Statistical Outlier Detection for Jury Based Grading Systems

Prof. Mary Kathryn Thompson, Technical University of Denmark

Mary Kathryn Thompson is an Associate Professor in the Department of Mechanical Engineering at the Technical University of Denmark. Her research interests include the development, improvement, and integration of formal design theories and methodologies; assessment in project-based engineering design courses; and numerical modeling of micro scale surface phenomena. From 2008 - 2011, Prof. Thompson was the Director of the KAIST Freshman Design Program, which earned her both the KAIST Grand Prize for Creative Teaching and the Republic of Korea Ministry of Education, Science and Technology Award for Innovation in Engineering Education in 2009. She earned her B.S., M.S., and Ph.D. from the Massachusetts Institute of Technology, Department of Mechanical Engineering.

Dr. Line H Clemmensen, Technical University of Denmark

Line H. Clemmensen is an Assistant Professor in the Department of Applied Mathematics and Computer Science at the Technical University of Denmark. She is engaged in statistical research of models for high dimensional data analysis including regularized statistics and machine learning. She is also interested in educational research and is involved in various projects on teaching and learning assessment at the Technical University of Denmark. She earned her M.S. and Ph.D. from the Technical University of Denmark, Department of Informatics and Mathematical Modeling.

Dr. Harvey Rosas, Valparaiso University

Harvey Rosas is an Associate Professor and Director of Undergraduate Studies in the Department of Statistics at the Valparaiso University, Chile. He is currently working on the application of machine learning techniques to the dimensionality reduction problem; Multidimensional Item Response Theory; and text classification. He received his BS in mathematics with an emphasis on the theory of computation from the National University of Colombia and his MS and PhD from the Department of Mathematical Sciences at the Korea Advanced Institute of Science and Technology (KAIST).

A Statistical Outlier Detection Algorithm for Jury-Based Grading Systems

Abstract

This paper presents an algorithm that was developed to identify statistical outliers from the scores of grading jury members in a large project-based first year design course. The background and requirements for the outlier detection system are presented. The outlier detection algorithm and the follow-up procedures for score validation and appeals are described in detail. Finally, the impact of various elements of the outlier detection algorithm, their interactions, and the sensitivity of their numerical values are investigated. It is shown that the difference in the mean score produced by a grading jury before and after a suspected outlier is removed from the mean is the single most effective criterion for identifying potential outliers but that all of the criteria included in the algorithm have an effect on the outlier detection process.

Introduction

Engineering design courses commonly employ jury-based grading systems[1-5] where student projects are evaluated by a group of professionals or experts in the field. Students generally prefer "the involvement of external examiners and jurors" because they believe that this increases the objectivity of the evaluation[6]. The inclusion of multiple raters also helps to balance the differences of opinion that naturally occur during subjective evaluations[7]. However, grading juries also introduce the potential for disagreements in scoring. The use of rubrics can improve the consistency of scores across multiple raters[8-9], but an "extreme reaction to a project by any one particular jury member"[10] can still skew the final score enough to be "problematic"[11].

A number of strategies exist to resolve disagreements in scoring. High stakes assessment typically employs two raters to "independently review an examinee's response and assign a score representing the perceived level of proficiency"[12]. If the two scores are within some pre-defined level of agreement, the scores are averaged to produce the final score. Otherwise, a third party is added to the assessment process. In some cases, the third party reviews the examinee's work and the original evaluations and assigns one of the two original scores to be the final score. In other situations, the third party independently rates the examinee's work. The third score may replace both original scores, be averaged with the two original scores, or be averaged with the closer of the two original scores[12].

More complex techniques are used in athletics. For example, martial arts competitions have four judges and a single referee. Points are awarded, fouls are called, and penalties are given when a majority of the judges agree. However, the referee can "overturn any decision made by the four judges, and as well, assign points and fouls independently, without benefit of consultation or discussion with any other person"[11]. In contrast, the judging system for the International Skating Union employs 9 judges. Of the 9 judges' scores, "the highest and lowest score of each element or program component are ignored" and the remaining 7 scores are averaged[13].

Large-scale jury-based assessment systems require a way to determine when raters disagree and which scores fall outside the pre-defined level of agreement. This is especially important when more than two raters are used and when the rating scale is complex. This paper presents an

algorithm that was developed to identify statistical outliers from the scores of grading jury members in a large project-based first year design course[14-15]. The paper begins with a brief description of the jury-based grading system to provide context for the development of the algorithm. Next, the requirements of the outlier detection system are presented. The outlier detection algorithm and the follow-up procedures for score validation and appeals are described in detail. Finally, the impact of various elements of the outlier detection algorithm, their interactions, and the sensitivity of their numerical values are investigated.

