International Journal of Engineering Education Vol. 29, No. 6, pp. 1490–1502, 2013
Printed in Great Britain

0949-149X/91 $3.00+0.00

Effect of Rubric Rating Scale on the Evaluation of Engineering Design Projects*

MARY KATHRYN THOMPSON
Department of Mechanical Engineering, Technical University of Denmark, Produktionstorvet, Building 425, Room 217, Lyngby, Denmark, 2800. E-mail: mkath@mek.dtu.dk

LINE HARDER CLEMMENSEN
Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 305, Room 123, Lyngby, Denmark, 2800. E-mail: lkhc@dtu.dk

BEUNG-UK AHN
Department of Computer Science, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong Yuseong-gu, Daejeon, 305-701, Korea. E-mail: elaborate@sparcs.kaist.ac.kr

This paper explores the impact of the rubric rating scale on the evaluation of projects from a first year engineering design course. A small experiment was conducted in which twenty-one experienced graders scored five technical posters using one of four rating scales. All rating scales tested produced excellent results in terms of inter-rater reliability and validity. However, there were significant differences in the performance of each of the scales. Based on the experiment's results and past experience, we conclude that increasing the opportunities for raters to deduct points results in greater point deductions and lower overall scores. Increasing the granularity of the scale can reduce this effect. Rating scales that use letter grades are less reliable than other types of scale. Assigning weights to individual criteria can lead to problems with validity if the weights are improperly balanced. Thus, heavily weighted rubrics should be avoided if viable alternatives exist. Placing more responsibility for the final score on the grader instead of the rubric seems to increase the validity at the cost of rater satisfaction. Finally, rater discomfort can lead to intentional misuse of a rating scale. This, in turn, increases the need to perform outlier detection on the final scores. Based on these findings, we recommend rating scale rubrics that use simple 3 or 4-point ordinal rating scales (augmented checks) for individual criteria and that assign numerical scores to groups of criteria.

Keywords: engineering design; evaluation; rating scale; rubric; psychometrics

1. Introduction

Project-based design courses are an integral part of engineering education. In addition to teaching students about design, these courses expose students to the engineering profession; encourage them to be active learners; allow them to apply their knowledge to real-world problems; and help them to develop professional skills related to teamwork, communication, and project management. Many also provide opportunities to work in an interdisciplinary environment. As a result, these courses often lead to increased student motivation [1, 2], satisfaction [3], and creativity [4], a greater sense of community [3], and higher program retention rates [3].

The way that a design course is evaluated affects how the students approach the subject [4] and how much they learn [5]. Good evaluation also helps faculty members understand how they can improve their courses [3, 6]. Unfortunately, evaluation in project-based design courses is challenging at best. The open-ended and subjective nature of design projects makes these courses poorly suited to examinations with right and wrong answers [7–8]. As a result, faculty members are often unsure of how to evaluate their students [6, 9–11]. In addition, multiple supervisors and/or graders are often necessary for large project-based courses, which can lead to concerns about fairness and consistency in evaluation [2, 12–13].

It is increasingly common to use rubrics to evaluate student deliverables in project-based engineering design courses [6, 10, 13–14]. Scoring rubrics reduce the evaluation subjectivity [15–19] and grading time [20] by explicitly defining the evaluation procedure and criteria. However, the choice of rubric rating scale can significantly impact the reliability and validity of a rubric. It has been shown that differences in age [21], education [22], knowledge [23], experience [24], and motivation [25] can affect the inherent severity of raters, their tendency to choose (or eschew) extreme values on a rating scale, and their tendency to exhibit a yes- or no-saying bias [21]. Raters also vary in how they interpret a rating scale [17], how closely they can and will follow a rubric that has been provided [12, 24, 26–27], and how well they are able to withhold judgment [24].

* Accepted 11 July 2013.

Poster Formatting and Style: (_____ / 15 points)	
The poster was easy to read	0 / ✓- / ✓ / ✓+
The poster was attractive	0 / ✓- / ✓ / ✓+
The poster distributed graphic/blank space/text effectively	0 / ✓- / ✓ / ✓+
The poster made effective use of visual aids	0 / ✓- / ✓ / ✓+

Fig. 1. Example Poster Grading Criteria, Fall 2009.

Since the inherent variation between raters is difficult or impossible to eliminate, much work has been dedicated to understanding and improving rating scale design. Researchers in a variety of fields have explored the influence of rating scale length [22, 28–34], the presence or absence of a scale mid-point [29, 35–36], the use of augmentation to indicate in-between scale ratings [27, 37], the nature of rating scale labels [22–23] and rater training [24, 34] on rating scale accuracy [34], reliability [22, 31, 33–34], validity [22, 28, 33, 36], sensitivity [31], rater response time [29, 33], administration time [22, 29], the proportion of scale used [29], administrator preference [22], and user satisfaction [21, 33]. Unfortunately, their findings have been highly contradictory with no consistent recommendations beyond the encouragement of augmentation.

Payne concluded that rating scales should be chosen based on "the nature of the task and the sophistication of the raters" [38]. This implies that rating scales may perform differently in an educational context and may also depend on the course being offered. Unfortunately, most of the research on ratings scales in education has focused on high stakes assessments where student essays are evaluated on 3 to 6 point Likert scales by 1 to 3 raters [32, 37, 39]. It is unclear how those results translate to the assignment of letter grades on a 100-point scale, when more raters are used, and when engineering design projects are evaluated. In addition, most of the literature examines the influence of various rating scale parameters on single item questions instead of looking at the overall effect on multi-item scales [40]. Thus, the results may have limited applicability to multi-item constructs such as grading rubrics.

The goals of this work were to explore the impact of rating scale on the evaluation of engineering design projects and to choose a rating scale for use in a large mandatory first-year design course in South Korea [41–42]. The paper begins with the background and motivation of the project. This is followed by a description of an experiment in which five technical posters were evaluated by twenty-one experienced raters using four rating scales. A detailed analysis of the experiment results and the follow-up survey are presented. Finally, the various factors that could explain the differences between the rating scales and the limitations of the study are discussed. The paper ends with a summary, conclusions, and recommendations.

