Subjective Evaluation in Engineering Design

Richard Bannerot
Department of Mechanical Engineering
University of Houston

Abstract

Engineering students may not receive sufficient experience and help in making and understanding subjective decisions. Two studies are presented that imply engineering students have poor subjective evaluation skills. The artifacts resulting from two design studies were evaluated by two different classes of sophomore engineering students and their instructor. One study represented innovative design, and the artifact was designed and fabricated by teams composed of four students each. The other study represented routine design and was addressed by individual students working alone. Both the instructor and students, acting individually, evaluated both the team and individually produced design artifacts. Compared with the instructor's evaluations of the artifacts, the students undervalued the best artifacts and greatly overvalued the worst artifacts. Most demonstrated very little discrimination in their evaluations. However, the better designers were also better "graders" in that their evaluations more closely followed those of the instructor. Self-evaluations followed the same trends with many of the poorest designers grossly overvaluing their own artifacts. The results indicated that designing ability correlates positively with critical evaluation skills, and it is proposed that the teaching of such skills should be included in design courses.

Introduction

Evaluation is an important part of the design process. Most engineering design textbooks (see for example references 1-6) devote considerable attention to the evaluation process. Much of this attention is directed to concept evaluation (i.e., selection), but evaluation procedures are also provided for comparing performance and other measures. However, most evaluation is at some level subjective. In engineering design the subjectivity of decisions involving a large number of parameters is usually reduced by subdividing the "big decision" into many "smaller decisions", e.g., Pugh concept selection process or criteria functions. In theory these subdividing processes are good, but they do not remove the subjectivity; they simply move it to another level. Engineering design textbooks normally do not provided much help for the student in making these subjective decisions except for case study examples where students may get a "feel" for the small decisions.

It is a concern that engineering student may not receive sufficient experience and help in making and understanding subjective decisions that motivated the work presented in this paper. Two studies are presented. In the first study, the individual evaluations of team produced artifacts in what could be classified as innovative design[7] are "evaluated." In the second study, individual evaluations of individually produced artifacts in what could be classified as routine design[7] are "evaluated." In both studies the students were given explicit instructions on how the ratings were to be linked with various attributes of the artifacts. They were even given instructions on the desired "grade" distribution. The student evaluations were compared to those developed by the instructor. Details of the results are presented in the paper. Both studies produced similar conclusions.

Evaluation of Team Produced Artifacts in an Innovative Design Environment

An experiment was conducted in fall 2002 in the sophomore design class in the Department of Mechanical Engineering at the University of Houston to assess the skills of young engineering students in evaluating the artifacts that resulted from the major class project for the semester. The 41 students self-selected into 13 teams. All teams had four members except for one three-member team. These artifacts resulted from a two-month long, team project: design, build and test. This project was the major component of the course grade (50%). The "value" (or quality) of the artifact itself represented 20% of the project grade, i.e., 10% of the entire course grade. The remaining 80% of the project grade was provided from various reporting requirements (45% for two progress reports, a final report and an oral presentation), and performance measures (35% for the initial testing and a final competitive testing). The problem statement for the fall 2002 class is given in Fig. 1. The students were asked to evaluate all the artifacts themselves. They were given the instructions provided in Fig. 2. The artifacts were on display in the design workshop, and students in groups of eight were given about fifteen minutes to view and record their evaluations of the designs including their own. The students had also viewed all the designs in operation during the testing phases of the project. All the submissions were tabulated and compared to the instructor's. The summary of that comparison is shown in Table 1.

The artifacts are ordered from "best" (in the instructor's judgment) to "poor" in Table 1 as indicated by the numbers from 104 to 32 in the second column that can be referred to the rating system described in Figure 2. Note that these grades are completely subjective, and only reflect the testing (figure of merit) to the extent that the evaluator chooses. The 104 grade was awarded for one team's extraordinary success in achieving the goals of the project (routinely depositing all ten ping pong ball in 1.2 second intervals utilizing only gravity) and its use of special fabrication techniques. The third column is the grade that each team gave itself, i.e., self evaluation of the team determined by averaging the individual evaluations of the team members. The fourth column, presented as a per cent, represents the amount that the group "overrated" itself compared to the instructor's grade. It was determined by subtracting the instructor's grade (column 2) from the team's self evaluation grade (column 3) and dividing by the instructor's grade. The trend is very clear that the better teams are definitely better critics of their own (and presumably others) work than the poorly performing teams.