Background

The outlier detection system presented in this paper was developed to ensure the fair and consistent evaluation of 6 deliverables for up to 100 design teams per semester in ED100: Introduction to Design and Communication at the Korea Advanced Institute of Science and Technology. Each grading jury consisted of 2 faculty members and up to 4 teaching assistants who evaluated the work of students from other sections of the course. This system resulted in approximately 3600 scores per semester to be evaluated and averaged.

Jury members in ED100 evaluated the students' worked independently using grading rubrics in an online grading environment[15]. Although grading assignments were not anonymous, discussion among jury members was actively discouraged. For this reason, raters were not permitted to serve on juries with other members of their section. Individual rater's scores were confidential and could be viewed only by the course administration. The average score of each jury was made available to jury members after the initial grading was completed. The online grading pages were re-opened briefly after the preliminary grades were computed to allow the jurors to revised their scores if desired. Outlier detection was performed after the grading pages closed for the final time.

Requirements of an Outlier Detection Scheme for Grading Juries

The outlier detection system developed for ED100 was intended to satisfy three functional requirements (FRs):

FR1: Identify scores that are potential outliers
FR2: Confirm or reject flagged scores as true outliers
FR3: Adjust the final grade to compensate for the presence of outlying scores

To respect the time, effort, and expertise of the raters, two main constraints (Cs) were included:

C1: Jury members cannot be required to score more deliverables than necessary
C2: Scores cannot be indiscriminately removed from the data set

Finally, the selection criteria (SCs) for the outlier detection system were chosen to maximize the accountability and the overall efficiency of the grading system:

SC1: Minimize the number of outliers that go undetected (i.e. minimize false negatives)
SC2: Minimize the number of unnecessarily flagged scores (i.e. minimize false positives)
SC3: Automate the outlier detection and removal process to the extent possible
SC4: Accommodate an arbitrarily large number of design teams and grading jury members

The constraints prohibit score resolution options that automatically drop the highest and lowest scores. The third and fourth selection criteria eliminate the possibility of using a referee or adjudicator for every evaluation like the martial arts competitions described above. The fourth selection criterion also excludes consensus criteria based on identical and adjacent scores like those found in high stakes assessment[11] since this becomes increasingly unlikely as more raters and longer rating scales are used. To fulfill the functional requirements while satisfying the constraints and selection criteria, a more sophisticated statistical approach was chosen instead.

Description of the Outlier Detection Algorithm

The outlier detection algorithm developed for ED100 has one base rule: scores are flagged as potential outliers if they fall outside of 1.5 times the standard deviation of the jury members' scores.

$$\mu - 1.5\sigma < \text{Typical Scores} < \mu + 1.5\sigma \quad (1)$$

Outlier detection based on standard deviation is possible because the distributions of the scores produced by most of the design juries in ED100 are normally distributed. For example, a Shapiro-Wilk's test for normality with a 5% significance level revealed that the null hypothesis was rejected for only 46 out of 564 sets of scores in the Fall 2010 semester (8.2%). If 5% of those observations were due to the statistical significant level, then only 3.2% of score sets were non-normal.

A range of +/- 1.5σ on a normal distribution should flag a maximum of 13.4% of the scores in the course as potential outliers. This is a large enough percentage to minimize the risk of false negatives. However, it could easily flag more scores than could be reviewed in a large course. The standard deviation multiplier value could be further increased to reduce the number of flagged outliers. However, this would also increase the number of true outliers that go undetected, violating SC1. Instead, three additional conditions were added to reduce the number of flagged scores and to satisfy SC2.

First, there can be no outliers if the jury has come to an agreement on the final score. Agreement was defined based on the standard deviation of the jurors' scores. Thus, there can be no outliers if the standard deviation of all scores within the grading jury σ is less than 5%. The cut-off value for jury standard deviation was determined empircally.

$$\text{Typical Scores: } \sigma < 5\% \quad (2)$$

Second, in order to respect the graders and their efforts, no score can be removed if this will not substantially affect the students or their final grades. Thus, there are no outliers if the removal of a flagged score will change the jury mean μ by less than 2.25%. The cut-off value for the change in jury mean was partially determined based on the US 100 point letter grade system: 100 - 97% = A+, 96 - 94% = A, 93 - 90% = A-, etc. A change in mean score greater than or equal to 2.25% was likely to result in a change in letter grade (for example, B+ to A-) and thus was expected to be meaningful to the students. The cut-off value for the change in mean in the 2% to 3% range was determined empirically.

$$\text{Typical Scores: } |\ \mu_{new} - \mu_{old}\ | = \Delta\mu < 2.25\% \quad (3)$$

