
AC 2012-4614: TOWARDS THE DEVELOPMENT OF AN OBJECTIVE ASSESSMENT TECHNIQUE FOR USE IN ENGINEERING DESIGN EDUCATION

Dr. Scarlett R. Miller, Pennsylvania State University, University Park

Scarlett Miller is an Assistant Professor of engineering design and industrial engineering at the Pennsylvania State University where she holds the James F. Will Career Development Professorship. She received her Ph.D. in industrial engineering from the University of Illinois and her M.S. and B.S. in industrial engineering from the University of Nebraska.

Prof. Brian P. Bailey, University of Illinois, Urbana-Champaign

Brian Bailey is an Associate Professor in the Department of Computer Science at the University of Illinois, Urbana-Champaign. His research interests include creativity support tools, attention management systems, and, more generally, improving interactions between people and technology. Bailey received the NSF CAREER award in 2007 for his research in the area of human-computer interaction.

Prof. Alex Kirlik, University of Illinois, Urbana-Champaign

Towards the Development of an Objective Assessment Technique for use in Engineering Design Education

Abstract

Design is now considered to be the central or distinguishing activity of engineering and thus is being widely integrated throughout engineering education. However, engineering programs face special challenges in assessing student design abilities because of the subjective nature of design. Research has shown that current evaluation methods are insufficient in evaluating student design performance because quantitative tests of students design skills are not indicators of students design performance and open-ended project evaluations are muddled by the subjective biases of the single human evaluator. The purpose of this paper is to evaluate a new evaluation metric that utilizes a novel application of the Bayesian Truth Serum (BTS) in an effort to increase the accuracy of design skill evaluations and minimize subjective bias. Through a series of experiments our results indicate that BTS is a better indicator of idea generation ability than traditional evaluation methods. Although more work is needed to validate and understand how BTS can be used as an evaluation technique in engineering design, the results from this study are promising and provide a first step at deriving quantitative measures of student design performance in engineering.

Background

Assessment of student design skill is an important part of engineering education, especially with the ever-increasing role of design in the engineering classroom ^[1-3]. However, engineering programs face special challenges in assessing students design capabilities and providing meaningful feedback because engineering design is mostly subjective in that there are no mathematical proofs or conclusive experiments to grade. In fact, most engineering design course grades are structured around subjective grading of open-ended design projects that require students to harness and *apply* the knowledge he or she has gathered during the semester in the development of their project^[4].

Although open-ended projects are critical for teaching and evaluating design skill, the open-ended nature of these projects makes it difficult to develop objective criterion to assess performance. Because of these difficulties, many design project grades are structured around evaluations performed by the course instructor, a design ‘expert’ ^[5]. One of the main issues with this assessment method is that design tasks do not typically have concrete right or wrong answers and therefore must be *subjectively* graded by the course instructor or teaching assistants. These judges, however, are subject to cognitive biases and limitations based on their own beliefs and expectations. For instance, evaluators can be biased based on the order in which they evaluate the projects (cue primacy and anchoring effects), or their inability to look at the problem from a perspective

other than their own (framing bias) ^[6]. There is also no way to assure the same criterion for evaluation is utilized across all projects evaluated.

Traditional approaches to solving the problem of reducing bias in subjective bias in design evaluations include developing evaluation rubrics ^[7, 8] or using peer-ratings of design skill ^[9]. However, these methods are time consuming and do not truly minimize subjective biases as they still rely wholly on a *single* human input. The problem of assessing design skill is particularly problematic in educational settings where there is an inverse relationship between class size and project evaluation time. In other words, as class sizes increase, instructors have less time to assess individual student design skill and therefore less time to sit and methodically grade design project deliverables. Therefore, new methods are needed that objectively evaluate design competence and scale to growing class sizes. These new methods need not completely replace existing methods of evaluation, but instead provide a more objective evaluation of student design performance that re-weigh or replace single subjective evaluations.

