^2 \\ & A \succ 0 \end{aligned} \quad (4)$$

We will next rewrite the constraints. Note that (4) essentially computes a matrix A to fit a set of ellipsoids $E_j = \{x_j : x_j^T Ax_j = \tilde{y}_j\}$ centered at the origin. The notion of distance in this context is to be interpreted with respect to the matrix A , that is, $\|x\|_A = \sqrt{x^T Ax}$. Keeping this in mind the set of ellipsoids can be expressed as $E_j = \{x_j : \|x_j\|_A = \sqrt{\tilde{y}_j}\}$ in other words we are linking a set of vectors

x_j with a set of prescribed but varying distances \tilde{y}_j in a space defined by the matrix A . Calafiore (2002) addresses a similar problem fitting a single ellipsoid to a set of vectors x_j , where each vector is assigned a constant distance $\tilde{y}_j = \text{const.}$. Following a similar approach we can rewrite the ellipsoidal constraints and define:

$$h_j(A) = \begin{pmatrix} x_j^T & 1 \end{pmatrix} \begin{bmatrix} A & 0 \\ 0 & -\tilde{y}_j \end{bmatrix} \begin{pmatrix} x_j \\ 1 \end{pmatrix} = x_j^T A x_j - \tilde{y}_j, \quad \forall j = 1, \dots, l. \quad (5)$$

Using Schur's complement as discussed by Vandenberghe & Boyd (1996) the expression $t_j - h_j^2(A) \geq 0$ is equivalent to:

$$\begin{bmatrix} 1 & h_j(A) \\ h_j(A) & t_j \end{bmatrix} \succ 0, \quad \forall j = 1, \dots, k. \quad (6)$$

Finally, let $t = (t_1 \ \dots \ t_k)^T$ and let $e \in \mathfrak{R}^k$ denote a vector of ones $e = (1 \ \dots \ 1)^T$, the optimization problem can be formulated as:

$$\begin{aligned} & \min \quad t^T e \\ & \text{subject to} \\ & \begin{bmatrix} 1 & h_j(A) \\ h_j(A) & t_j \end{bmatrix} \succ 0, \quad \forall j = 1, \dots, k \end{aligned} \quad (7)$$

$$h_j(A) = \begin{pmatrix} x_j^T & 1 \end{pmatrix} \begin{bmatrix} A & 0 \\ 0 & -\tilde{y}_j \end{bmatrix} \begin{pmatrix} x_j \\ 1 \end{pmatrix}, \quad \forall j = 1, \dots, k$$

$$A \succ 0.$$

Interior point methods (den Hertog, 1993) can be used efficiently to solve semidefinite programming problems (Vandenberghe & Boyd, 1996). The above semidefinite optimization problem can be implemented in Matlab (<http://www.mathworks.com>) using SeDuMi (Sturm, 1999), a Matlab toolbox, which solves semidefinite programming problems based on interior point methods, and Yalmip (Löfberg, 2004), an interface, which allows for convenient control and implementation of semidefinite programming problems.

Findings and Conclusions:

We have introduced an optimization formulation, which serves as a basis for calculating a positive definite matrix A and subsequently a function $f(x)$ quantifying student rating ability. Consider as a preliminary example the three vectors $x_1 = (2 \ 1 \ 1)^T$, $x_2 = (-1 \ 1 \ -1)^T$ and $x_3 = (5 \ 4 \ 5)^T$ representing differences in rating ability between students s_1 , s_2 , and s_3 with respect to the reference score r . That is, for the reference score $r = (3 \ 2 \ 3)^T$ and a student score

$s_1 = (5 \ 3 \ 4)^T$ the difference vector x_1 equals $x_1 = (2 \ 1 \ 1)^T$. For computational simplicity we consider here $d = 3$ instead of $d = 9$. Suppose that an expert has labeled the three deviations between student score and reference score with $y_1 = 0.9$, $y_2 = 0.5$, and $y_3 = 0.4$. Note that the expert decides that even though x_2 is rather close to the reference score (recall that for $x_j = 0$ we have $s_j = r$), x_2 is considered only marginally better than x_3 , a rater who is consistently overestimating the reference values and doing so by a significant margin ranging from 4 to 5. Converting the labels y_j to log-scale one finds $\tilde{y}_1 = -2\ln(y_1) = 0.21$, $\tilde{y}_2 = 1.39$, and $\tilde{y}_3 = 1.83$. Finally, solving problem (7) results in the matrix:

$$A = \begin{bmatrix} 0.3183 & -0.1308 & -0.2951 \\ -0.1308 & 0.4505 & -0.1888 \\ -0.2951 & -0.1888 & 0.5684 \end{bmatrix}$$