Finally, it was determined that there can be no outliers if the removal of a score will not meaningfully change the agreement of the jury. Thus, a third criterion was added: there are no outliers if the removal of a flagged score will not change the standard deviation of the jury's scores by more than 2%. The cut-off for change in standard deviation was also determined empirically.

$$\text{Typical Scores: } |\ \sigma_{new} - \sigma_{old}\ | = \Delta\sigma < 2\% \quad (4)$$

The base rule and the three lower bound exemption conditions are sufficient most of the time. However, exceptionally large jury standard deviations can still mask true outliers. Thus, it is recommended to flag juries (rather than individual scores) with standard deviations much greater than 8%. The value for this upper bound can range between 9% and 15% depending on the quality of the grading rubrics, the experience of the graders, the performance of the students, and other factors that can influence the final grading distributions. True outliers that have been masked by the high standard deviation should be obvious by inspection.

$$\text{Hand Check for Atypical Scores: } \sigma > 9 - 15\%$$

An Example of the Outlier Detection Algorithm

The following example demonstrates the outlier detection algorithm. Table 1 shows the scores assigned by 3 different grading juries to 3 different teams during the Fall 2011 semester. Each jury consisted of 2 professors and 3 or 4 teaching assistants. Descriptive statistics for the jury scores (mean and standard deviation) and the calculations for the base rule (+/-1.5σ) are shown in table 2.

For team 1, the mean score is 88.40 and the jury standard deviation is 5.13. Since this deliverable was graded out of 100 points, the standard deviaton and the % standard deviation are the same. The +1.5σ limit is at 96.09 and the -1.5σ limit is at 80.71. The scores from all 5 jury members fall within these boundaries so no outliers are present. For team 2, the jury standard deviation (4.59) is less than 5%. The jury has come to an agreement about this team's score and thus no outliers can be present. However, for team 3, one score (the 48 from grader 3) lies below the lower 1.5σ boundary (54.84). This results in a mean score of 77.83 (C+) and a standard deviation of 15.33 which well above the expected upper limit for a team standard deviation. The score from grader 3 is clearly an outlier and should be removed from the team's average.

Table 1. Scores Assigned by 3 Grading Juries in the Fall 2011 Semester

		Professors			Teaching Assistants		
		Grader 1	Grader 2	Grader 3	Grader 4	Grader 5	Grader 6
Jury 1	Team 1	86	92	87	82	95	
Jury 2	Team 2	85	90	82	85	83	76
Jury 3	Team 3	76	81	48	87	87	88

Table 2. Base Condition Outlier Parameters for Scores Assigned by 3 Grading Juries

		μ	σ	+1.5 σ	-1.5 σ	
Jury 1	Team 1	88.40	5.13	96.09	80.71	No Outliers
Jury 2	Team 2	83.50	4.59			σ < 5%
Jury 3	Team 3	77.83	15.33	100.83	54.84	1 Outlier Detected

The new mean, standard deviation, and +/- 1.5 σ limits for all three teams after outlier removal are show in table 3. Since no scores were removed from the first two sets, these parameters are unchanged. The new mean for team 3 is 83.80 (B) and the new standard deviation is 5.17. This is a substantial change in the students' final grades and affirms the need for outlier detection.

Table 3. Base Condition Outlier Parameters for Scores Assigned by 3 Grading Juries After Outlier Detection and Removal

		μ	σ	+1.5 σ	-1.5 σ	
Jury 1	Team 1	88.40	5.13	96.09	80.71	No Outliers
Jury 2	Team 2	83.50	4.59			σ < 5%
Jury 3	Team 3	83.80	5.17	91.55	76.05	1 Outlier Removed

Reviewing Flagged and Final Scores

Outlier detection algorithms are capable of identifying anomalies in patterns of data, but they cannot determine whether or not a flagged score is truly invalid. After potential outliers are identified, an expert grader must review each flagged score. In ED100, all flagged outliers were reviewed by the course director who was the most experienced grader available.

The review of flagged scores in ED100 is a two-stage process. First, the jury scores and the outlier detection parameters are examined. If a score is obviously an outlier (as seen in the example above), then it is removed from the jury average without further consideration and the students receive the new grade. However, if there is any reason to question either the flagged

score or the jury average, then the original student submission is re-opened and re-evaluated by the expert grader. Based on the findings, a flagged score may be removed as an outlier, one or more scores may be removed as part of an expert grade adjustment, or the expert may choose to replace all existing scores with a new grade.

