Bitácora urbano/territorial 27 (4) 2017: 19 - 30
Universidad Nacional de Colombia, Bogotá

Measuring design innovation for project-based design assessment:

CONSIDERATIONS OF ROBUSTNESS AND EFFICIENCY

La medición de la innovación en diseño para la evaluación del diseño basado en proyectos

Medição da inovação em design para avaliação de projeto baseada em projetos

Andrea Goncher
PhD in Engineering Education
Charles Sturt University
andreagoncher@gmail.com

Joel Chan
PhD in Psychology
Carnegie Mellon University
joelchuc@cs.cmu.edu

Christian D. Schunn
PhD in Psychology
University of Pittsburgh
schunn@pitt.edu

Recibido: 11 de marzo de 2017
Aprobado: 2 de octubre de 2017
https://doi.org/10.15446/bitacora.v27n4Esp.68959

Abstract

Instructional approaches that support the acquisition of innovation design process skills for engineering and other design students are critical to developing design competencies. Resources that enable efficient and valid evaluation of design outcomes are needed; however current evaluation methods do not apply well to the heterogeneous projects found in authentic project-based design classes such as capstone design. We develop a robust and efficient measure of design outcome innovation and validate our measure with a large and diverse set of design outcomes from a project-based design class. The measure is based on expert judgments of design concepts' value and functionality derived from a set of importance-weighted design requirements. In the context of an engineering design class, the innovation score was a statistically significant predictor of success in terms of implementation status by the client. Thus, the measure's validity was supported by its ability to predict design concepts' implementation by clients in the context of a product realization class. New design outcome assessment measures in the context of authentic project-based design environments, such as the one developed in the present study, should interface with process-based metrics to create higher-quality input into the overall assessment of design team performance.

Keywords: capstone design, innovation, assessment, evaluation.

Resumen

Los enfoques instructivos para la adquisición de habilidades en procesos de diseño innovativo para ingenieros y diseñadores son fundamentales para desarrollar competencias de diseño. Se necesitan recursos que permitan una evaluación eficiente y válida de los resultados del diseño, sin embargo, los métodos de evaluación actuales no son adecuados para los proyectos heterogéneos de las clases de diseño basadas en proyectos auténticos como diseño final. Por ello, nosotros desarrollamos una medida robusta y eficiente, y la validamos con un conjunto grande y diverso de resultados de diseño de una clase de diseño basada en proyectos. La medida se basa en juicios expertos sobre el valor y la funcionalidad de los conceptos de diseño derivados de una serie de requisitos de diseño ponderados por su importancia. Las nuevas medidas de evaluación de los resultados en el contexto de entornos de diseño auténticos basados en proyectos, como el desarrollado en el presente estudio, deben interactuar con las métricas basadas en procesos para crear una entrada de mayor calidad en la evaluación general de rendimiento del equipo de diseño.

Palabras clave: diseño final, innovación, valoración, evaluación.

Ressumo

As abordagens instrutivas para adquirir habilidades em processos de design inovadores para engenheiros e designers são fundamentais para desenvolver competências de design. Os recursos são necessários para permitir uma avaliação eficiente e válida dos resultados do projeto, no entanto, os métodos de avaliação atuais não são adequados para projetos heterogêneos de classes de design com base em projetos autênticos como design final. Portanto, desenvolvemos uma medida robusta e eficiente e a validamos com um conjunto grande e diversificado de resultados de design de uma classe de projeto baseada em projetos. A medida baseia-se em julgamentos especializados sobre o valor e a funcionalidade dos conceitos de design derivados de um número de requisitos de design ponderados pela sua importância. Novas medidas para avaliar os resultados no contexto de ambientes de projeto autênticos baseados em projetos, como a desenvolvida no presente estudo, devem interagir com métricas baseadas em processos para criar uma entrada de maior qualidade no Avaliação geral do desempenho da equipe de design.

Palavras-chave: capstone design, inovação, avaliação.

1. Introduction

A key concern of design education research is to discover ways to support the training of future innovators in design (i.e., engineering or industrial designers) who have the tools and skills to produce novel artifacts that add significant value to stakeholders (Bransford, 2007; Dym, et al., 2005). Researchers and educators increasingly recognize the importance of "authentic" learning contexts (e.g., learning-by-doing, problem/project-based learning) for design students to develop the skills needed for innovation (Crawley, et al., 2007; Dym, et al., 2005; Litzinger, et al., 2011; Mills and Treagust, 2003). Authentic learning contexts often take the form of a project-based class, such as capstone design. In such a class, students (individually or in teams) tackle a design problem, moving through major early phases of professional design, from problem formulation and understanding, to requirements definition, to concept generation and selection, to prototyping (stopping short of fabrication).