2. Background and motivation

The experiment presented in this paper was performed to choose and validate the rubric rating scale used to evaluate the student deliverables in ED100: Introduction to Design and Communication at the Korea Advanced Institute of Science and Technology (KAIST). Teams in ED100 produced two midterm and four final deliverables per semester. Each deliverable was assessed by a jury that consisted of two faculty members and up to four teaching assistants using grading rubrics. The original rubrics used a 4-point ordinal rating scale (zero, check minus, check, or check plus) with integer scores for each category of criteria (Fig. 1). After grading was complete, all scores were analyzed [43] and scores that were flagged as statistical outliers were hand checked by the course director. Scores that were deemed invalid were removed from the grading data set. The remaining scores were averaged. In the Fall 2009 semester, this process resulted in an outlier removal rate of between 2% and 5% for the final deliverables (Table 1). This rate is representative of a typical semester in the course.

Based on feedback from course faculty and staff members, a 7-point letter grade rating scale (A+, A, B, C, D, F, 0) for the individual criteria was introduced in the Spring 2010 semester (Fig. 2). Each category still received an integer score as before. Although the new rating scale was intended to be easier to use, its impact on grading performance was dismal. Outlier removal rates increased between 160% and 900% (Table 1). All grades in the course had to be hand checked and adjusted by

Table 1. Number (percentage) of outliers removed from the deliverable grading averages in the Fall 2009 and Spring 2010 semesters

	Poster	Paper	Technical	Prototype
Fall 2009	8 (2%)	20 (5%)	23 (5%)	12 (3%)
Spring 2010	93 (18%)	93 (18%)	41 (8%)	77 (15%)

Poster Formatting and Style: (_____ / 15 points)							
The poster was easy to read	A+	A	B	C	D	F	0
The poster was attractive	A+	A	B	C	D	F	0
The poster distributed graphic/blank space/text effectively	A+	A	B	C	D	F	0
The poster made effective use of visual aids	A+	A	B	C	D	F	0

Fig. 2. Example poster grading criteria, Spring 2010.

expert graders. Despite these efforts, the final grades in the course had to be curved for the first and only time in its history.

The only explanation for the abrupt change from a seemingly successful grading system to one that was nearly non-functional was the change in rubric rating scale. Since the performance of a given rating scale could not be predicted a priori, the course administration was left with no choice but to determine it experimentally.

3. Methods

3.1 Experiment participants

In order to understand better how rating scales affect grader behavior and to choose a rating scale for use in ED100, an experiment was conducted in which 45 experienced course faculty and staff members (23 professors and 21 TAs) were asked to grade five final posters from previous semesters using a grading rubric with one of four rating scales. All grading was done using an online platform [13] that was developed for the course. All summing operations in all four grading pages were performed automatically. The letter grade equivalents of the assigned numerical scores were calculated and displayed for each category and for all final scores to eliminate cross-cultural number-to-letter grade conversion bias. Scores could be revised and re-saved.

The experiment had a 47% participation rate for a total of 21 raters in groups of four to six. Each rater had taught the course for at least two semesters and had served as a grader at least once. The four grading juries were balanced for age, gender, experience, and nationality to the extent possible. All graders were aware that they were participants in an experiment, that participation was optional, that responses were not anonymous, that the results would not affect the grades of any current students, and that the results would determine the rating scale used in future semesters. Participants were not compensated for their time.

3.2 Deliverables to be graded

The final technical posters were chosen as the deliverable to grade in the experiment because they were generally the fastest and easiest to grade and required the least expert knowledge in design. The five posters used in the experiment were at least two semesters old in order to minimize the chance that the graders would remember the poster, the project, or their final scores. The posters were selected to represent the widest possible range of performance in the course and were expected to receive the following grades: A/A– (poster 3), A–/B+ (poster 5), B+/B (poster 2), B/B– (poster 1), and C or lower (poster 4). It was assumed that the poster number would influence the order in which the participants would download and grade them. Thus, the posters were numbered randomly (except for the requirement that the best and worst posters were neither first nor last) in an attempt to disguise the expected poster rank order.

3.3 Evaluation criteria

The poster evaluation criteria used in the rubric were the same as the ones that were used in the course with one exception. The final category in the poster rubric evaluated the student presentations and their question-and-answer sessions. Since this could not be evaluated online, this group of criteria was removed. As a result, the posters in the experiment were scored out of 85 points instead of 100. The detailed evaluation criteria are shown in the Appendix.

3.4 Rating scales to be tested

The variables considered in this study were: rating scale labels (number, letter grade, or augmented checks), scale length (4, 5, 10, or 12 points), whether the score was calculated by item or by category (i.e. whether individual criteria weights existed), whether the weights of individual criteria were visible or hidden, and the total number of allowable final scores (12 or 101). Given the limited number of graders, these variables could not be tested independently. Only four scales (A–D) were tested to ensure a sufficient number of graders per jury.

Scale A was a numerical interval scale. It required graders to provide an integer score (out of 5 or 10 points) for each assessment criterion (Fig. 3). The sub-totals, final score, and letter grade equivalents were automatically calculated and displayed.

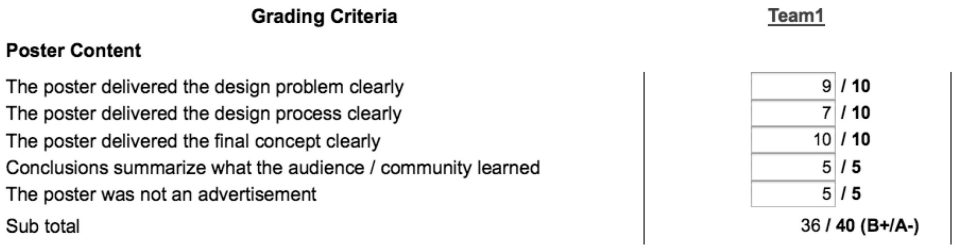

Fig. 3. First group of criteria for rating Scale A.

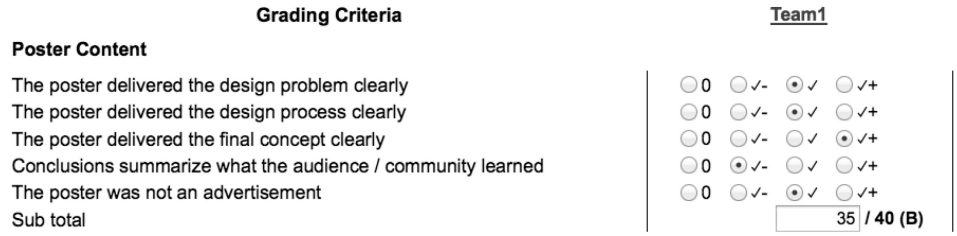

Fig. 4. First group of criteria for rating Scale B.