Although outlier detection and the score reviewing are done with the greatest of care, mistakes are still possible. Thus, two additional rounds of review are included in ED100. First, all course faculty and staff members have the opportunity to review the scores assigned to the students in their sections. Any score can be challenged and a re-grade can be requested. When this occurs, the original student submission is re-opened and re-evaluated by an expert grader. An outlier may be identified and removed, the final score may be adjusted, or the final score may be replaced based on the expert's findings. Finally, students may submit a grade challenge via their faculty adviser after the final grades are released. (The faculty members moderate the process to minimize frivolous requests.) The review process for student challenges is the same as for a faculty and staff challenge.

The possible scenarios during and after the outlier detection process can be summarized as follows:

1. No outliers are flagged. No adjustment is made. All scores are averaged.
2. A score is flagged during outlier detection, reviewed, and validated. No adjustment is made. All scores are averaged.
3. A score is flagged, reviewed, and removed during outlier detection. The remaining scores are averaged.
4. If the expert is suspicious of an individual score or a jury average, the jury scores are reviewed. Outlier removal, expert adjustment (individual scores removed), and a re-grade are possible.
5. If students or their adviser challenge a final score, all jury scores are reviewed. Unflagged outliers may be identified and removed. Expert adjustment and a re-grade are also possible.

Influence of the Lower Bound Conditions in Outlier Detection

As noted above, the multiplier value for the base condition and the cut-off values for the three lower bound conditions were all chosen empirically. This section explores the impact of those decisions on the number of outliers flagged using the scores produced in Fall 2010. We begin by examining the individual impact of each lower bound condition. Next, we examine the impact of pairs of lower bound conditions. Finally, we vary the limit values of the three lower bound conditions. This is not meant as an exhaustive study to determine the optimal values for the outlier detection algorithm. It is merely intended to provide insight into how they affect the total number of flagged outliers and their interplay in this setting.

Effect of the Base Rule Standard Deviation Multiplier

Figure 1 shows the number of scores that would be flagged as outliers for 5 outlier detection schemes over a range of base rule standard deviation multiplier values:

1. No lower bound conditions. Outliers determined by base rule
2. Base rule plus minimum jury standard deviation ($\sigma < 5\%$)
3. Base rule plus minimum change in jury mean ($\Delta\mu < 2.25\%$)
4. Base rule plus minimum change in jury standard deviation ($\Delta\sigma < 2\%$)
5. All lower bound conditions

The first condition shows that multiplier values less than 1.5 provide insufficient screening and result in an excessive number of false positives while multiplier values greater than 2 provide little benefit.

Effect of Individual Lower Bound Conditions

Examination of the individual plots in figure 1 shows that the impact of the conditions based on the minimum jury standard deviation σ and the change in jury standard deviation $\Delta\sigma$ are roughly the same. The lower bound condition based on the change in jury mean $\Delta\mu$ has approximately double the impact of the other two conditions. The minimum change in mean is especially effective with no or low standard deviation multipliers. Combining all the lower bound conditions shows that few additional outliers are flagged when compared to only using only the minimum change in jury mean.