Andrea Goncher

Andrea Goncher is originally from the US. She worked on this project while a graduate student at the University of Pittsburgh. She then went on to obtain her PhD in Engineering Education from Virginia Tech, and is now a Lecturer at Charles Sturt University, conducting research on students conceptual understanding of engineering concepts.

Joel Chan

Joel Chan is originally from Malaysia. He worked on this project while a graduate student at the University of Pittsburgh, where he continued to obtain his PhD in Cognitive Psychology with research on innovation processes. He is just beginning an Assistant Professor position at the University of Maryland, after completing a postdoc position on crowd sourced innovation at Carnegie Mellon University.

Christian D. Schunn

Christian Schunn is originally from Canada. He obtained his PhD in Cognitive Psychology from Carnegie Mellon University. He is now a Professor of Psychology, Learning Sciences, and Intelligent Systems as well as Senior Scientist at the Learning Research and Development Center at the University of Pittsburgh. His research builds on the synergy of study design teams and building innovative design-based learning curricula.

Assessing student innovation in project-based courses presents unique challenges to design educators. While traditional methods (e.g., exams, written reports, etc.) can be effectively leveraged to assess students' understanding of the design process and skill in executing key elements of the process (Atman, et al., 2014), assessment of design outcomes (e.g., the creativity or quality of the team's final product) remains a difficult challenge. Authentic design problems are ill-structured and admit multiple solutions. These properties invalidate key assumptions of traditional methods of assessment (e.g., existence of one "gold standard" answer) and consequently render them unusable. Little guidance exists in the literature as to how to design and implement design outcome measures that are objective, reliable, and robust across multiple contexts. We argue that this lack of guidance is a major reason that educators tend to either neglect assessment of design outcomes, or implement them in an *ad-hoc* manner (McKenzie, et al., 2004; Pembridge and Paretti, 2010; Sobek and Jain, 2004).

The lack of robust design outcome assessment practices in project courses is a major barrier to effective instruction on innovation. While process adeptness and conceptual and technical knowledge are important components of innovative skill, how will we know if students are able to innovate if we have no robust quantitative way of measuring the innovativeness of their designs? Presumably good design processes lead to good design outcomes, but the correlation is imperfect. Design outcome measures provide an important complement to process measures for calibrating understanding of how students are developing the ability to innovate, providing educators and students more precise information with which to focus learning efforts. In this paper, we contribute to addressing this gap by presenting a new objective measure of design

innovation for use in project courses. We adapt a well-known measure of engineering design innovation from Shah and colleagues (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003), addressing key issues in adapting it to the project course context, including specifying a robust and principled process for devising rating scales across multiple problems with few available solutions for comparison.

1.1 Existing approaches for assessing capstone design and other project-based design courses

In evaluating capstone design and student performance, faculty who teach design use a wide range of assessments that typically include written reports, presentations, as well as the technical quality of the design. Pembridge and Paretti (2010) report that several stakeholders contribute to the assessment of a project, including the course instructor(s), project advisors, as well as other students, although less than half of the faculty surveyed in their study reported using the involvement of industry partners in assessment.

Most importantly, consistent assessment measures across evaluators and projects are critical, yet difficult to apply in a design project course with multiple design contexts (one for each student or student team), each design context involving different stakeholders. Rubrics that focus on important characteristics within each activity that can be assessed, e.g. design logbooks, final presentations, are used in order to communicate expectations to students and apply consistent evaluation (Nassersharif and Rousseau, 2010). McKenzie, et al. (2004) found that faculty predominately use the overall combination of written reports, final product technical quality, and design deliverables to determine individual student performance. However, faculty indicated that they lacked the information or knowledge on how to develop rubrics that work for all users in the capstone design setting. Further, the reported success of a project is mainly composed of delineated assessments, that may or may not have had client or industry input on the assessment of the final design (i.e., little emphasis on the overall success of the design outcome—the value provided to the stakeholders).

