Title:
Assessing Student Design Work using the Intelligence of the Crowd

Authors:
Les A. Piegl, lespiegl@mail.usf.edu, University of South Florida
Zachariah J. Beasley, zjb@mail.usf.edu, University of South Florida
Paul Rosen, prosen@usf.edu, University of South Florida

Keywords:
CAD education, assessment, crowdsourcing, peer review

DOI: 10.14733/cadconfP.2019.117-121

Introduction:

The first author has been offering a design class to upper-level undergraduate as well as graduate students for well over a decade. The design term project is vaguely defined so that it gives the students the maximum freedom for exploration of a fairly large space of potential solutions. As design is a form of self-expression, we are keenly interested in seeing personalities woven into each project. Two typical outcomes are shown in Figure 1. These two examples provide a glimpse into the vast array of products the students produce each semester. Some are functionally oriented, yet others are artistic with no usefulness in mind whatsoever. One of the challenges we faced over the past decade is how to assess these design works for a grade that is a fair reflection of the student's ability to complete design tasks with various requirements as well as constraints. Multiple answer quizzes and tests involving numerical results or requiring the regurgitation of facts are easy to grade; however, open-ended, highly creative works reflecting different personalities are significantly more difficult to assess. The added wrinkle is that we are promoting highly human elements such as co-creativity, brainstorming as well as social intelligence.

The idea we wanted to investigate is to use the intelligence of the crowd, instead of relying exclusively on the professor and the teaching assistant. To make this work, we needed to answer a few important questions: (1) how to use the intelligence of the crowd for assessment, (2) what should be the relevant education model, (3) how to aggregate the opinion of the crowd, and (4) how to convert the crowd opinion to a grade. In this paper, we answer these questions in some detail.

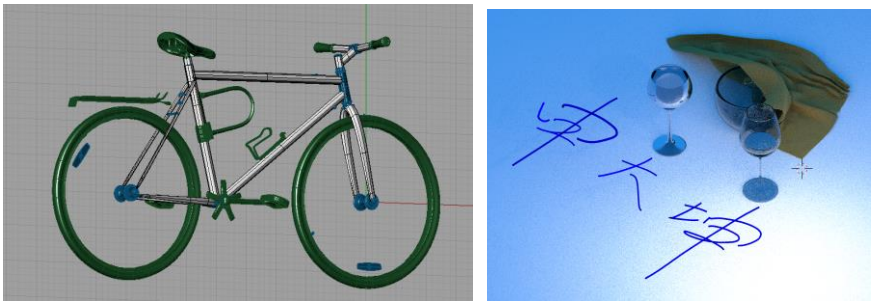

Fig. 1: Typical design project outcomes.

The intelligence of the crowd:
Sourcing a task to a reasonably large crowd is termed "crowdsourcing" or "group sourcing." It has been around for quite some time; however, it has not gained much attention until recently [2]. Observations have confirmed that, when appropriately used, groups tend to be smarter than the smartest person in the group. According to Pentland's research [1], the pattern of communication among people is as significant as intelligence, personality, skill, and achievement combined! In other words, what matters is not what happens between our ears, but rather, what happens between people.

So why should we turn to the crowd when we have the professor, the expert? Unfortunately, experts tend to be poor decision makers: they tend to think alike, they are overconfident and they perform quite poorly in group settings (think of the outcomes of faculty meetings).

In general, individuals are also poor decision makers: they are emotional, they are biased by social norms, they do not handle risks well, and they have short-term vision.

We also need to establish the fact that learning is a social process: we continuously learn from one another as we observe the behavior of others. Learning is practiced in a social context; i.e., it is done in a team using co-creativity, brainstorming, and teamwork.

If experts have a dismal record, humans are poor decision makers, learning requires a social context, and the group as a whole is more intelligent than the smartest of its members, then perhaps we should ask the group to do the assessment. This is precisely what we have done, and the results were amazingly positive.

There are a few considerations to make crowd assessment work: (1) there must be diversity among the members, (2) people in the group should not be biased by the opinions of others, (3) individuals should not be under the thumb of the professor, i.e., opinions should come from local knowledge, (4) the individual judgements must be aggregated, i.e., turned into a collective decision, and (5) decisions should be made simultaneously to avoid influence one over the other.

Finally, we needed to choose the crowd. For assessment of student work, the members of the crowd need to be somewhat knowledgeable of the subject, so we picked the students from the class where the design course was offered. We could have chosen a collection of experts, potential employers, or design enthusiasts. We have no data on how the system would work outside the classroom; however, this is an aspect of the work we intend to investigate in the future.

The educational model:
To successfully assess highly creative design work using the crowd, we needed a new educational paradigm that relied on the innate ability of students to explore potential solutions in an organic, non-structured and non-standardized, manner. The model is shown in two parts in Figure 2 [3]. The left part moves the students away from the traditional cover-test-sort paradigm and puts them in the driving seat to decide what to learn and how to use it. First, we assume minimal background knowledge and encourage the discovery of information on an as-needed basis.