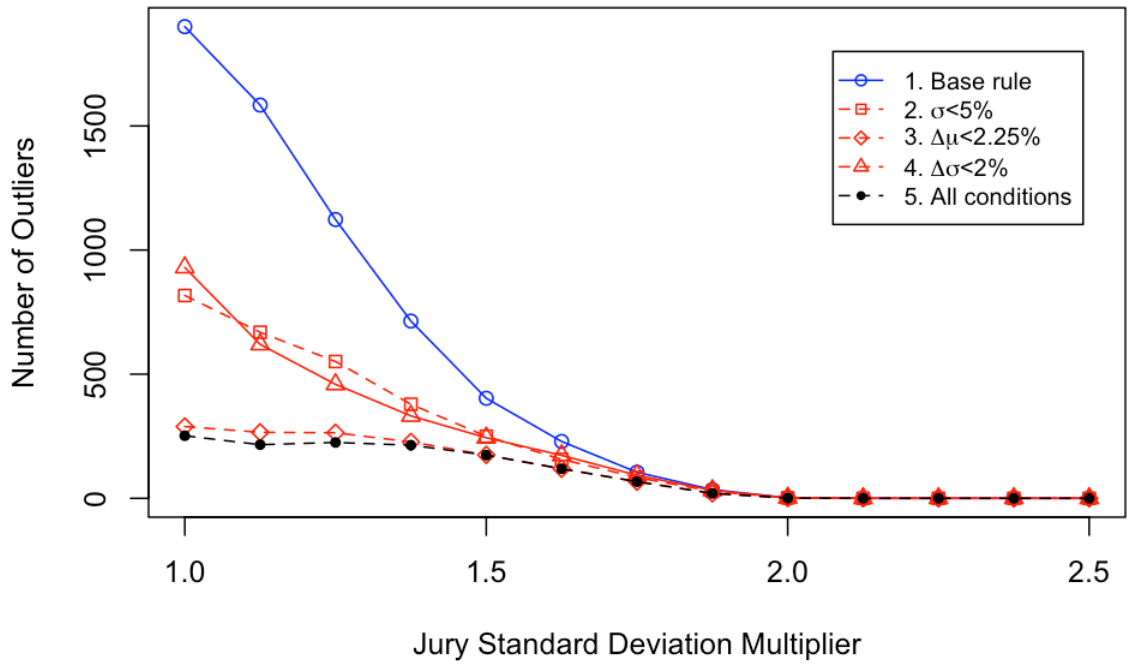

Figure 1. Number of Flagged Outliers as a Function of the Standard Deviation Multiplier Value with One Lower Bound Condition Active at a Time

Effect of Pairs of Lower Bound Conditions

Next, we examine the effect of combining pairs of lower bound conditions. Figure 2 shows the number of scores that would be flagged as outliers for different combinations of lower bound conditions over a range of base rule standard deviation multiplier values:

1. No lower bound conditions. Outliers determined by base rule
2. Base rule plus $\sigma < 5\%$ and $\Delta\mu < 2.25\%$
3. Base rule plus $\sigma < 5\%$ and $\Delta\sigma < 2\%$
4. Base rule plus $\Delta\mu < 2.25\%$ and $\Delta\sigma < 2\%$
5. All lower bound conditions

The results illustrate that combining the standard deviation cut off with the change in jury standard deviation ($\sigma < 5\%$ and $\Delta\sigma < 2\%$) gives the smallest effect in relation to all other combinations of two conditions. The remaining combinations of conditions are only marginally different for a σ multiplier equal to 1.5.

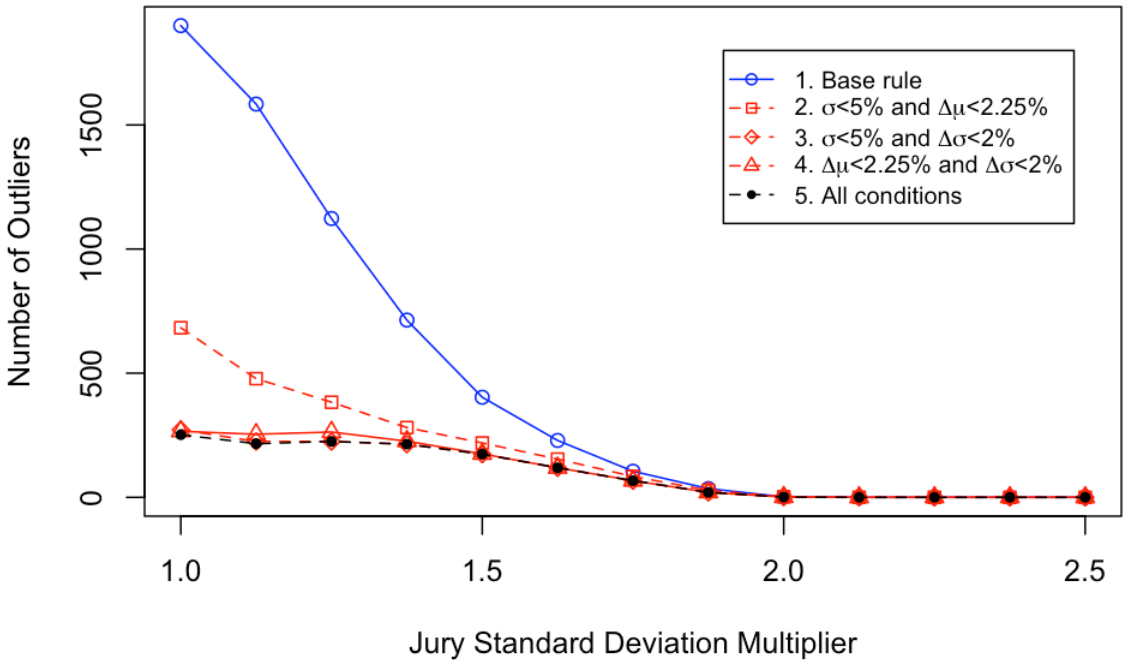

Figure 2. Number of Flagged Outliers as a Function of the Standard Deviation Multiplier Value with Two Conditions Active at a Time

Effect of Minimum Jury Standard Deviation

We now vary the size of each condition while fixing others at their assigned values ($\sigma < 5\%$, $\Delta\mu < 2.25\%$, $\Delta\sigma < 2\%$,). Figure 3 shows the number of scores that would be flagged as outliers for different cut off values of minimum jury standard deviation σ over a range of base rule standard deviation multipliers. The plots demonstrate that the standard deviation cut off value has a marginal effect on the total number of flagged outliers for increasing standard deviation

multiplier values. However, increasing the cut off value to 7% has a noticeable effect even for standard deviation multiplier values larger than 1.5.