Few principled approaches to assessing design innovation in project courses exist. A notable exception is Sobek and Jain's (2004) Design Quality Rubric (DQR) and Client Satisfaction Questionnaire (CSQ). These assessments were developed evaluate projects based on the design outcome *per se* rather than the process used. The DQR focused on key dimensions of design innovation synthesized from various engineering curricula and design competitions, including meeting technical requirements, and feasibility, novelty, and simplicity of the design. The CSQ focused on a number of different dimensions, including technical quality of the final design, benefits of the design for the company, level of involvement with the design team, complexity of the design task, and quality of final deliverables (e.g., report, presentation, engineering drawings, prototypes).

This work provides a good starting point, but leaves opportunities for improvement. We argue that holistic Likert-like ratings of technical quality do not adequately capture the inherent complexity of design, where design problems address multiple (and sometimes competing) design requirements of varying importance. Holistic rating risks collapsing performance into just one or two requirements, which can mask worthwhile advances for other aspects of the problem the students might have produced.

1.2 Existing quantitative measures of design innovation outcomes

To address the need for function-focused assessment of design outcomes, we look to the literature on design innovation research for measures that might apply in an assessment context.

1.2.1 Consensual Assessment Technique (CAT)

Amabile's (1982; 1983; 1996) Consensual Assessment Technique (CAT), in which groups of domain experts judge the creativity of a set of creative products on an unanchored, global creativity scale (typically on a Likert-type 1-6 scale). This approach is used often in studies of creativity in various domains other than design. Its face validity is high, since its foundational assumption that domain expert consensus judgments on a product's creativity are a valid and reliable measure matches that of real-world judgments of creative achievement (e.g., Academy Awards, Nobel Prizes, etc.). Reliability of the average judgments across the group of experts is often high, with typical Cronbach alphas ranging from 0.7 to 0.9. However, this validity and reliability is critically contingent on obtaining both the right *kind* of experts— validity and reliability are compromised when raters are not experts in the relevant domain (Kaufman, et al., 2008; Kaufman, Baer and Cole, 2009) —and *number* of experts— typically seven or more experts to obtain acceptable measure reliability (Amabile, 1982). In an engineering education context, it should be relatively easy to obtain the right *kind* of experts, but may be prohibitively challenging to obtain the right *number* of experts.

1.2.2 Creative Product Semantic Scale (CPSS)

Another approach is the Creative Product Semantic Scale (CPSS) (Besemer 1998; Besemer & O'Quin, 1999; O'Quin and Besemer, 1989). This method consists of providing 1-7 Likert ratings for 55 items, with each item anchored at the top and bottom end of the scale by bipolar adjectives. The adjectives are clustered according to three critical dimensions of creative products, with each dimension composed of sub-dimensions: *novelty* (composed of originality and surprise), *resolution* (composed of logical, useful, valuable, and understandable), and *elaboration and synthesis* (also called style; composed of organic, well-crafted, and elegant). These dimensions are based on Besemer and Treffinger's (1981) Creative Product Analysis Model. Validity has been established for the novelty dimension of the scale, which has been shown to be capable of satisfactorily discriminating between more and less novel known products; validity for the resolution and elaboration and synthesis sub-dimensions remains to be established convincingly. Reliability has also been shown to be generally good. However, similar to the CAT, a potential barrier to adoption in any area of education is its cost. On average, 10 mi-

nutes are required for the rating of a single product, per rater, and Besemer and colleagues (The CPSSAcademic, n.d.) recommend that at least 50-60 knowledgeable raters provide ratings for each product in order to achieve good reliability and validity. In addition, the current version of the scale is proprietary and is pay-per-use. Based on the current fee structure, the cost of obtaining the recommended minimum 50 ratings for a given product would be US$450 per product (The CPSSAcademic, n.d.). Few university departments have the budgetary resources for this approach.

A further issue shared by both CAT and CPSS is the lack of insight into the domain-specific (function-focused) dimensions of design innovation. If utilized for assessment, students will gain a global/holistic sense of their innovation performance (similar to Sobek and Jain's DQR) but it may be difficult to use that feedback to diagnose and fix deficiencies in knowledge or skill required to innovate.

1.2.3 Shah and colleagues' system of ideation metrics

A final approach to consider is the system of design innovation metrics proposed by Shah and colleagues (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003). Their system includes detailed versions of four traditional metrics of creative processes and products: quantity, variety, novelty, and quality. Because the focus of this paper is on approaches for measuring innovation output rather than process characteristics, we focus on Shah and colleagues' novelty and quality metrics.