Scale B was a 4-point ordinal scale (zero, check minus, check, or check plus) with integer scores for each group of criteria (Fig. 4). The raters provided the sub-total values. The final score and letter grade equivalents were automatically calculated and displayed. This scale was successfully used in the course in 2008 and 2009.

Scale C was the same as Scale B, except that the final score had to be chosen from a total of 12 options using a drop down box (Fig. 5). This effectively added a high-pass filter to the final grades and required the grader to reflect on his or her previous choices before assigning the final score.

Scale D provided each of the 12 letter grade options from Scale C for each individual criterion. The raters chose the score for each criterion using a drop down box. The percentage chosen by the rater (100%, 96%, etc.) was multiplied by the weight given to each criterion (5 or 10 points) to determine the numerical score for each criterion. The sub-totals, final score and letter grade equivalents were automatically calculated and displayed.

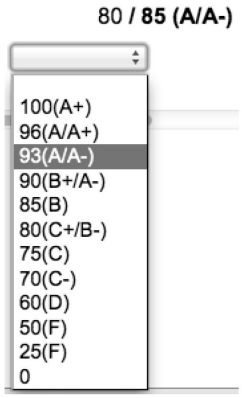

Fig. 5. Drop down box for final score in rating Scale C.

3.5 Surveys

After grading was complete, each grader was asked to fill out a short online survey about their experi-

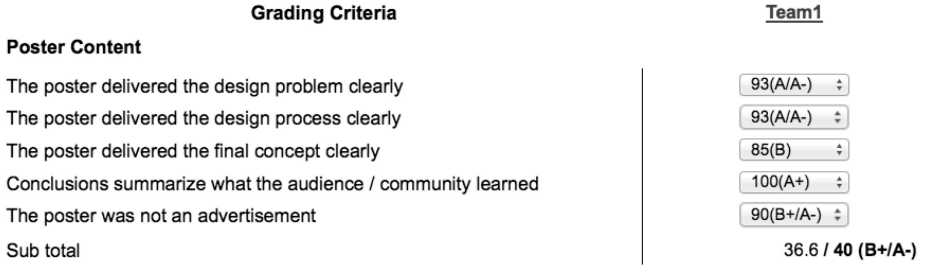

Fig. 6. First group of criteria for rating Scale D.

ences. Graders were asked if the rating scale was easy to use and if they enjoyed using the rating scale (1 = Not at all, 5 = Very much). They were asked if their rating scale should be used for the coming semester (Yes/No), what they liked and disliked most about the rating scale (free response), how they would improve the rating scale (free response), and if they had any additional comments. Survey participation was optional and responses were anonymous.

3.6 Analysis

After the grading data were collected, the scores were normalized so all four scales were calculated out of 100 points. The normalized scores were analyzed using the algorithm presented in [43] to identify potential statistical outliers. Additional scores were flagged by hand. Flagged scores were hand checked and scores that were determined to be 'true' outliers (as opposed to extreme but valid view points) were removed.

A two-way analysis of variance (ANOVA) was performed to determine whether the posters and rating scales could be distinguished statistically. The four rating scales were then assessed by examining their inter-rater reliability using Cronbach's alpha [44] and their validity using expected and actual poster rank order, a T-test, analysis of variance (ANOVA) tests [45], and two-sample F-tests. Finally, the affect of restricting the final grading options in Scale C was explored.

The statistical tests used to examine the differences between the grading scales in this experiment depend on the assumption that there were no differences between the juries. This assumption cannot be tested for the four juries employed in this study. However, an analysis of variance (ANOVA) of a mixed effects model that examined the difference between the assigned grade and the jury mean for all deliverables from the Fall 2010 semester revealed no significant effects from any of the juries. Since these juries were also balanced for age, gender, experience, and nationality to the extent possible, we believe that it is reasonable to assume no jury effect in this work.

4. Results

The final scores assigned by each grader for each poster are shown in Table 2. Of those scores, nine scores were flagged and seven scores were removed as outliers, resulting in a total outlier removal rate of 6.6%. The details of the outlier detection and removal and the impact of their removal on the mean and standard deviation of the final scores are shown in Table 3.

4.1 Distinguishing between posters and rating scales

To determine whether the posters and ratings scales could be distinguished from one another, a two-way analysis of variance (ANOVA) with F-tests for each

Table 2. Final scores by grader, poster, and scale

Scale	Poster	Grader 1	Grader 2	Grader 3	Grader 4	Grader 5	
A	1	82.4	91.8	64.7	74.1	91.8	
A	2	78.8	89.4	78.8	71.8	82.4	
A	3	83.5	94.1	96.5	84.7	92.9	
A	4	63.5	72.9	62.4	58.8	51.8	
A	5	75.3	97.7	83.5	74.1	74.1	
		Grader 6	**Grader 7**	**Grader 8**	**Grader 9**		
B	1	92.9	88.2	69.4	82.4		
B	2	82.4	85.9	72.9	89.4		
B	3	100	97.7	71.8	87.1		
B	4	74.1	75.3	77.7	74.1		
B	5	96.5	68.2	87.1	85.9		
		Grader 10	**Grader 11**	**Grader 12**	**Grader 13**	**Grader 14**	**Grader 15**
C	1	70	75	93	90	90	80
C	2	70	50	75	90	85	90
C	3	96	60	100	93	96	85
C	4	50	25	60	85	60	90
C	5	90	70	90	90	80	90
		Grader 16	**Grader 17**	**Grader 18**	**Grader 19**	**Grader 20**	**Grader 21**
D	1	90.5	92.7	89.7	90.8	84.1	90.2
D	2	90.4	89.6	85.1	88.2	92.9	94.6
D	3	97.5	94.4	96.8	90.9	96.1	95.7
D	4	61.5	84.1	81.4	83.4	79.7	74.1
D	5	83.7	91.5	96.4	90.8	87.8	94.8

Table 3. Statistical outliers flagged and removed from score set

Grader	Score	Old mean	New mean	Old st. dev.	New st. dev.	Flagged by	Removed
2	97.7	80.9	76.8	10.1	4.5	Algorithm	No
7	68.2	84.4	89.8	11.8	5.8	Hand	Yes
8	69.4	83.2	87.8	10.2	5.3	Hand	Yes
8	71.8	89.1	94.9	12.9	6.9	Hand	Yes
11	50	76.7	82	15.4	9.1	Algorithm	Yes
11	60	88.3	94	14.8	5.6	Algorithm	Yes
11	25	61.8	69	23.8	17.5	Algorithm	Yes
11	70	85	88	8.4	4.5	Algorithm	Yes
16	61.5	77.3	80.5	8.6	4	Algorithm	No

effect was performed on the normalized results after outlier removal treating rating scale and poster as fixed effects. This revealed that both the rating scale used (p-value: 0.0001) and the poster being rated (p-value: < 1e–6) had a significant effect on the final poster score. Thus, further investigation of the rating scales seemed warranted.