One way to minimize subjective bias in design skill evaluation and reduce the timeliness of grading design projects is to remove the single expert judge and determine new methods for evaluating a student's ability to analyze, create and evaluate design solutions. A study by Bonnardel and Marmèche ^[10] showed that there is a relationship between design expertise and one's ability to identify useful design sources. Their study found that individuals with more design expertise were more able to use a wider variety of sources because they are able to identify the positive and negative aspects of each design. If an individual's ability to judge design examples is in fact related to their design expertise then ratings of example quality could be used as a way to objectively test one's design skill level. This would allow for the creation of new design skill evaluation assessments where individuals' design abilities would no longer be based purely on a single expert evaluation, but instead would be based on an individual's ability to rate existing solutions. However, Bonnardel and Marmèche's ^[10] work only focused on understanding how experts and novices solicited examples and not how ratings of example quality relates to design expertise. Therefore, the purpose of this paper is to understand how ratings of examples can be used to determine one's design skill level in engineering design education through the use of the Bayesian Truth Serum.

The Bayesian Truth Serum

The Bayesian Truth Serum (BTS) is a method for eliciting subjective information when objective truth is intrinsically or practically unknowable ^[11]. This method is used when subjective judgments of multiple respondents are the only evidence available, and majority opinion may be wrong. Since design is subjective by nature, and one can only determine if a design is appropriate or not by majority opinion, the BTS method could provide a way to accurately predict genuine design knowledge and design skill levels

The BTS method consists of an information scoring system that incentivizes truthful answers from a sample of rational (i.e. Bayesian) expected value-maximizing respondents. Unlike earlier approaches to “test theory without an answer key”^[12], the Bayesian Truth Serum does *not* privilege the consensus answer but instead boosts scores to answers that are more common than collectively predicted, with predictions drawn from the same population that generates the answers. These responses are deemed ‘surprisingly common’ and the associated numerical index is called an information score. Those who have demonstrated superior meta-knowledge – knowledge of others opinions – receive higher BTS scores and thus are considered to have more knowledge than others in the sample population. This adjustment in the target criterion which does not privilege the consensus answer removes respondents bias towards answering what they feel is the likely group mean. In other words, truthful responding remains the correct strategy even for someone who is sure his or her response represents a minority view.



Figure 1: Example used in BTS survey

To calculate a BTS score, participants are provided with a series of survey questions and asked to provide personal answers and also a prediction of the percentage of the sample population that they feel will endorse each answer. For example, n participants are provided with questions that have m answers: i.e. Which of the following statements best describes Figure 1 ($m=3$):

- (a) This is an effective example of the visual salience design principle
- (b) This example is not relevant in reference to the visual salience design principle
- (c) This example depicts a violation of the visual salience design principle

Respondents are asked to provide a personal answer (i.e. a-c) as well as a prediction of the proportion of the sample that will endorse each answer: e.g. I think 30% of other respondents will think this it is an effective example, 2% will

say it is not relevant, 68% will think it's a violation. Once the answers and predictions are collected, a BTS score can be calculated.

The calculation of the BTS score involves four steps. First, Let $x_k^r \in \{0, 1\}$ indicate whether respondent r has endorsed answer k and $y = (y_1^r, \dots, y_m^r)$ be her prediction of the sample proportions ($y_k^r \geq 0$, $\sum_k y_k^r = 1$). The BTS algorithm then proceeds in four steps.

Step 1- Calculate the average $\bar{x}^k$ of the endorsement and the geometric mean $\bar{y}^k$ of the predictions:

$$\bar{x}^k = \frac{1}{n} \sum_{r=1}^n x_k^r, \log \bar{y}^k = \frac{1}{n} \sum_{r=1}^n \log y_k^r$$

Step 2- Calculate the BTS score of each individual r :

$$u^r = \text{information score} + \text{prediction score}$$

$$u^r = \sum_{k=1}^m x_k^r \log \frac{\bar{x}^k}{\bar{y}^k} + \sum_{k=1}^m x_k^r \log \frac{y_k^r}{\bar{x}^k}$$

Step 3- For each answer k , calculate the average BTS score $\bar{u}^k$ of all individuals endorsing answer k :

$$\bar{u}^k = \frac{1}{n \bar{x}^k} \sum_{r=1}^n x_k^r u^r$$

Step 4- Select the k that maximizes $\bar{u}^k$

The prediction score calculated in Step 2 scores the accuracy of respondent's predictions of the empirical distribution of answers. It scores how well the respondent's predictions match the empirical frequencies. The best prediction score is zero, attained only when prediction exactly matches reality, $y_k^r = \bar{x}^k$. The information score, on the other hand, looks at respondent's personal answers giving a higher score to those whose answer is 'surprisingly common' or more common than collectively predicted. For example, if an answer is endorsed by 30% of the population ($\bar{x}^k$) but the population predicted only 5% ($\bar{y}^k$) of respondents would select that response then those who selected that answer would receive a higher information score because their response is 'surprisingly common' ($\bar{x}^k > \bar{y}^k$). On the other hand, if the predictions averaged 75% ($\bar{y}^k$) and only 30% of the population ($\bar{x}^k$) selected the answer, the answer would be deemed surprisingly uncommon and those who selected it would receive a lower information score ($\bar{x}^k < \bar{y}^k$). Students that have a higher BTS score after completing the BTS survey would be considered to have a higher expertise than the other survey respondents.