Similar to the CPSS, the Shah and colleagues approach was developed specifically within the context of engineering design research (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003). As such, it has a heavy focus on functions and requirements, which are important in many (but not all) areas of design. Measurement begins with a functional decomposition of the overall product. Any whole product (a system) can be divided into functional subsystems. For instance, a car can be divided into a propulsion subsystem, a steering subsystem, a load-carrying subsystem, and a safety subsystem. Design outputs are then evaluated separately on each of the functions (e.g., its propulsion, its steering). Variations in designs that do not impact these main functions are considered irrelevant. Moreover, the functions may not be equal in overall importance. Thus, the overall novelty or quality score is a weighted-by-function-importance average across sub-scores for each function.

Within the functional decomposition, *novelty* is the novelty of particular function feature(s). The exact novelty calculation is a variation of an approach used by Torrance (1962), and can be determined using empirically derived or *a priori* estimates of the novelty of particular features. In the *a priori* method, the experimenter determines (before conducting the experiment) how common different choices for each function have been in the past—this method has questionable reliability and validity for complex real design applications because it requires very deep knowledge of every design task by the instructor or researcher. More importantly, in many authentic design-problem contexts, prior solutions for a given design problem may not even exist. In the empirical method, essentially focusing on ease-of-generation rather than novelty *per se*, data is derived from a large set of participant responses to a fixed design task: ideas are novel to the extent that few participants generated them. This method is very convenient for fixed design tasks given to many participants (e.g., in an experiment or at a design competition), but is not useful for evaluation of novelty when each designer/team tackles a different design task, such as in capstone design classes and other authentic project-based design classes.

The approach used by Shah and colleagues (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003) to measure idea quality borrows heavily from common engineering evaluation metric approaches such as Quality Function Deployment (Huang, 1996) and Decision Tables (Pahl and Beitz, 1996). In these approaches, design concepts are evaluated on the degree to which they meet the main functional criteria of the overall product (e.g., strength, cost, manufacturability, ease-of-use). An overall score is determined by a weighted sum of each functional criterion, with the weights reflecting the importance of each functional criterion (e.g., perhaps ease-of-use is more important in one design context, but cost is more important in another context). In contrast to simple holistic quality ratings, this method for evaluating quality is likely to be reliable because it is more objective, and it is likely to be valid because it is directly linked to design functionality.

The Shah and colleagues (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003) quality and novelty metrics have generally been found to have good validity and reliability: inter-rater agreement for coding of ideas for functional elements (for novelty calculations) is often high (Cohen' kappa of 0.8 or higher) (Chan, et al., 2011; Tseng, et al., 2008), and inter-rater agreement is similarly high for quality ratings (Pearson's correlation of 0.7 or higher) (Chan, et al., 2011; Linsey, et al., 2010). The quality metric has the additional advantages of high face and construct validity due to its adaption from industry-wide methods of concept screening (e.g., Pugh, 1991). Cost barriers are also relatively low, as the method does not require a large number of experts to achieve satisfactory reliability, and it is not proprietary pay-for-use. Perhaps for this reason, these metrics have been widely used to good effect in engineering design research (e.g., Chan, et al., 2011; Kurtoglu, Campbell and Linsey, 2009; Linsey, et al., 2010; 2011; Tseng et al., 2008; Wilson, et al., 2010).

However, some important details are left unspecified by Shah and colleagues (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003), which have potential impacts on its translation into the capstone design assessment context; these issues are summarized in Table 1. First, regarding overall functional decomposition, what guidelines should be followed in decomposing a product into functional sub-systems? Ideally, we would like this design outcome measure to be comparable across projects within a class, across semesters of a class, and across instructors to support various scales of formative evaluation and empirical research. Secondly, when gauging the

novelty of a final product, or set of final products, it is not clear how to establish the universe of ideas within which to estimate a given product's rarity. In capstone design courses, observed idea spaces are typically small because teams often solve different problems, and if they do solve the same problem, the number of teams is typically small. In these situations, estimates of the baseline frequencies of various concept types may either be circular in the identity case or not be stable enough to generate valid and reliable estimates of a final product's relative novelty. Further, there are often few outside benchmarks from which to draw estimates of the novelty of concepts. This is especially the case if the teams are addressing authentic design problems, which are likely to be unsolved, and therefore by definition do not have an established universe of possible solutions to compare against. Finally, there are questions surrounding the use of sub-scales for the quality metric. From where should these sub-scales come, and how should the sub-scales be weighted? Who should generate these sub-scales and assign weights, and by what criteria? The rating scale is also vaguely specified. How should instructors determine the size of the rating scale, and appropriate anchors for each portion of the rating scale? Without answers to these questions, applying this assessment approach in a capstone context may lead to invalid or unreliable assessments, potentially obstructing rather than facilitating accurate assessment of students' skill development.