4.2 Inter-rater reliability of rating scales

To determine the inter-rater reliability of the four rating scales, Cronbach's alpha was computed for each scale and for the combined scores from all of the rating scales before and after outlier removal (Table 4). Removed outliers were replaced with average values for the second calculation. Values of Cronbach's alpha above 0.8 are generally considered to be good and values above 0.9 indicate excellent rater agreement [47]. Thus, all of the rating scales and the combined scores (with the exception of the pre-outlier Scale B) show very good to excellent inter-rater reliability.

To ensure that the low pre-outlier agreement in Scale B was due to one aberrant grader (#8) rather than the rating scale, Cronbach's alpha was also calculated for the raw scores from Scale C. The raw scores were obtained by summing the numerical values for each category rather than by using the final scores from the drop-down menu (Table 6). This resulted in a Cronbach's alpha of 0.84 and confirmed the inter-rater reliability of Scale B.

4.3 Validity of rating scales: poster ranking

A first estimate of the validity of the four rating scales was determined by comparing the rank-order of the five posters from each of the four grading scales to the rank order that was expected by the experiment's designers (Table 5). This shows that all four rating scales were successfully able to identify the best (3), second best (5) and worst (4) posters. However, three of the four scales (A, B, and C) disagreed on the relative ranking of posters 1 and 2. One scale (A) was totally unable to distinguish between posters 2 and 5. In addition, two scales (A and D) provided very poor differentiation between posters 1, 2, and 5.

The apparent disagreement in poster ranking seems to be due to the nature of the posters themselves rather than a failing of the grading rubric or the individual rating scales. For example, Poster 1 has major weaknesses in content, mechanics and presentation. However, the information that it does provide is exceptionally clear. In contrast, Poster 2 has much more and better technical content but lacks clarity. Similarly, Poster 5 provides an excellent description of the design process and outstanding visual aids, however it does not include any references (a major point deduction) and would benefit from additional text. It has been observed that different raters 'observe and value different things' [44] and that despite 'similar training, different scorers may focus on

Table 4. Cronbach's alpha for each scale and for all scales with and without outlying scores

	Scale A	Scale B	Scale C	Scale D	Combined
Cronbach's alpha (all scores)	0.92	0.22	0.84	0.91	0.96
Cronbach's alpha (outliers removed)	0.92	0.9	0.88	0.91	0.97

Table 5. Expected and actual poster rankings and mean scores (outliers removed)

Expected	Scale A	Scale B	Scale C	Scale D	Combined
3	3 (90.4)	3 (94.9)	3 (94)	3 (95.2)	3 (93.6)
5	5 (80.9)	5 (89.8)	5 (88)	5 (90.8)	5 (87.3)
2	1 (80.9)	1 (87.8)	1 (83)	2 (90.1)	1 (85.2)
1	2 (80.2)	2 (82.7)	2 (82)	1 (89.7)	2 (84.1)
4	4 (61.9)	4 (75.3)	4 (69)	4 (77.3)	4 (71)

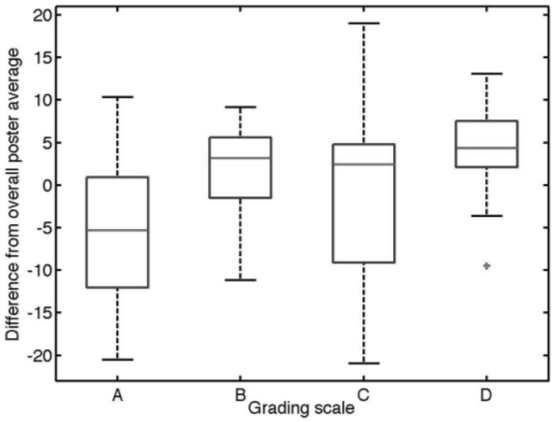

Fig. 7. Box plot showing differences from the overall mean by grading scale.

different . . . features' including some 'that are not cited in the scoring rubric' [24]. Thus, it is possible that three very different posters could receive similar or interchangeable scores even with the aid of a good grading rubric and a fully functional rating scale—and that these scores could be equally valid.

4.4 Validity of rating scales: significant effects in the differences from the overall mean

An inspection of Table 3 reveals that Scale A produced the minimum score in 5 out of 5 cases while Scale D produced the maximum score in 5 out of 5 cases. To determine if these differences were significant, the mean scores from each rating scale were compared with the overall mean ('true' score) for each poster after outlier removal (Fig. 7). A difference of means test (t-test) confirmed that the differences between the means of Scale A (p = 0.0032) and Scale D (p = 3.1e–5) and the overall mean were statistically significant, while the difference between the means of Scales B (p = 0.18) and C (p = 0.65) were not. Thus, Scale A seemed to consistently under-estimate the final score, while Scale D seemed to consistently over-estimate it.

4.5 Validity of rating scales: comparison of means and variances

For a more rigorous investigation of the differences between the four rating scales and the overall mean, two additional series of ANOVA tests were performed. First, a series of ANOVAs with F-tests was performed with the null hypothesis that the means of the differences between the assigned and 'true' scores for each of the four rating scales are equal. The analyses showed that the mean of Scale A was significantly different from Scales B and D (p = 0.0028 and p = 1.2e–6) but was indistinguishable from Scale C (p = 0.0906). The mean of Scale C was significantly different from Scale D (p = 0.0076), but was indistinguishable from Scale B (p = 0.26). The means of scales B and D were indistinguishable (p = 0.0975).

Second, a series of F-tests was performed with the hypothesis that the variances of the differences between the assigned and 'true' scores for each pair of the four rating scales are equal. The variance of Scale A was significantly different from that of Scale D (p = 0.0085) whereas it was indistinguishable from the variances of Scales B and C (p = 0.076 and p = 0.48). The variance of Scale B was significantly different from that of Scale C (p = 0.019) but not from Scale D (p = 0.67). The variances of Scales C and D were significantly different (p = 8e–4).