Each group will design and fabricate a structure and all the auxiliary systems (hereafter called the "device") to transport sequentially at least two, but up to ten, ping pong balls to a height of at least twelve inches before depositing them in one second or longer intervals into a provided container, within 30 seconds in a limited space and without operator intervention. The device will operate on top of a table provided by the instructor. The device shall weigh less than five pounds and shall be deployable from a six-sided container with each edge length less than 19.0 inches (corresponding to the cube of volume 4.00 cubic feet). Both the device and the balls shall remain at all times within the space above the plane of the top of the table and not higher than 24 inches above the table, i.e., within the parallelepiped whose base is the table top and whose height is 24 inches. Gravity is the preferred prime mover but any source of energy may be utilized as long as it is self-contained (e.g., batteries) and counted as part of the device. The above requirements must be fulfilled for the device to be "successful", but the goal is to maximize the figure of merit, FM, defined as

$$FM = (30 - \sigma) + 2\beta - 4\lambda + 3(5 - \mu) + 2(4 - \delta)$$

where

- σ is the total run time in seconds ($\sigma \leq 30.0$).
- β is the number of balls successfully transferred ($2 \leq \beta \leq 10$).
- λ is the number of balls "dropped" during the run (not in the final or initial container or the device at the end of the run) ($\lambda \leq 8$).
- μ is the weight of the device in pounds ($\mu \leq 5.0$).
- δ is the volume (in cubic feet) of the container (actually the cube of its longest edge dimension) ($\delta \leq 4.0$).

Figure 1: Problem Statement for Innovative Design Problem

grade all four categories on the basis of 0 to 100+:
100 =A; 75= B; 50 = C; 25 = D; 0 = F

	85-100	**65-85**	**40-65**	**15-40**	**below 15**
appearance:	looks fantastic, well engineered, and robust	looks good; probably works most of the time	looks ok and probably works half the time	unacceptable appearance probably doesn't work	failure
quality of idea: preference to those that move ten balls with gravity only, less for fewer balls and less gravity	one of the best	good idea; but I like a couple others better	average idea	could do much better	poor
execution of idea:	great design and exceptional craftsmanship	some good ideas and well made	average ideas but could have been better	below average design	poor
overall score:	one of best two or three in the class	good work	average results	below expectations	unacceptable

Figure 2: Grading Instruction for Team Innovative Design

Table 1: Instructor and Student Evaluations for Team Project

team no.	instructor's grade	team self grade	over est	class given grade	class grade std dev	class grade range	class over est
1	104	96	-8%	96	6	73 - 100	-8%
2	99	92	-7%	90	8	74 - 98	-9%
3	98	93	-5%	91	8	58 - 100	-7%
4	97	95	-2%	90	5	72 - 95	-7%
5	96	92	-4%	91	5	73 - 95	-5%
6	93	91	-2%	93	5	80 - 100	0
7	85	90	6%	86	7	71 - 89	1%
8	83	90	8%	79	10	50 - 91	5%
9	66	85	29%	80	10	55 - 90	21%
10	47	89	89%	60	20	15 - 90	28%
11	43	90	109%	79	10	45 - 95	84%
12	40	87	118%	61	19	15 - 90	52%
13	32	69	116%	65	17	32 - 91	106%

Average Student Assigned Grade: 82.2 (65.8 to 93.1)
with a standard deviation of 11.5 (3.0 to 25.6)
Instructor's Average Grade: 75.6 with a std div of 26.3