Second, the newly found information needs to be analyzed and used for the primary purpose: problem-solving. The figure demonstrates that a problem-solving fueled exploration is a highly iterative process that may require the update of one's background, the discovery of new information, a better analysis, and perhaps the design of a different solution. The cycle can continue indefinitely and may never terminate, i.e., a design may never be completed; only the best solutions are accepted.

The exploration part of the model is shown on the right of Figure 2. First, the boundary of the space to be explored is fuzzy; it is intentionally vaguely defined. Second, the exploration is done organically, i.e., a random walk fills the space and the level of randomness, and the space-filling property of the walk determines how much the student learns and how much time is spent on filling the space till a sufficiently broad coverage is obtained.

A critical aspect of random space exploration is a set of constraints. In nature, each plant grows randomly, i.e., no two plants of the same type are identical but instead grow within certain constraints. For example, no two carrots are the same. However, carrot seeds turn into carrots, not potatoes. We simulate this constraint randomness by placing soft, hard, partial, and full constraints inside the domain to be explored. A soft constraint is a hint or a suggestion; a hard constraint is a well-defined

entity that needs to be satisfied, e.g., dimensions or weights, a full constraint must be included in the solution, whereas only a part of the partial constraint needs to be satisfied. For a design task, these constraints have a dual purpose: (1) they are required characteristics of the final product, and (2) they guide the organic space exploration.

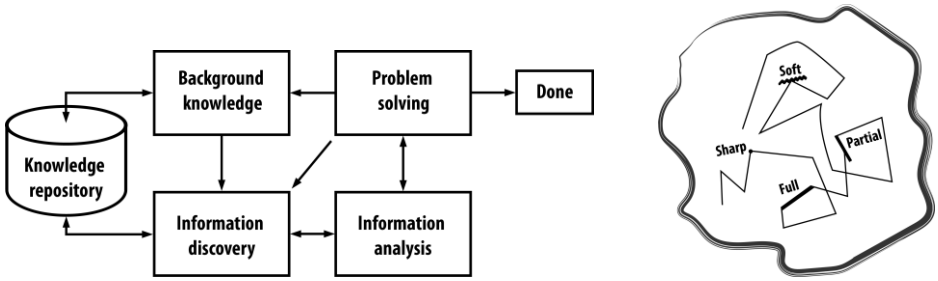

Fig. 2: The organic learning model used in this work.

Our design term project uses both models shown in Figure 2. For example, we do not teach students how to use a modeling system. They fill the gap in their background knowledge by reaching into the knowledge repository, e.g., educational videos on how to use Blender, to learn what is needed to complete the project. The space of exploration is vaguely defined, e.g., design any household item, and we sprinkle in a few constraints such as minimum functionality or appeal to a certain customer group. Then we turn the students loose, i.e., we let the natural born creativity in each student go to work.

The assessment process:

The most challenging task in our research was the design of an intelligent review form and the interpretation of the results coming from the answers. We did not know what kind of questions to ask, how many of them to put in the form, how to group them, and how to weight the answers. The guiding principle was this: when in doubt, ask the crowd, i.e., we crowdsourced the entire process using a seed-growing algorithm, intelligent data combing, and topological text reduction. This is how it worked.

First, we started with a straightforward review form that asked the students to share with us their overall impression about the work, tell us what they thought about comprehensibility, originality, significance, etc., and, most importantly, use free-form comments to give the work a thorough evaluation. Then we combed through all this information and identified relevant words and phrases that could be used in the next iteration of the review form. Below are two examples of how we reduced the student feedback to words and phrases.

*"This presentation was very well done. The presenters **understood the material,** and that was shown in their delivery. The organization of the content was such that it **promoted engagement** and **triggered discussion**. It was **technically accurate** and provided a **plethora of resources** to be used in the development of the final project. To me, this presentation marks an important milestone (with regards to the information it covers), and I am glad that the presentation enabled a clear understanding of the material."*

***"Clear, educational, and straight to the point**. This presentation was easy to understand and presented its information in such a way that it trimmed a lot of wordiness that other presenters had. However, their presentation **lacked a lot of uniqueness** - the slides were all the same style, with the same layout and minimal variation. It works well as an outline to teach the topic but doesn't inspire me to want to read the whole thing, just skim. **The organization is spot-on**, however, and the group planned their sections well."*

Using words, phrases, and partial sentences, we then established the next round of questions. Some examples are shown below.

*"They responded well to the questions they received." (**QUESTIONS**)*

*"The presenters seemed knowledgeable of the topic." (**PREPARATION**)*

*"Their presentation was informative and created good thought-provoking questions that triggered some good conversation in class." (**IDEA FLOW**)*

*"The presenters did a good job of communicating the information clearly." **(COMMUNICATION)***

*"Your essay made good use of cited sources for information ..." **(REFERENCES)***

*"I also liked that the team split the presentation up and were each able to speak on the topics, it shows that they worked as a team and put effort into this presentation helping each other learn the material so they could teach us." **(TEAMWORK)***

*"I liked that it was interactive and I got to ask questions." **(INTERACTION)***

*"… had real-life experience in the subject" **(PRACTICALITY)***

The key words above formed the new review form with three choices for each question. For example, to check PREPARATION, we gave the students three options: (1) very knowledgeable, (2) somewhat prepared, and (3) unprepared.