This analysis showed that Scales A and C were statistically indistinguishable when comparing their ability to produce scores close to the overall mean due to their large variances. Likewise, Scales B and D were indistinguishable due to their similar means and small variances.

When these three perspectives on validity are viewed as a whole, it seems logical to conclude that all four rating scales are valid. However, Scale B is able to distinguish between all five posters (unlike Scales A and D), has a close correlation to the overall mean (unlike Scale A), and has a small variance (unlike Scales A and C). Thus, it seems reasonable to conclude that Scale B is perhaps a bit more valid than the others.

4.6 Influence of reduced options for final scores

To better understand the impact of reducing the total possible final scores from 101 to 12, we compared the raw scores from Scale C (Table 6) with the ones that were submitted using the drop-down menu (Table 7). Two of the graders (10 and 11) systematically rounded their raw scores up or down to the closest permissible score. Two graders (14 and 15) both chose to round in the opposite direction once but otherwise were very faithful to

Table 6. Raw scores from Scale C (all scores)

Poster	Grader 10	Grader 11	Grader 12	Grader 13	Grader 14	Grader 15
1	70.6	72.9	92.9	87.1	88.2	80
2	70.6	44.7	71.8	87.1	84.7	87.1
3	96.5	62.4	100	92.9	96.5	84.7
4	52.9	27.1	32.9	78.8	68.2	88.2
5	89.4	65.9	84.7	84.7	78.8	87.1

Table 7. Final scores from Scale C (all scores)

Poster	Grader 10	Grader 11	Grader 12	Grader 13	Grader 14	Grader 15
1	70	75	93	90	90	80
2	70	50	75	90	85	90
3	96	60	100	93	96	85
4	50	25	60	85	60	90
5	90	70	90	90	80	90

their original scores. The last two graders (12 and 13) had some major departures from their raw scores, always rounding up. The average difference between the raw and final scores was quite small: 1.69 points per poster. In no case did the adjustment affect poster order. These results indicate that limiting the final score options does successfully force some reflection and revision on the part of the rater. However, the increased variance in Scale C compared with Scale B (which is otherwise equivalent) implies that rater reflection is occurring anyway and that reflection at a finer resolution (looking at individual criteria or groups of criteria rather than the poster as a whole) leads to better final scores.

4.7 Grader satisfaction and feedback

Finally, we examined the results from the rater surveys to better understand the raters' perspectives (Table 8). Scale D received the best feedback on the survey, with respondents saying that it was both easy (4.40/5) and enjoyable to use (4.00/5), followed by Scale A. However, these were also the only two scales that did not receive unanimous recommendations for use in the following semester.

In the free response section, raters from Scales A, C, and D expressed discomfort with their assigned rating scale and suggested changes to their scale or the use of another scale. However, there was no consensus about which scale to use or what changes to make. (Scale B received no feedback.) This is consistent with other faculty and staff surveys in ED100 over the years and emphasizes the need to choose rating scales experimentally.

4.8 Selection and performance of Scale B

As noted above, the primary motivation of this work was to choose a rating scale for use in a large project-based engineering design course. Based on the results above and its previous success in the course, Scale B was chosen for use in the Fall 2010 semester. After its reinstatement, the percentage of removed outliers ranged from 0% to 6% for all of the course deliverables. We consider this to be final validation of the choice of Scale B.

Table 8. Summary of survey results

	Easy to use	Enjoyable	For/Against
Scale A	4.2	3.6	3 to 2
Scale B	3.3	3.3	3 to 0
Scale C	3.7	3.3	3 to 0
Scale D	4.4	4	5 to 1

5. Discussion

There are four major limitations of the study: the sample sizes of the juries, the lack of a control across the four juries, the subjective nature of outlier detection and removal, and the fact that the raters were aware that they were participating in an experiment.

5.1 Insufficient sample size

By experimental standards, the sample size (four to six raters) for all four juries is small. The sample size is particularly concerning for Scale B, which had a total of four respondents, and only three scores for three of the five posters after outlier removal. It is reasonable to question whether scale B is comparable to the other scales, which had up to twice as many responses. It is also logical to conclude that the sample size is not large enough to definitively prove anything about the rating scales. However, we believe that general conclusions can still be offered for three reasons.

First, in the past, ED100 juries have consisted of: (a) pairs of expert raters who discussed and compared results after completing their initial evaluations (3 semesters), (b) teams of four raters who evaluated the projects independently (1 semester—Fall 2009), and (c) teams of six (+/–1) raters who evaluated the projects independently (4 semesters). Teams with four raters tend to produce higher standard deviations than larger teams. This makes outlier detection more difficult. However, teams with four raters are still able to reach a consensus, as indicated in Table 1. We may question whether the particular four-rater jury for Scale B functioned well, but there is sufficient evidence from the Fall 2009 semester that juries with only four members perform reasonably well in general.

Second, the results from Scale B could be (and were) supplemented with the raw scores from Scale C as shown in Tables 6 and 7. When combined, the scores from Scale B and the raw scores from Scale C represent between 8 and 10 data points, depending on the outlier detection. These combined scores are still able to distinguish between all five posters and

still have a close correlation to the overall mean. (The combined scores do have a larger variance than Scale B but the larger number of responses and the reduced reflection on the part of the graders probably cause this.) Thus, the combined scores also support the recommendation of Scale B.

Finally, Scale B had been used successfully during the four semesters before and the three semesters after this experiment was performed. Thus, its performance and stability is well understood. The primary purpose of the experiment was to determine if any of the other scales performed better than Scale B. Although we cannot say definitively that Scale B performed better than the other scales, there seems to be sufficient evidence to conclude that the other rating scales did not outperform Scale B.

5.2 No inter-jury control

The experiment described in this work did not have a formal control across all of the juries using a separate poster and rating scale. This was done because of concerns that the interpretation and usage of a second rating scale would be influenced by the first. As noted in Section 3.6, no significant effects were observed on grading from jury formulation in other semesters of the course so it seems reasonable to assume no jury effect in this work as well.

5.3 Subjective outlier detection

In this study, outliers were flagged automatically by a formal algorithm and manually by an expert rater. The decisions to keep or remove a flagged outlier were also made by an expert rater. Because of the subjective nature of design projects, outlier detection and removal will always partially depend on personal judgment. In addition, the algorithm that was used for outlier detection was developed for a four-rater jury but optimized for a six-rater jury. Thus, additional judgment was required for outlier detection in the smaller juries (namely the jury for Scale B).