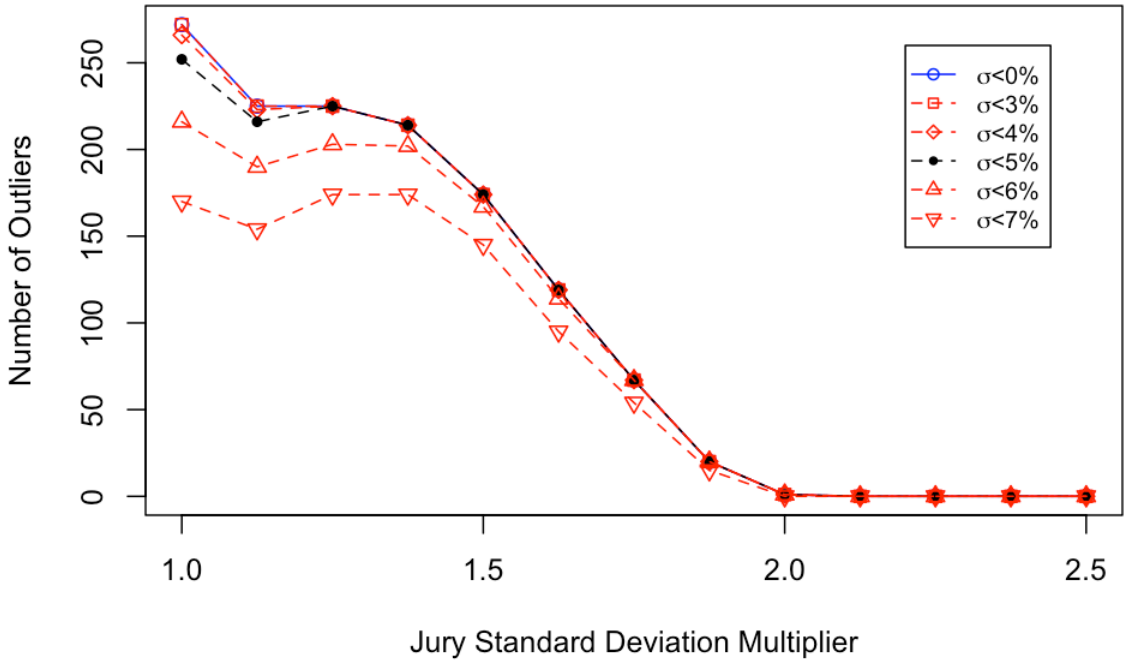

Figure 3. Number of Outliers Flagged For Various Standard Deviation Cut Off Values

Effect of Minimum Change in Jury Standard Deviation

Figure 4 shows the number of scores that would be flagged as outliers for different minimum changes in jury standard deviation $\Delta\sigma$ over a range of base rule standard deviation multipliers. There are notable differences when varying the change in $\Delta\sigma$ for standard deviation multiplier values less than or equal to 1.5. However, substituting $\Delta\sigma < 3$ with $\Delta\sigma < 5$ when the standard deviation multiplier value equals 1.5 only shows a marginal difference in the number of flagged outliers.

Effect of Minimum Changes in Jury Mean

Figure 5 shows the number of scores that would be flagged as outliers for different minimum changes in jury mean $\Delta\mu$ over a range of base rule standard deviation multipliers. These plots illustrate that the minimum jury mean $\Delta\mu$ has an effect for values of the standard deviation multiplier less than 2. Even small changes in the minimum change of mean (for example, from 2.25 % to 2.5%) produce notable changes in the number of flagged outliers for a multiplier of 1.5.