BTS can be used to identify expert respondents if knowledge correlates across multiple questions^[13]. BTS suggests two possible indexes for rating the expertise of a single respondent: the individual and pooled index. The *individual index*, is the BTS score u^r of each individual averaged across all responses while the *pooled index*, is the average BTS score of the *answer* endorsed by respondent r averaged across all questions, $\sum_{k=1}^m x_k^r \bar{u}^k$. The pooled index filters out individual differences in prediction competence (i.e. long run calibration of predictions).

Therefore, if two respondents selected the same answers for all of the BTS survey questions, they would receive the same pooled index score but different individual index scores because the individual index takes into account the individuals prediction score for each survey question while the pooled index does not. The study presented in this article seeks to understand if these indices can predict design skill levels or “expertise” in engineering design.

The Bayesian Truth Serum has been validated both theoretically and empirically and has proven to be a solid way to identify experts when subjective judgments remains the only source of evidence available, and there is a possibility that most people may be wrong ^[11, 13]. Since design is subjective, the Bayesian Truth Serum could provide a way to accurately predict genuine design knowledge and design skill levels.

Experimental Design

In order to understand if BTS can be used to more accurately predict design skill levels in engineering design compared to traditional assessment methods like course grade and exam average an empirical study was completed at a large public institution. Twenty-three junior and twenty-four senior engineering students (47 total) in an upper-level human factors engineering design class were recruited and participated in this study. The study was completed in two phases. In phase 1 a BTS score was tabulated for each participant based on a series of responses to survey questions, which asked participants to rate the design quality of pictorial design examples. They were also asked to predict the empirical distribution of responses of the other participants in each of these questions. In phase 2, participants were asked to develop design ideas for an engineering design problem and select their best idea. This idea was then judged for its ability to solve the design problem by 10 other participants. The details of the experiment are provided below.

Procedure

During Week 1, 47 participants were given the following instructions:

“For the following 30 questions you will be asked to rate design examples based on 10 human factors design principles. We ask that you answer the questions as truthfully as possible, even if you think your opinion represents a minority view (i.e. you think others will answer different than you will). In addition to giving a personal answer, you will also be asked to predict the percentage of other participants in the study currently taking the Human Factors Engineering Design Course that would endorse each answer. Remember to answer to the best of your ability!”

Next, participants were directed to www.surveymoz.com where the survey was conducted and given 60 minutes to complete the BTS survey.

The BTS survey required participants to rate 30 pictorial examples as an effective example of a human factors principle, a violation of a principle, or one in which the principle does not apply based on one of the following human factors principles: color coding, comfort, physical affordances, analogies, population stereotypes, safety, salient feedback, shape coding, stimulus-response compatibility and visual salience. The design principles tested in this study were selected because they are core to the human factors design course curriculum and students were expected to be able to identify problems with existing solutions according to these principles when they complete the course. Once participants had selected their personal response to each survey question, they were then asked to estimate the empirical distribution of participants (their fellow classmates) that would endorse each response. Three pictorial design examples were selected for each of the ten human factors principles listed leading to a total of 30 BTS survey questions, see Figure 1 for example question. The order in which the 30 questions were presented was randomized for each participant.

A week after the BTS assessment, 47 participants were brought back into the lab and were informed of the goals of week 2 of the study and participants were provided with the following instructions:

“For today’s experiment we ask that you do the following generate (and sketch) *as many* design ideas as possible for the design problem described below and write comments on each design such that an outsider could look at your idea and understand the concept being depicted.”