Judgments related to outlier flagging and detection in this work were accepted because they were made by the most expert rater in ED100—an individual who had participated in the grading of almost 800 ED100 projects over four years. This does not guarantee that the decisions made were free of bias. But there exists no better alternative for outlier detection at this time.

5.4 Low stakes evaluation

Finally, all raters in this work knew that they were participating in an experiment and that an unfair rating would not negatively impact students. This can be seen by rater #11's willingness to assign 'harsh', 'misleading' and 'low' scores (see Section 6.6 below). Evaluation done in an authentic context is likely to be done more carefully and with more reflection and revision. This is evidenced by the fact that the outlier detection rate in this experiment was slightly higher than normally observed in ED100. However, we believe that most of this effect was removed through the outlier detection process and it is not expected to have had a major impact on the experiment's results.

6. Discussions

Based on the results of the experiment described above and our experiences with rating juries in ED100, we offer seven potential conclusions and topics for future research.

6.1 More opportunity for point reductions leads to lower grades

One of the participants in the experiment correctly predicted that Scale A would consistently undervalue the students' work. She hypothesized that raters' standards for perfect scores would be very high and that they would tend to deduct at least one point for each criterion rather than assigning the full score. Thus, providing raters with more opportunities to assign (or deduct) points on a rubric would generally lead to a lower total score. Since Scale A offers 14 opportunities for point reduction, this reasoning implies that most scores should be lower than 86/100 and indeed this is the case. This behavior is consistent with Dolnicar's [21] observation that Asian respondents (specific to this study) and individuals with higher education levels (most graders) tend to use the extreme options on rating scales less than others.

6.2 A finer rating scale reduces the impact of increased opportunity for point deductions

Although Scales A and D provide an equal number of opportunities to deduct points, Scale D did not exhibit the same score ceiling that Scale A did. This may be because Scale D uses a finer rating scale than Scale A. Scale D allows the rater to deduct a minimum of 0.2 points (4% of 5 points) from each criterion instead of a full point (20% of 5 points). With 14 opportunities for point reduction at the finer scale, this reasoning implies that most scores from Scale D should be lower than 96/100 and indeed this is also the case.

6.3 Letter grade based rating scales may be unsuitable for grading rubrics

It is well known that if raters are 'presented with a scale in which their attitude is non-mid-point, then they will subjectively divide the range between the values that they recognize as being consistent with their own attitude range and, as a consequence,

exhibit a narrower response range than the presented scale intends' [31]. The mid-point of a US-type letter grading scale is usually in the B range (85) rather than at the mid-point of the scale (50), which represents a clear failing grade. As a result, raters naturally limit their usage to the upper most 30% of the letter grade scale (85 ± 15). This produces grade that naturally fall well above 70/100 and makes a failing grade nearly impossible to receive using a letter grade scale—even if it is well deserved.

We tried to account for this behavior both in this experiment and in the scale offered during the Spring 2010 semester by presenting the raters with a letter-grade scale that offered more options at the higher end of the scale, however neither scale functioned as expected. The raters used too few of the options in Scale D, while the raters from Spring 2010 exhibited strong variations in their interpretation of the scale, which lead to extreme disagreement between the raters. It may be that letter grades are simply too culturally dependent, imprecise, and ambiguous to be used in this context.

6.4 Unbalanced criteria weights lead to unbalanced grades; heavily weighted scales should be avoided

Scale D assigned weights to the rater's responses by transforming their letter grades into percentages. It also assigned weights to the individual criteria in the rubric by offering 5 or 10 points for each response. Both weights needed to be chosen correctly to ensure that the sub-totals and final scores match the rater's intention. Given the high mean and low variance of the scores from Scale D, it seems reasonable to conclude that this scale did not adequately capture the opinions of the raters or the performance of the students and thus was not properly weighted.

With enough time and data, the relative weightings of each letter grade in Scale D could be adjusted. However, there is a substantial risk that students will receive unfair and inconsistent grades while the weighting data are being collected. We assert that it is not reasonable or practical to use a scale where the weightings are difficult to determine a priori when viable alternatives exist.

6.5 Increased responsibility of the rater decreases comfort but increases reflection and improves validity

Scales A and D share an additional disadvantage: both reduce the responsibility of the rater for the final grade. Scales A and D required the raters to assign scores only to individual criteria. The sub-totals and final scores were calculated by the online grading system. In the survey comments, some raters expressed dissatisfaction with the scores that these rubrics produced. However, it appears that only some of the raters (for example, graders #2 and #20) attempted to correct for this systematically. Other raters (for example, graders #3, #16, and #21) only seemed to make corrections in the extreme cases (Posters 3 and 4). In most cases, the raters seemed to defer to the rubric in determining the final score.

In Scales B and C, there were no weights given to the individual criteria and no guidance was given about how to convert their responses into a numerical grade. Thus, the rater—and not the rubric—was ultimately responsible for the final grade. Based on the examination of the raw and final scores from Scale C, there is evidence that the raters in both Scales B and C were actively engaging in reflection during the grading process. We believe that this reflection is partially responsible for the success of those two scales.

Unfortunately, this improvement in validity is not without a cost. As noted in the survey comments, increased ambiguity in the rating scale and increased responsibility of the grader is uncomfortable. This leads to lower ratings for ease of use and scale satisfaction.

6.6 Rater discomfort can lead to intentional misuse of the rating scale

In extreme cases, discomfort with increased responsibility can lead raters to intentionally misuse the rating scale. For example, in a post-experiment e-mail, rater #11 said that he developed his own rating-scale-to-score conversion system by dividing the group point values between the criteria in the group and then further sub-dividing those points by the three options in the rating scale (check minus, check, and check plus). He described this as a 'straightforward approach that [did] not require any . . . thinking.' He acknowledged that his grades were 'harsh,' 'misleading,' and 'low', but since they were assigned systematically, he did not correct them.

6.7 Misuse of the rating scale increases outliers; outlier detection is important

The misuse of a rating scale is likely to lead to the assignment of unfair or invalid scores. This is evident by the fact that four of rater #11's five scores were removed from the data set. It also underscores the importance of outlier detection in any jury-based grading system.