Indeed, we can verify that for x_1

$$\begin{aligned} f(x_1) &= \exp\left\{-\frac{1}{2}x_1^T Ax_1\right\} \\ &= \exp\left\{-\frac{1}{2}(2 \ 1 \ 1) \begin{bmatrix} 0.3183 & -0.1308 & -0.2951 \\ -0.1308 & 0.4505 & -0.1888 \\ -0.2951 & -0.1888 & 0.5684 \end{bmatrix} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\right\} = 0.9 = y_1 \end{aligned}$$

Analogous calculations can be demonstrated for (x_2, y_2) and (x_3, y_3) . We therefore have obtained a function $f(x)$, which introduces a metric based on the expert labeling that can now be applied to additional observed student ratings x_j .

Recommendations:

Student peer evaluations are highly subjective and can potentially result in a significant bias in the overall score for a team member. In this paper a differential score, derived from peer evaluation scores and a reference score using a hypothetical vignette, is being evaluated by an expert. We present a semidefinite programming formulation for calculating a function, which provides a means to quantifying student rating ability. Rating ability is measured in the range from 0 to 1 with a value of 1 representing the best score. This process introduces a metric, emphasizing certain directions over others when evaluating student rating with respect to a reference score. The function, derived in this process, can thereafter be employed to evaluate additional student scores providing constructive rating feedback to student team members.

Acknowledgments

This research has been supported by the National Science Foundation, NSF Grant DUE-0512776.

References

ABET (2000). Criteria for Accrediting Engineering Programs. The Engineering Accreditation Commission of The Accreditation Board for Engineering and Technology. <http://www.abet.org/eac/eac.htm>.

- ABET (2002). Engineering Criteria 2002-2003. Accreditation Board for Engineering and Technology, <http://www.abet.org/criteria.html>.
- ASEE Deans' Council and Corporate Roundtable (1994), Engineering Education for a Changing World, Washington, DC: American Society for Engineering Education, October.
- ASEE (1995). Educating Tomorrow's Engineers. *ASEE Prism*, 11-15, May/June.
- Calafiore, G. (2002). Approximation of n-Dimensional Data Using Spherical and Ellipsoidal Primitives, *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 32 (2), 269-278.
- Dahir, M. (1993). Educating Engineers for the Real World. *Technology Review*, 14-16, Aug/Sept.
- den Hertog, D. (1993). *Interior Point Approach to Linear, Quadratic and Convex Programming*, Kluwer.
- Engelhard, G. (1992). The Measurement of Writing Ability With a Many-Faceted Rasch Model, *Applied Measurement in Education*, Vol. 5 Issue 3, p171, 21p.
- Engelhard, G. (1994). Examining Rater Errors in the Assessment of Written Composition With a Many-Faceted Rasch Model, *Journal of Educational Measurement*, 31 (2) , 93–112.
- Engelhard, G. (1996). Evaluating Rater Accuracy in Performance Assessments, *Journal of Educational Measurement*, 33 (1), 56–70.
- Hissey, T.W. (2000). Education and Careers 2000. *Proceedings of the IEEE*, 88(8), 1367-1370.
- Imbrie, P.K., Maller, S.J., and Immekus, J.C. (2005) Assessing Team Effectiveness, Proceedings of the 2005 American Society for Engineering Education Annual Conference & Exposition++, Portland, Oregon.
- Löfberg, J. (2004). YALMIP: A Toolbox for Modeling and Optimization in Matlab, In: *Proceedings of the CACSD Conference*, Taipei, Taiwan.
- MATLAB. <http://www.mathworks.com>.
- Saal, E. F., Downey, G. R., & Lahey, M. A. (1980). Rating the Ratings: Assessing the Psychometric Quality of Rating Data. *Psychological Bulletin*, 88 (2), 413-428.
- Sherrard, W. R., F. Raafat, and R. R. Weaver. (1994). An empirical study of peer evaluations: Students rating students. *Journal of Education for Business* (Volume 70, No. 1) 43-47.
- Smith, D. E. (1986). Training programs for performance appraisal: A review, *Academy of Management Review*, 11 (1), 22-40.
- Sturm, J. F. (1999). Using SeDuMi 1.02, A Matlab Toolbox for Optimization over Symmetric Cones, *Optimization Methods and Software*, 11-12, 625-653.
- Valenti, M. (1996). Teaching Tomorrow's Engineers. *Mechanical Engineering Magazine*, 118(7).
- Vandenbergh, L. & Boyd, S. (1996). Semidefinite Programming, *SIAM Review*, 38 (1), 49-95.
- Weigle, S. C. (1994). Using FACETS to model rater training effects (Draft). Paper presented at the Language Testing Research Colloquium, Washington DC.