Once participants read the instructions, any questions were answered. Next, participants were given the design task and asked to brainstorm ideas. The following task description was read aloud to participants as they followed along on their own copy:

“There has been an increase in student accidents on campus in recent years from student’s texting and/ or talking on cellphones or listening to music on an MP3 player while walking around campus. While using these devices, students become distracted, and can trip, fall or even run into something. In fact, in 2008 over 1,000 pedestrians visited emergency rooms due to accidents from using these devices while walking. There are reports of concussions, sprained ankles, broken appendages and even fatalities from these accidents. These numbers do not include the countless number of unreported incidents involving walking into something (i.e. a parked car) without an ER visit. This increase in accidents has been substantial on college campuses because of the number of students on campus and the increase of MP3 players and cellular usage (texting and talking), all of which are distracting.

As a Human Factors Engineering designer, you've been tasked with developing AS MANY solutions that would address all or some of the issues associated with this increased accident rate. These ideas can include

either alterations to current technologies or the design of new technologies to help reduce these accidents.”

Participants were then given one hour to generate as many potential solutions to the problem as possible. Once the hour had expired, participants were asked to select, by marking it with a sticker, what they felt was their ‘best idea’.

A week later, the 47 participants were brought back in to the lab and were informed of the goals of week 3 of the study and then provided with the following instructions:

“For this study, you will be asked to evaluate the ideas produced by other participants from the Human Factors Design Class on a scale from 1 (poor human factors design) to 4 (excellent human factors design). In addition to rating the quality of the design in terms of HUMAN FACTORS DESIGN PRINCIPLES, you will also be asked to provide a design critique.”

Participants were then handed a packet of 15 anonymized ideas that other participants had self-selected as their ‘best idea’ at the end of week 2. Each participant received a different set of 15 ideas in a randomized order, and participants did not rate their own idea. Participants were then directed to www.surveygizmo.com where an online survey was developed to help guide them through the critique process. Participants were given 1 hour to complete as many of the 15 idea critiques as possible without compromising their critique quality. For this paper, we will only be discussing the ratings of the idea quality. This concluded the study.

Statistical Analysis

There were two main questions to be addressed in this experiment; Do the BTS indexes of expertise correlate with human factors knowledge as measured in this study, and Can the BTS indexes of expertise better predict one’s ability to generate high quality design ideas over traditional measures such as course grade and exam average?

In order to answer these questions, the pooled and individual index scores calculated from the survey in week 1 were categorized into 3 groups (high, medium and low) for each participant, with the cut-off scores being half a standard deviation away from the mean. An ‘idea score’ score was also computed for each participant based on the peer ratings (1-poor design idea, 4-excellent design idea) for each self selected ‘best idea’ from week 2 where:

$$\text{idea score} = \frac{\sum \text{idea ratings}}{\text{total points possible}} = \frac{\sum \text{idea ratings}}{4 * \text{number of idea ratings}}$$

In order to understand if the BTS indexes of expertise correlated with human factors design knowledge, a partial correlation test (controlling for class standing)

was performed between the indexes of expertise, the benchmarks of conventional wisdom and response accuracy, and the participant's course grade. Similarly, in order to compare the predictive value of BTS indexes of expertise on a student's ability to generate high-quality design ideas, an additional partial correlation was performed (controlling for class standing) between idea score and the following metrics: pooled index, individual index, response accuracy, conventional wisdom, exam average and course grade.

SPSS v. 20 was used for all analysis with a level of significance (α) of 0.05.

Results

Do the BTS indexes of expertise correlate with genuine knowledge?

The first question we sought to address was if the BTS indexes of expertise were related to the current evaluation metrics used in engineering design education as prior studies have shown that the pooled index score is an excellent predictor of genuine knowledge correlating significantly better than the measure of conventional wisdom with response accuracy^[13]. Therefore, we tested the indexes for correlation with the course grade and exam average. As a benchmark, we also tested these measures correlation with conventional wisdom and response accuracy. The results from this study can be seen in Table 1.

Table 1: The results from our correlation test.

Control Variables			Individual Index	Pooled Index	Conventional Wisdom	Response Accuracy
Class standing	Exam Average	Correlation	0.17	0.24	0.02	-0.01
		Significance	.28	0.12	0.86	0.96
		df	47	47	47	47
	Course Grade	Correlation	0.13	0.27	0.02	-0.01
		Significance	0.38	0.07	0.86	0.96
		df	47	47	47	47

Our results indicate that the measures of human factors knowledge (exam average and course grade) did not correlate significantly with the individual index ($p=0.28$ and $p=0.38$, respectively) or the pooled index ($p=0.12$ and $p=0.07$, respectively). However, as can be noted by the results in Table 1, the pooled index was a better predictor of human factors knowledge than the measures of conventional wisdom, response accuracy and the individual index as it was trending towards correlations with the course grade ($p=0.07$). Although more research is needed to understand how the pooled index score relates to genuine knowledge, this finding confirms prior research that showed that the pooled index is a better indicator of genuine knowledge compared to the measures of conventional wisdom and response accuracy.