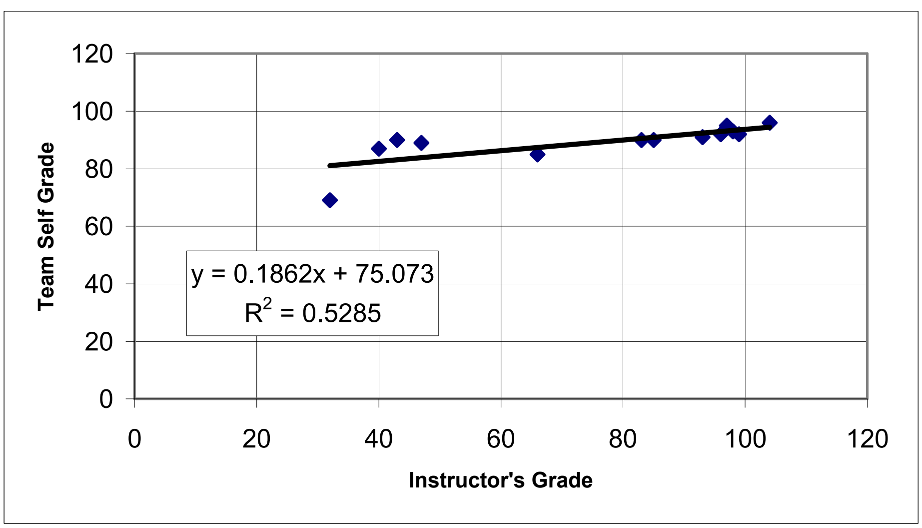

Figure 3: Plot of Each Team's Self Grade as a Function of the Instructor's Grade

The fifth column is the average grade for each artifact given by all the students except those students on the team being evaluated; the standard deviation for those grades is listed in column six, and the range of these grades, in column seven. The “over estimate” (as defined above) for each team’s self evaluation compared to the class determined grade is given in column eight. As noted below the figure, the class determined average grade for all thirteen artifacts was 82 with a standard deviation of 11.5 which was not much higher that the instructor’s average grade of 76, but the standard deviation for the instructor’s grades was 26.3, greater than that of any one student. (The most discriminating student posted a standard deviation for his grading of 25.6.) This averaging out of performance is also noticeable by examining the difference between columns three and five and column two.

Figure 3 depicts the team self grades as a function of the instructor’s grades. A perfect correlation between these grades would results in a slope of 1.0 for the curve fit of the data. As seen in the curve fitting equation on the figure, the slope is actually 0.18. Points to the right of the line of slope 1.0 through the origin correspond to teams that self graded their artifact less favorably than the instructor, i.e., teams 1 through 6 as illustrated in Table 3. The extremely shallow slope of the correlation (nearly horizontal) illustrates the lack of quality discrimination of the class as a whole.

In summary, these students assigned grades that are more “conservative” than the instructor did, assigning fewer and lower “high” grades and no “low” grades. In short, the artifacts were graded without much discrimination, and most were simply graded “above average” by the students.

Evaluation of Individual Artifacts Produced in a Variant Design Environment

In the same sophomore design class, but in the spring 2003, an individual (everyone worked by him/herself) design project was assigned. The general problem statement read: “Design, fabricate and test a device that will use the sun to determine <u>local</u> time in Houston, Texas, between February 6 and 8 as accurately as possible”. Students were encouraged to visit various websites or to use other resources to learn about the history, design and construction of such devices. Specifically they were instructed to design and construct two devices: a vertically mounted device (for a south-facing surface) and a portable device for use on a horizontal surface, that use the sun for determining the local time. Both devices were to read the time <u>directly</u>, i.e., without any “correction,” from the position of the sun’s shadow at times between 9 AM February 6th and 5 PM February 8th in Houston, Texas. Detailed instructions followed concerning the testing procedure and the alternatives for non-sunny days. The performances of the artifacts were to count for 30% of the project grade, and the process to evaluate them was spelled out, i.e., the time accuracy required for given point totals. Fifty per cent of the grade was assigned to a technical report. Twenty per cent of the grade was to be assigned by the instructor based on his evaluation of the artifact.

As in the previous case, all the artifacts were placed on display in the workshop and members of the class were assigned to evaluated them according to a scheme provide by the instructor and provided in Figure 4. There were thirty-nine students in the class. Thirty-seven submitted artifacts on time for evaluation and thirty-five students participated in the evaluation process. All students were requested to evaluate all artifacts, but eleven failed to provide their self-evaluation grade. The tabulation of the results for this part of the study is found in Table 2. Table 2 is laid out exactly as Table 1 except for the content of column 8, and the artifacts are listed in descending order of quality as determined by the instructor's grading. The instructor's grades are listed in column 2. The student's self grade of his/her own artifact is listed in column 3. (As noted above eleven students declined to evaluate themselves.) The student's inflated opinion of his/her artifact is represented by the per cent over the instructor's grade listed in column 4. The evaluations by the entire class (less the producer of the specific artifact) are given in column 5 as well as the standard deviation for the distribution of the grades. The range of the grades is given in column 7. The evaluations of the instructor and the class are

	90-100	80-90	70-80	60-70	below 60
Overall Impression What was your first impression? Were you interested?	looks fantastic, it really got my interest	looks good	goods like the others	not as impressive as most	inferior; not much effort
Creativity Wow! That's a great idea.	really different, unexpected a very good idea	different from the rest; a good idea	different but the idea didn't work out	looks like the rest	looks worse than the rest
Engineering Does it look like it will perform as intended over and over. It is this good quality?	robust and accurate; expect it to provide correct time to a few minutes	not so robust; probably not as accurate as the top five	a little flimsy; expect reading to the nearest half hour	alligned improperly; too flimsy to be trusted	very poorly engineeered
Execution of idea: Did designer spend the time to make it accurate and to make it look good?	great design and exceptional craftsmanship	some good ideas there; well made	ok, but could have improved on appearance	below average design	poor
overall score: not necessarily an average of your four scores, but an final assessment	one of best five in the class, very impressive	good work; impressive; one of the next five to ten best	average results; as good as most	below expectations	unacceptable