7. Concluding remarks

This work presented the results of an experiment that was designed to choose a rating scale for use in a large project-based engineering design course and to improve our understanding of the influence of rating scales on design rubric performance. In the experiment, twenty-one experienced graders scored

five technical posters using one of four rating scales. It was shown that all four rating scales tested produce excellent results in terms of inter-rater reliability and validity. However, the rating scale that required a numerical score for each criterion in the rubric (A) consistently under estimated the final score, while the scale that required a letter grade estimate for each criterion in the rubric (D) consistently over estimated the final score. In addition, reducing the number of possible final scores from a 101-point continuous scale (B) to a 12-point discontinuous scale (C) (i.e. introducing a high pass filter) significantly increased the variance of the final scores.

Based on the experiment's results and past experience, we conclude that increasing opportunities for raters to deduct points results in greater point deductions and lower overall scores. Increasing the granularity of the scale can reduce this effect. Rating scales that use letter grades are less reliable than other types of scales. Assigning weights to individual criteria can lead to problems with validity if the weights are improperly balanced. Thus, heavily weighted rubrics should be avoided if viable alternatives exist. Placing more responsibility for the final score on the grader instead of the rubric seems to increase validity at the cost of rater satisfaction. Finally, rater discomfort can lead to intentional misuse of a rating scale. This, in turn, increases the need to perform outlier detection on the final scores.

The final scale selected for this work was a simple four-point ordinal rating scale using augmented checks that assign numerical scores to groups of criteria instead of to individual criteria. This scale appeared to have the best ability to distinguish between different posters, had the closest correlation to the overall mean, and the smallest variance of the four scales tested. The chosen scale continued to perform well for several semesters after the conclusion of the experiment and we believe could be used in other design courses with equal success.

Acknowledgements—The authors would like to thank Dr. Harvey Rosas, Ms. Monica Pena, and Ms. Eleonora Ibragimova for their help in designing the experiment described in this work; all of the participants for their involvement with the rating scales experiment; KAIST President Emeritus 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 support of the KAIST Freshman Design Program and related research. This work was partially supported by a KAIST High Risk High Return Grant.

References

1. P. C. Powell, Assessment of team-based projects in project-led education, *European Journal of Engineering Education*, **29**(2), 2004, pp. 221–230.
2. H. Kilic, M. Koyuncu and M. Rehan, Senior graduation project course for computing curricula: An active learning approach, *International Journal of Engineering Education*, **26**(6), 2010, pp. 1472–1483.
3. S. M. Nesbit, S. R. Hummel, P. R. Piergiovanni and J. P. Schaffer, A design and assessment-based introductory engineering course, *International Journal of Engineering Education,* **21**(3), 2005, pp. 434–445.
4. D. Oehlers and D. Walker, Assessment of deep learning ability for problem solvers, *International Journal of Engineering Education,* **22**(6), 2006, pp. 1261–1268.
5. J. W. Pellegrino, N. Chudowsky and R. Glaser (eds), *Knowing What Students Know: The Science and Design of Educational Assessment*, Committee on the Foundations of Assessment, Center for Education, National Research Council, 2001.
6. R. Bailey and Z. Szabo, Assessing engineering design process knowledge, *International Journal of Engineering Education*, **22**(3), 2006, pp. 508–518.
7. P. Wellington, I. Thomas, I. Powell and B. Clarke, Authentic assessment applied to engineering and business undergraduate consulting teams, *International Journal of Engineering Education*, **18**(2), 2002, pp. 168–179.
8. K. J. Reid and E. M. Cooney, Implementing rubrics as part of an assessment plan, *International Journal of Engineering Education,* **24**(5), 2008, pp. 893–900.
9. L. McKenzie, M. Trevisan, D. Davis, S. Beyerlein and Y. Huang, Capstone design courses and assessment: A national study, *Proceedings of American Society for Engineering Education Annual Conference*, Salt Lake City, UT, 2004.
10. D. Davis, M. Trevisa, R. Gerlick, H. Davis, J. McCormack, S. Beyerlien, P. Thompson, S. Howe, P. Leiffer and P. Brackin, Assessing team member citizenship in capstone engineering design courses, *International Journal of Engineering Education*, **26**(4), 2010, pp. 771–783.
11. J. McCormack, S. Beyerlein, D. Davis, M. Trevisan, J. Lebeau, H. Davis, S. Howe, P. Brackin, P. Thompson, R. Gerlick, M. J. Khan and P. Leiffer, Contextualizing professionalism in capstone projects using the IDEALS Professional Responsibility Assessment, *International Journal of Engineering Education*, **28**(2), 2012, pp. 416–424.
12. J. Jawitz, S. Shay and R. Moore, Management and assessment of final year projects in engineering, *International Journal of Engineering Education*, **18**(4), 2002, pp. 472–478.
13. M. K. Thompson and B.-U. Ahn, The development of an online grading system for distributed grading in a large first year project-based design course, *119th ASEE Annual Conference and Exposition,* San Antonio, TX, 2012.
14. M. Steiner, J. Kanai, C. Hsu, R. Alben and L. Gerhardt, Holistic assessment of student performance in multidisciplinary engineering capstone design projects, *International Journal of Engineering Education*, **27**(6), 2011, pp. 1259–1272.
15. B. M. Moskal and J. A. Leydens, Scoring rubric development: Validity and reliability, *Practical Assessment, Research and Evaluation,* **7**(10), 2000.
16. C. A. Mertler, Designing scoring rubrics for your classroom, *Practical Assessment, Research and Evaluation*, **7**(25), 2001.
17. S. E. Stemler, A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability, P*ractical Assessment, Research & Evaluation,* **9**(4), 2004.
18. R. Tierney and M. Simon, What's still wrong with rubrics: Focusing on consistency of performance criteria across scale levels, *Practical Assessment, Research & Evaluation,* **9**(2), 2004.
19. J. A. Marin-Garcia and C. Miralles, Oral presentation and assessment skills in engineering education, *International Journal of Engineering Education*, **24**(5), 2008, pp. 926–935.
20. L. Anglin and K. Anglin, Improving the efficiency and effectiveness of grading through the use of computer-assisted grading rubrics, *Decision Sciences Journal of Innovation Education*, **6**(1), 2008, pp. 51–73.
21. S. Dolnicar, Are we drawing the right conclusions? The dangers of response sets and scale assumptions in empirical tourism research, *Proceedings of the 5th Conference on Consumer Psychology in Tourism and Hospitality*, 2005.