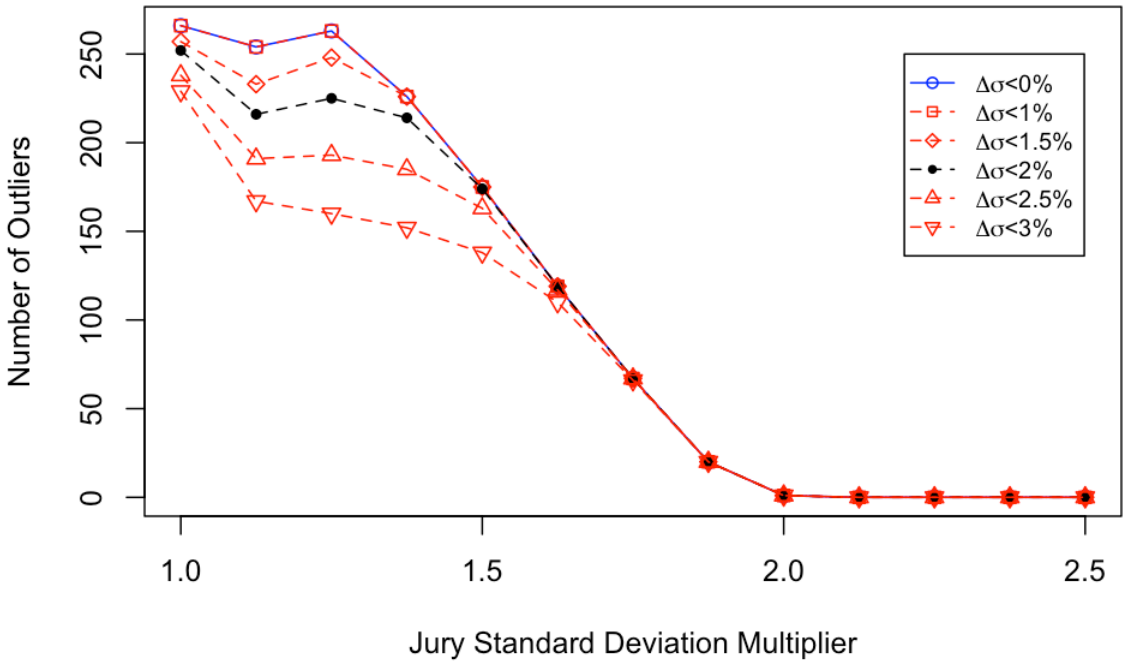

Figure 4. Number of Outliers Flagged For Various Minimum Changes in Jury Standard Deviation