This new review form replaced the old one, and we collected a new set of detailed comments that we processed precisely the same way as before. If we discovered that the students were looking for something relevant, it was added as a new question and the iteration continued. After about 3-4 iterations, no further questions emerged; it appeared that the process has converged and we got the answers to our opening questions: what are the questions and how many are there.

Next, we had to group the questions and weight each possible answer. The grouping came somewhat naturally, and it formed four groups in our review form:

1. G_1 overall impression: outstanding, very good, good, acceptable, fair, marginal, unacceptable.
2. G_2 technical details as in soundness, balance, idea flow, etc.
3. G_3 personality, e.g., curiosity, uniqueness, and creativity.
4. G_4 detailed and free-form comments

The second important question was how to weight each answer to every question. For example, in the PREPARATION question above one could use 1.0, 0.5 or 0.0 for *very knowledgeable*, *somewhat prepared* and *unprepared*. Or it could be 1.0, 0.4 and 0.0. The *somewhat prepared* answer needs to be put in the proper context based on what is being measured in what subject.

Once the weights are established, the partial grades $G_1,\cdots,G_3$ need to be computed. While there are many statistical methods to massage this data, we found that using the mean produced perfectly acceptable results (we tried to calculate the final results using many different methods, but the difference was negligible). The final score was established by the following formula: $G = w_1G_1 + w_2G_2 + w_3G_3 \pm f(G_4)$ where the weights $w_1,\cdots,w_3$ are set based on the level of importance in each category. For example, if the overall impression is what we are after, then the weights (0.5, 0.25, 0.25) may be used. If the technical content is the most important, then perhaps (0.2, 0.6, 0.2) could be

applied. What we found over several years of using this system is that once the student sentiments are correctly captured, the weight does not matter much, unless we use extreme values.

We are left with dealing with the last part G_4, the score given to the detailed comments. This is an on-going area of research that does not yet have any acceptable results due to either the absence of meaningful comments (e.g., "great job!") or comments that conflict the technical part of the form (e.g., a poor evaluation may be given to a technically brilliant work if the student reads from the slides). So, at this point, we use G_4 for validation as well as adjustment.

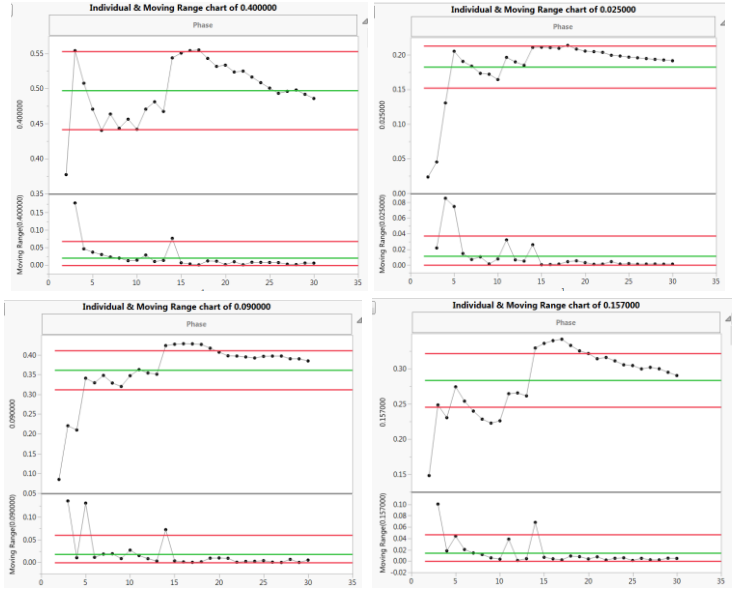

Fig. 3: Standard deviation as a function of the number of reviews.

The last question that we need to answer: how large should the crowd be? Do we need a lot of students or only a few (3-4) will suffice? According to our research, you need more than a handful of 3-4. We considered any score statistically relevant if the standard deviation was less than 10%. A score with a standard deviation of less than 5% was considered very good but was quite rare. Figure 3 shows the standard deviation as a function of the number of reviews used for overall (top left), technical (top right), personal (bottom left) and final (bottom right) scores. In our classes of 40+ students, after roughly 20 reviews the standard deviation did not move much.

Conclusions:

A crowd-sourced based assessment system has been outlined in this paper. The system is based on an iterative method to find relevant questions and every part, ranging from finding the questions to completing the actual assessment, has been sourced to the crowd, the students. Comparing the crowd-based grades to ours, the system seems embarrassingly accurate, beating our combined 40+ years of grading experience.

References:

[1] Pentland, A.: A New Science of Building Great Teams, Harvard Business Review, April 2012.
[2] Surowiecki, J.: The Wisdom of Crowds, Anchor Books, New York, NY, 2005.
[3] Cummings, M. L.: Learning in the 21st Century: Principles, Models, Environments, U-turn Press, 2019.