Figure 4: Grading Instruction for Individual Project Evaluation

Table 2: Instructor and Student Evaluations for the Individual Project

Artifact no.	instructor grade	self eval	per cent self over instructor	class grade	std dev	class range	per cent class over instructor	Artifact no.
35	103			95	4	85-100	-8%	35
29	101	97	-4%	94	7	75-100	-7%	29
3	98	100	2%	92	7	80-100	-6%	3
34	91			87	6	73-100	-4%	34
18	89	96	8%	90	6	80-100	1%	18
14	85	100	18%	89	7	75-100	5%	14
27	81	93	15%	90	5	80-100	11%	27
37	81	89	10%	89	6	70-98	10%	37
16	79	85	8%	87	7	60-95	10%	16
28	79			86	7	75-95	9%	28
2	77			90	5	80-95	17%	2
5	76	90	18%	87	7	65-95	14%	5
23	75	80	7%	85	6	70-95	13%	23
25	75			85	6	75-95	13%	25
9	74	85	15%	84	8	70-100	14%	9
20	74			85	7	60-100	15%	20
7	72	100	39%	88	6	75-100	22%	7
10	71			86	8	70-98	21%	10
12	69	100	45%	86	8	70-100	25%	12
8	69	95	38%	87	6	70-99	26%	8
4	67			85	7	65-100	27%	4
30	65	85	31%	81	5	70-90	25%	30
6	64			84	7	60-95	31%	6
13	64	90	41%	84	6	70-90	31%	13
19	64			84	7	70-100	31%	19
32	62	90	45%	81	7	60-100	31%	32
15	59			87	7	70-97	47%	15
17	54	80	48%	84	8	60-100	56%	17
33	52	93	79%	83	6	72-92	60%	33
11	51	95	86%	86	8	60-95	69%	11
36	51	80	57%	83	6	70-95	63%	36
22	50	65	30%	81	8	65-95	62%	22
1	48	100	108%	83	8	70-95	73%	1
21	43	93	116%	82	10	50-95	91%	21
24	37	95	157%	82	10	55-100	122%	24
31	35	95	171%	83	7	70-95	137%	31
26	31	100	223%	82	9	60-100	165%	26
AVG	**68**	**91**	**34**	**80**	**7**			
std dev	**17.8**	**8.4**		**3.5**				

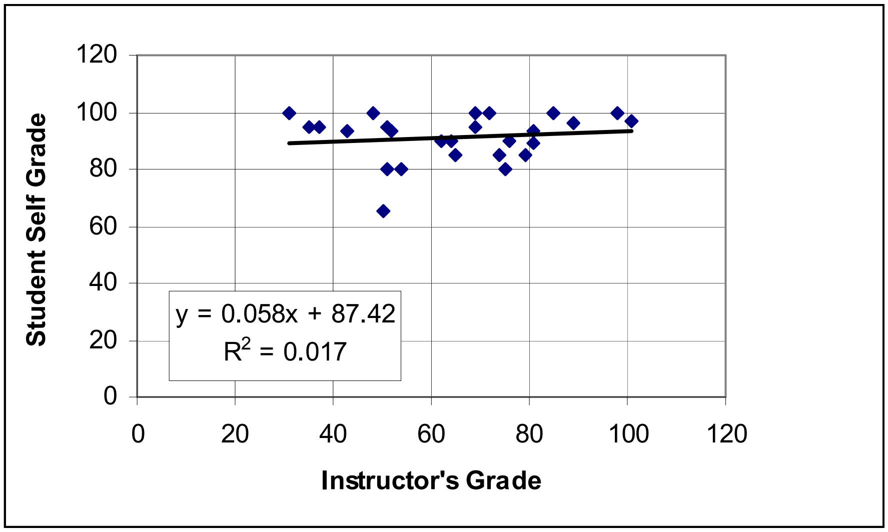

Figure 5: Plot of Each Student's Self Grade as a Function of the Instructor's Grade

compared in column 8 where the over evaluations of the class relative to the instructor are listed as a per cent.

Figure 5 depicts the individual self grades for the routine design project. The self evaluated grade for the student's own artifact is plotted against the instructor's grade for the same artifact. The figure should be interpreted just as Figure 4 was. It is seen that the correlation with the instructor's grades is even less than for the team self evaluations as the slope of the curve fit is only about one-third of the slope found in Figure 4. The student evaluations are essentially random. Only the fabricator of artifact 29 evaluated his/her device lower than the instructor.