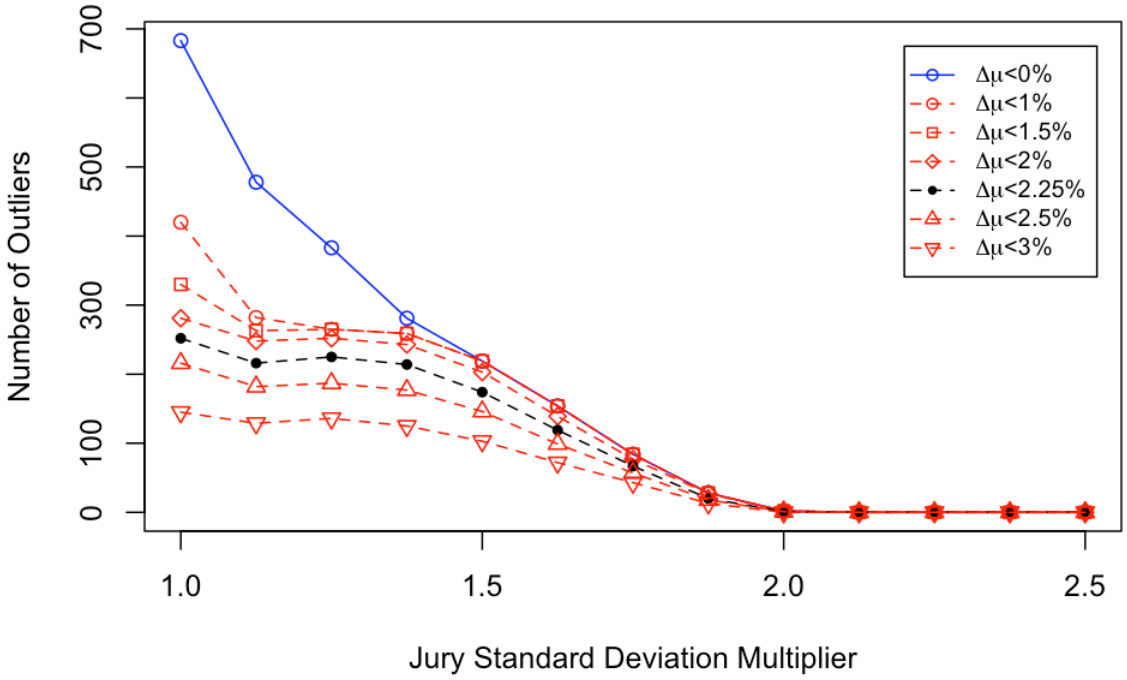

Figure 5. Number of Outliers Flagged For Various Minimum Changes in Jury Mean

Summary and Conclusions

This paper presented an outlier detection algorithm to flag potentially invalid scores produced by groups of up to 6 graders in a large project-based engineering design course. It was shown that for jury standard deviation multipliers less than 1.5, the change in jury mean is the single most effective criterion for identifying potential outliers, but that all of the criteria presented have an effect on the outlier detection process. Increasing the jury standard deviation cut off value from 5% to 7% has a noticeable effect even for standard deviation multiplier values larger than 1.5. This provides some validation for the assumption that atypical scores are usually present for jury standard deviation values above 8% (i.e. atypical scores present when $\sigma > 9$ - 15%). There are notable differences when varying the change in jury standard deviation $\Delta\sigma$ for standard deviation multiplier values less than or equal to 1.5, but this has little or no effect for multiplier values of 1.75 and above. Finally, all considered values of the minimum change in jury mean $\Delta\mu$ produce notable changes in the number of flagged outliers for all jury standard deviation multipler values less than 2. These observations can be used to suggest refinements to the outlier detection algorithm in the future.

Acknowledgements

The authors would like to thank KAIST President Nam P. Suh and the Republic of Korea for creating and sponsoring the KAIST Freshman Design Program and Dean S. O. Park, Dean K. H. Lee, Dean G. M. Lee and Dean S. B. Park for their unwavering support for the program. The authors also would like to acknowledge the ED100 faculty project advisors and teaching assistants for their exceptional dedication. Without their help, the jury-based grading system described in this work would not have been possible. This research was partially supported by a KAIST High-Risk High-Return Research Grant.

References

1. Parker, J., Midkiff, C., Kavanaugh, S. (1996) Capstone senior design at the University of Alabama. *Proceedings of the 26th IEEE Frontiers in Education Conference,* 1, pp. 258-262.
2. Newman, D. J. and Amir, A. R. (2001) Innovative first year aerospace design course at MIT. *Journal of Engineering Education*, 90 (3), pp. 375-382.
3. Raucent, J. (2004) What kind of project in the basic year of an engineering curriculum. *Journal of Engineering Design*, 15 (1), pp. 107-121.
4. Song, S. and Agogino, A. M. (2004) Insights on designers' sketching activities in new product design teams. *Proceedings of the ASME Design Theory and Methods Conference*, pp. 351-360.
5. Saunders-Smits, G. N., Roling, P., Brügemann, V., Timmer, N., Melkert, J. (2012) Using the Engineering Design Cycle to Develop Integrated Project Based Learning in Aerospace Engineering. *Proceedings of the 2012 International Conference on Innovation, Practice and Research in Engineering Education*, pp. 18-20.
6. Salama, A. M. and El-Attar, M. S. T. (2010) Student Perceptions of the Architectural Design Jury. *International Journal of Architectural Research*, 4, (2-3), pp. 174-200.
7. Cook, D. A. and Beckman, T. J. (2009) Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Adv in Health Sci Educ*, 14, pp. 655–664.
8. Kryder L. G. (2003) Grading for Speed, Consistency, and Accuracy, *Business Communication Quarterly* 66 (1), pp. 90–96.
9. Taylor, S. S. (2007) Comments on Lab Reports by Mechanical Engineering Teaching Assistants: Typical Practices and Effects of Using a Grading Rubric, *Journal of Business and Technical Communication*, 21 (4), pp. 402-424.

10. Van Wezemael, J. E., Silbergerger, J. M., Paisiou, S. (2011) Assessing 'Quality': The unfolding of the 'Good' - Collective decision making in juries of urban design competitions. *Scandinavian Journal of Management*, 27 (1), pp. 167-172.
11. Johnson, R. L., Penny, J., and Gordon, B. (2000) The Relation Between Score Resolution Methods and Interrater Reliability: An Empirical Study of an Analytic Scoring Rubric, *Applied Measurement in Education*, 13(2), pp. 121-138.
12. Johnson, R. L., Penny, J., Fisher, S. and Kuhs, T. (2003) Score Resolution: An Investigation of the Reliability and Validity of Resolved Scores, *Applied Measurement in Education,* 16(4), pp. 299-322.
13. Summary of ISU Judging System. *The International Skating Union Official Website*, http://www.isu.org, accessed Jan. 4, 2012.
14. Thompson, M. K. (2012) Fostering Innovation in Cornerstone Design Courses, *International Journal of Engineering Education*, *Special Issue on Design Education: Innovation and Entrepreneurship*, 28 (2), pp. 325-338, 2012.
15. Thompson, M. K. and Ahn, B.-U. (2012) The Development of an Online Grading System for Distributed Grading in a Large First-Year Project Based Design Course (AC 2012-3467). *Proceedings of the 119th ASEE Annual Conference and Exposition.*