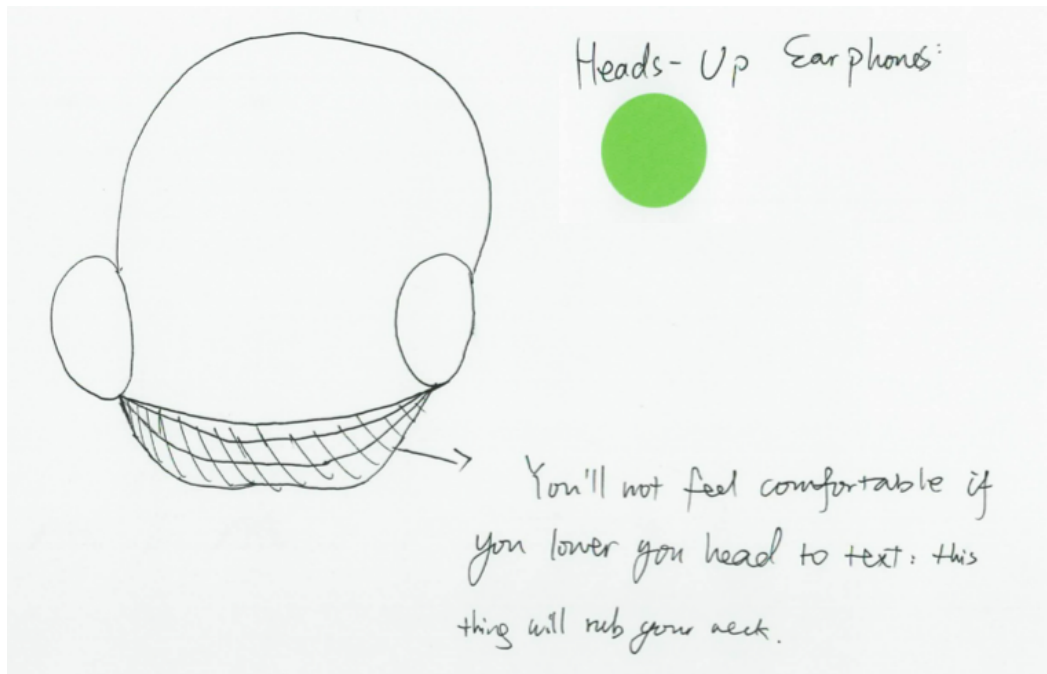


Figure 2: The idea that received the lowest idea score (0.30).