22. W. Davis, T. R. Wellens and T. J. DeMaio, Designing response scales in an applied setting, *Proceedings of the 51st Annual Conference of the American Association for Public Opinion Research*, Salt Lake City, Utah, 1996.
23. C. M. Myford, Investigating design features of descriptive graphic rating scales, *Applied Measurement in Education*, **15**(2), 2002, pp. 187–215.
24. E. E. Wolfe, The relationship between essay reading style and scoring proficiency in a psychometric scoring system, *Assessing Writing*, **4**(1), 1997, pp. 83–106.
25. L. J. Cronbach, *Essentials of Psychological Testing*, 5th edn, Harper & Row, New York, 1990.
26. L. Hand and D. Clewes, Marking the difference: an investigation of the criteria used for assessing undergraduate dissertations in a business school, *Assessment and Evaluation in Higher Education*, **25**(1), 2000, pp. 5–21.
27. E. W. Wolfe and D. H. Gitomer, The influence of changes in assessment design on the psychometric quality of scores, *Applied Measurement in Education*, **14**(1), 2001, pp. 91–107.
28. M. S. Matell and J. Jacoby, Is there an optimal number of Likert scale items? Study I: Reliability and validity, *Educational and Psychological Measurement,* **31**, 1971, pp. 657–674.
29. M. S. Matell and J. Jacoby, Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology,* **56**(6), 1972, pp. 506–509.
30. G. D. Jenkins and T. D. Taber, A Monte Carlo study of factors affecting three indices of composite scale reliability, *Journal of Applied Psychology*, **62**(4), 1977, pp. 392–398.
31. R. A. Cummins and E. Gullone, Why we should not use 5-point Likert scales: The case for subjective quality of life measurement, *Proceedings of the Second International Conference on Quality of Life in Cities*, National University of Singapore, Singapore, 2000, pp. 74–93.
32. R. L. Johnson, J. Penny and B. Gordon, The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric, *Applied Measurement in Education*, **13**(2), 2000, pp. 121–138.
33. C. C. Preston and A. M. Coleman, Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences, *Acta Psychologica*, **104**, 2000, pp. 1–15.
34. D. A. Cook and T. J. Beckman, Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX, *Advances In Health Science Education,* **14**, 2009, pp. 655–664.
35. R. Garland, The mid-point on a rating scale: Is it desirable? *Marketing Bulletin*, **2**, 1991, pp. 66–70.
36. T. A. Litzinger, S. H. Lee, J. C. Wise and R. M. Felder, A psychometric study of the index of learning styles, *Journal of Engineering Education*, **96**(4), 2007, pp. 309–319.
37. J. Penny, R. L. Johnson and B. Gordon, The effect of rating augmentation on inter-rater reliability. An empirical study of a holistic rubric, *Assessing Writing*, **7**, 2000, pp. 143–164.
38. D. A. Payne, *Applied Educational Assessment,* Wadsworth, Belmont, CA, 1997.
39. P. J. Congdon and J. McQueen, The stability of rater severity in large-scale assessment programs, J. *Ed. Measurement*, **37**(2), 2000, pp. 163–178.
40. J. A. Gliem and R. R. Gliem, Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales, *Proceedings of the Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, 2003.
41. M. K. Thompson, Increasing the rigor of freshman design education, *Proceedings of the International Association of Societies of Design Research Conference*, Seoul, South Korea, 2009.
42. M. K. Thompson, Fostering innovation in cornerstone design courses, *International Journal of Engineering Education*, **28**(2), 2012, pp. 325–338.
43. M. K. Thompson, L. H. Clemmensen and H. Rosas, Statistical outlier detection for jury based grading systems. *Proceedings of the 120th ASEE Annual Conference and Exposition,* June 23–26, 2013, Atlanta, GA
44. L. J. Cronbach, Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**(3), 1951, pp. 297–334.
45. R. Johnson, J. Freund and I. Miller, *Probability and Statistics for Engineers*, Pearson Education, 8th edn, 2011.
46. D. George and P. Mallery, *SPSS for Windows step by step: A simple guide and reference, 11.0 update*, 4th edn, Allyn & Bacon, Boston, 2003.
47. D. A. Cook, D. M. Dupras, T. J. Beckman, K. G. Thomas and V. S. Pankratz, Effect of rater training on reliability and accuracy of mini-CEX Scores: A randomized, controlled trial, *J. Gen. Intern. Med.*, **24**(1), 2008, pp. 74–79.

Appendix

Rating Scales Experiment Grading Rubric Criteria and Point Values

Poster Content: (____ / 40 points)	
The poster delivered the design problem clearly	10 points
The poster delivered the design process clearly	10 points
The poster delivered the final concept clearly	10 points
Conclusions summarize what the audience / community learned	5 points
The poster was not an advertisement	5 points
Poster Effectiveness: (_____ / 10 points)	
The poster was persuasive and convincing	5 points
The poster was able to 'stand on its own' without other help	5 points
Poster Formatting and Style: (_____ / 15 points)	
The poster was attractive and easy to read	5 points
The poster distributed graphic/blank space/text effectively	5 points
The poster made effective use of visual aids	5 points
Poster Mechanics: (_____ / 20 points)	
The poster was grammatically correct	5 points
The poster used appropriate word wrapping (no split words)	5 points
The poster contained references where appropriate	5 points
References were linked to information on the poster	5 points

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; education and assessment in project-based engineering design courses; and numerical modeling of micro scale surface phenomena. From 2008 to 2011, she 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 BS, MS, and Ph.D. from the Massachusetts Institute of Technology, Department of Mechanical Engineering.

Line Harder 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 also has an interest in educational research and is involved in various projects related to teaching and learning assessment at The Technical University of Denmark. She earned her MS and Ph.D. from the Technical University of Denmark, Department of Informatics and Mathematical Modeling.

Beunguk Ahn is an undergraduate student in the Department of Computer Science at the Korea Advanced Institute of Science and Technology. He is engaged in computer science research related to web content analysis, databases, and data mining. He is also interested in software engineering that integrates values from the humanities and social sciences with computer science. From 2008 to 2011, Ahn served as a teaching assistant and consultant for the KAIST Freshman Design Course. During this time, he helped to set up and run the university's Moodle e-learning system and developed custom capabilities for the freshman design course. He received an award for enhancing education at KAIST from the university in 2010 and a special award for dedicated service to the KAIST Freshman Design Course in 2011 in recognition for this work.