Discussion

In summary, students tend to grade more "conservatively" than the instructor, giving fewer and lower "high" grades and many fewer and much higher "low" grades. In fact many of the student grades bear little relationship with the quality of the artifacts as defined by the instructor. These results also provide evidence that the better designers are also the better self-critics. It is one thing to be unable to design well, but the apparent lack of appreciation of the differences between good and bad design is disappointing. This result seems to indicate that teaching only the mechanics of design may be for naught for students who lack the basic skills of critical evaluation.

The first grading scale (Figure 2) was unconventional, and some may comment that students are reluctant to grade each other below, say 60 which is the conventional failure point. However, the instructions were clear and verbally clarified in class. Also, a more convention grading scale was used in the second project (Figure 4), and the results (The slope of the correlation curve in Figure 5 was even less than in Figure 3.) supported the author's claim more strongly.

A second comment may be that all evaluations were compared only to that of the instructor and not to that of a panel. In retrospect perhaps additional credibility could be gained with a better established baseline. On the other hand there was no other willing evaluator, and this issue is not really about the ability to distinguish, for example, between the ninth and tenth best designs, nor to rate the designs on an absolute scale, but rather only to rank order the designs. The reader can be assured that there was no similarity in quality between the best and worst design. Further, few engineering faculty are likely to assign all As and Bs in their sophomore classes as the students were willing to do.

Conclusions

The conclusions are:

- Students and teams undervalue the best designs and greatly overvalue the worse designs which results in a grouping of "grades" from B to A for projects that the instructor rated from F- to A+.
- Self-evaluations (both for the teams and for the individuals) follow the same trend with many of poorest designers (or teams) grossly overvaluing their own artifacts.

In most cases these engineering students were unable to make the decisions to separate the good or poor designs from the average designs. Perhaps this result is not surprising since engineering students generally are not exposed to subjective decision making and critical evaluation, e.g., unlike a visual arts student. Since engineering students are expected to be able to make these judgments, perhaps we should provide more instruction and insight. In fact, as suggested in a previous paper[8] engineering students might benefit by interactions with design students and faculty in other design-based disciplines that tend to be more subjective in nature, e.g., interior design, visual arts, and architecture.

Recommendations

Engineering students (and perhaps design faculty) would benefit from an exposure to the design "process" in the visual arts. The suggestion is that it could be very beneficial to have artists discuss and demonstrate design in early engineering design courses. It would be even more interesting to have early engineering and art design classes meet together. As noted in a previous paper[9], there are projects that engineering and art students can work on together to the benefit of both groups.

References

1. George E. Dieter, Engineering Design, *A Materials and Processing Approach*, 3rd ed. McGraw-Hill Companies, New York, 2000, pp. 162, 182-195.

2. Clive L. Dym and Patrick Little, *Engineering Design: A Project-Based Introduction*, John Wiley and Son, Inc, New York, 2000, 150-154.
3. Barry Hyman, *Fundamentals of Engineering Design*, Prentice Hall, Upper Saddle River, NJ, 1998, pp. 322-331
4. Kevin Otto and Kristin Wood, *Product Design: Techniques in Reverse Engineering and New Product Development,* Prentice Hall, Upper Saddle River, NJ, 2001, pp. 336-338, 477-534.
5. David G. Ullman, *The Mechanical Design Process*, 2nd ed, The McGraw-Hill Companies, Inc. New York, 1997, pp. 152-165, 212-241, 243-283.
6. Karl T. Ulrich and Steven D. Eppinger, *Product Design and Development*, McGraw-Hill Companies, New York, 1995, pp. 106-127.
7. Clive L Dym, *Engineering Design: A Synthesis of Views*, Cambridge University Press, 1994, pp. 44.
8. Richard Bannerot, "It May be Engineering Design, but Is It Design?" Proceedings of the 2003 Annual Conference of the ASEE Gulf Southwest Section, March 19-21, 2003, Arlington, TX on CD.
9. Angela Patton and Richard Bannerot, "Chindogu" a Problem Solving Strategy for Transforming Uselessness into Fearlessness," Paper I A 2, *Proceedings of the 2002 ASEE Gulf-Southwest Annual Conference*, The University of Louisiana at Lafayette, Lafayette, LA March 20-22, 2002, available on the conference CD

Biography

RICHARD BANNEROT

Richard Bannerot is a professor in the Department of Mechanical Engineering at the University of Houston. His research interests are in the thermal sciences and in engineering design education. For the past thirteen years he has taught the required "Introduction to Design" course at the sophomore level and has recently become involved in teaching the college's capstone design course. He is a registered professional engineer in the state of Texas.