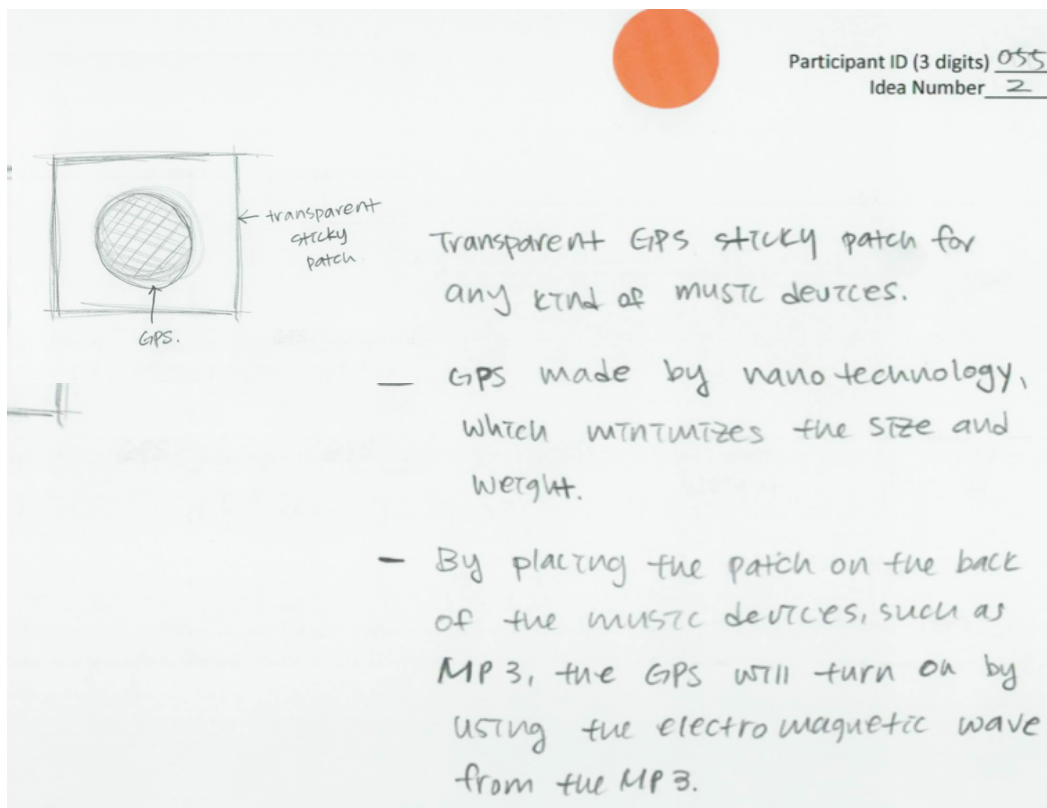


Figure 3: This is the idea that received the highest idea score (0.88).

Do the BTS indexes of expertise correlate with idea quality?

The next question we sought to understand was how well, if at all, the BTS indexes of expertise predicted ones ability to generate highly rated design ideas. In week 2, each participant generated between 1 and 7 ideas (average 3.5). The best idea, self-selected by the participant, was then rated in week 3 by 7 to 10 participants, average of 9.5, on a scale from 1 (poor human factors design) to 4 (excellent human factors design). The idea scores in this study ranged from 0.30 to 0.88 with an average of 0.58. Figure 2 and 3 show the ideas that received the highest and lowest idea scores, respectively. In order to determine if the BTS indexes of expertise could better predict the ability of respondents to *create* highly rated design ideas over traditional measures like course grade and exam average a partial correlation test was performed, see Table 2.

Table 2: Correlation test with the control variable of class standing.

		Exam Average	Final Grade	Conventional Wisdom	Response Accuracy	Individual Index	Pooled Index
Idea Score	Correlation	0.13	0.14	0.25	0.23	0.23	0.35
	Significance	0.40	0.35	0.10	0.14	0.14	0.02
	df	47	47	47	47	47	47

The results showed that the only metric that correlates significantly with the idea score is the pooled index ($r = 0.35$, $p < 0.05$). Not only did the traditional measures of exam average and final grade not correlate with the idea score, but they were the least predictive of the variables tested. This result again indicates that the pooled index is a better measure of design skill than traditional measures.

Discussion and Conclusion

With the ever increasing role of design in engineering education, it is important that we develop new metrics that quickly and accurately measure student competence in design while reducing the load on the course instructor as companies are now expecting graduates to contribute immediately to the design process post graduation^[14-16]. Current methods for assessing design skill levels in engineering education involve course instructors subjective grading of student design projects and objective tests of lower level design skills (i.e. memorization of principles). These methods are inadequate for current evaluation practices because they either require subjective judgments from one individual or only tap into the lower levels of learning and not the application of the principles, which are the skills most instructors aim to evaluate.

The purpose of this paper was to evaluate a new objective assessment techniques ability to determine student competence in design and compare this measure to

traditional course measures like exam averages and final course grades. We utilized a novel application of the Bayesian Truth Serum (BTS), an algorithm developed for financial engineering that has been proven to be a solid way of identifying *experts* when subjective judgments are the only source of evidence available. However, this algorithm had not been tested as an assessment technique and therefore it was unclear if, or how well, this method could be used to evaluate expertise in design.

The results from this experiment show some interesting findings with implications for engineering design skill assessment. First, BTS proved to be a better indicator of student ideation ability compared to the traditional measures of course grade and exam average. Although the exams used in the human factors course did not directly test idea generation ability, they did test the design principles necessary to effectively develop high quality design ideas. The exams in this course were graded by the course instructor and contained mostly multiple choice questions (with a few open ended design questions) like, “What type of stimulus-response compatibility does the following image depict: movement compatibility, congruence, movement proximity, or colocation.” This shows that our current evaluation methods only tap into students’ ability to memorize principles and not their ability to apply their design knowledge. Therefore, the results from this study show promise for new evaluation methods that employ BTS to tap into these deeper levels of learning like the development of ideas that properly employ these design principles. However, this study did not compare the BTS method to subjective evaluations of design projects completed by the course instructor, but future studies should be conducted that explore and compare these methods.

Our results also confirmed prior studies which showed that the pooled index is a better indicator of genuine knowledge compared to traditional measures of conventional wisdom with response accuracy ^[13]. More specifically, our results found that the pooled index was the only variable that correlated significantly with the course grade (although it was only moderately significant). Therefore, the BTS method and the course grade do test some of the same knowledge structures but our results revealed that BTS is able to more accurately identify individual who have attained a higher level of learning and the BTS method requires less input and evaluation of student achievements from the course instructor.

Although more work is needed to validate and understand how BTS can be used as an evaluation technique in engineering design, the results from this study are promising and provide a first step at deriving quantitative measures of student design performance in engineering. It also identifies the need for new quantitative measures of design skill in engineering education that scale to the ever-increasing class size.

Bibliography

1. Kilgore, D., C.J. Atman, K. Yasuhara, T.J. Barker, and A. Morozov, *Considering Context: A Study of First-Year Engineering Design Students*. Journal of Engineering Education, 2007. **96**(4): p. 321-334.
2. Morozov, A., D. Kilgore, and C.J. Atman. *Breadth in Design Problem Scoping: Using Insights from Experts to Investigate Student Processes*. in *Proceedings of the American Society for Engineering Education Annual Conference and Exposition*. 2007. Honolulu, HI.
3. Simon, H.A., *The Sciences fo the Artificial, 3rd Edition*1996, Mass: MIT Press.
4. Davis, D., M. Trevisan, S. Beyerlein, O. Harrison, and P. Thompson, *Engineering Capstone Design Course Assessments: Assessment Framework and Instruments*, 2007, Transferable Integrated Design Engineering Education (TIDEE).
5. Demirbas, O.O. and H. Demirkan, *Learning styles of design students and the relationship of academic performance and gender in design education*. Learning and Instruction, 2007. **17**(3): p. 345-359.
6. Wickens, C.D., J.D. Lee, Y. Liu, and S.E. Gordon Becker, *An Introduction to Human Factors Engineering*2004, Upper Saddle River: Pearson Prentice Hall.
7. Asunda, P.A. and R.B. Hill, *Critical Features of Engineering Design in Technology Education*. Industrial Teacher Education, 2007. **44**(1).
8. Davis, D., S. Beyerlein, P. Thompson, J. McCormack, O. Harrison, M. Trevisan, R. Gerlick, and S. Howe, *Assessing Design and Reflective Practice in Capstone Engineering Design Course*, in *American Society for Engineering Education Annual Conference and Exposition* 2009: Austin, TX.
9. Falchikov, N. and J. Goldfinch, *Student Peer Assessment in Higher Education: A meta-analysis comparing peer and teacher marks*. Review of Educational Research, 2000. **70**(3): p. 287-322.
10. Bonnardel, N. and E. Marmèche, *Evocation Processes by Novice and Expert Designers: Towards Stimulating Analogical Thinking*. Creativity and Innovation Management, 2004. **13**(3): p. 176-186.
11. Prelec, D., *A Bayesian Truth Serum for Subjective Data*, in *Science*2004. p. 462-466.
12. Batchelder, A.K. and A.K. Romney, *Markov chain estimation for test theory without an answer key*. Psychometrika, 1988. **68**(3): p. 373-389.
13. Prelec, D. and H.S. Seung, *An algorithm that finds truth even if most people are wrong*, in *Working Paper*2010, Massachusetts Institute of Technology.
14. May, E. and D.S. Strong, *Is Engineering Education Delivering what Industry Requires*, in *3rd Canadian Design Engineering Conference* 2006: Toronto. p. 204-212.
15. Newcomer, J.L. *A Broadened Perspective: Teaching Engineering Design in a Social Context*. in *Frontiers in Engineering Conference*. 1997.
16. Stouffer, W.B., J.S. Russel, and M.G. Oliva. *Making the Strange Familiar: Creativity and the Future of Engineering Education*. in *Proceedings of the American Society for Engineering Education Annual Conference*. 2004.