Aspect	Question
Functional decomposition	How to decompose?
Novelty	What universe of ideas serves as baseline?
Quality sub-scales	Who/how to define sub-scales?
Quality rating scale	Size and anchors for scale?

Table 1. Summary of key methodological ambiguities in Shah and colleagues' ideation metrics system

2. Description of proposed innovation measure

2.1 Overview

In order to evaluate the success of a design outcome in the context of capstone design course environment —where a variety of projects typically have one design outcome for each project—we developed a metric to address key issues in evaluating design outcomes for a course with a diverse set of projects. Our working definition of innovative design is the production of artifacts that add significant value over existing/prior artifacts and address a given need or want. A product that adds value to existing artifacts can be created through realizing new functions and properties (Pahl, et al., 2007) or meeting requirements in novel ways (Cropley and Cropley, 2005). Focusing on the performance of a product's functional components provides an evaluation measure that we can extend to a diverse set of design outcomes, include dimensions that are relatively qualitative.

Following Shah and colleagues' quality metric (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003), our innovation measure codes designs for innovation based on a set of project-specific subscales that relate to key functional requirements of the design, with each subscale weighted by importance to the overall design. Appropriate judges are selected to rate designs on their respective subscales, using a well-defined rating scale. The final innovation measure for a given design is a weighted combination of its subscale ratings.

Departing from Shah and colleagues (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003), we collapse novelty into this single outcome measure, leaving novelty implicit in the measure. Our motivation and justification for this is as follows. First, our theoretical assumption is that innovative design is the production of artifacts that add significant value over existing/prior artifacts that address a given need or want. This assumption privileges functionality and value-added over novelty *per se* (i.e., without functionality/value added), in a similar fashion to other theoretical conceptualizations of innovation (Chandy and Tellis, 1998; Garcia and Calantone, 2003). Novelty *per se* is not always beneficial, as the potential value of distinctiveness from the competition may be offset by additional costs required to bring that design to market, from manufacturing, supply chain, and user-support perspectives. Conversely, relatively small changes in functionality (i.e., low novelty) can sometimes lead to "game-changing" degrees of value-added. Second, in authentic design contexts, a minimal level of novelty is an implicit requirement of design briefs. Clients do not seek designed artifacts that are identical to existing competition; rather, they seek new designs that are different in some substantial way from the competition (usually in terms of value added). Finally, as discussed earlier, a frequency or *a priori*-based approach to estimating novelty of designs is potentially problematic in most design process-outcome research contexts. We believe that leaving novelty implicit in the measure, unless it is an explicit and separately specified client requirement, yields a clean and usable measure of design innovation, particularly in an engineering context, as well as other design contexts where functionality and value added are paramount considerations.

2.2 Deriving the sub-scales

2.2.1 The client-defined design problem case

For capstone projects in which the design problem is determined by an external agent (e.g., entrepreneur, an end-user, or company), requirements are derived directly from the design process. The process we specify assumes that the design projects to which this measure will be applied will begin with some sort of design brief, and go through an initial requirements clarification phase, where customer needs are translated into specific design requirements. Requirements continue to be iteratively refined, dropped, or added, via continued conversation and feedback loops with stakeholders as the design process progresses. We further assume that requirements will at least implicitly be ordered in an importance hierarchy, where certain requirements may be core/

critical, others less so, and still others optional. This iterative generating and refining of importance-weighted requirements is prescribed in many prominent engineering design texts (Otto and Wood, 2000; Ullman, 2002; Ulrich and Eppinger, 2008), and practiced in many (but not all!) capstone and experiential design courses.

Our proposed method extracts the final set of requirements for a given project and uses that set as the sub-scales for the innovation measure. Importance weights are extracted from project documents if they are explicitly specified. If design teams do not specify importance ratings for the set of design requirements, instructors using this method should query the team for explicit weightings. To ensure that requirements and importance weights are properly specified, we recommend an additional step prior to rating where a knowledgeable expert checks the requirements. With poorer-functioning/performing teams, one runs the risk with this method of obtaining importance-weighted requirements/specifications with serious flaws, such as incompleteness, poorly assigned importance weightings, and others. Even in higher-performing teams, however, there could be differences in how well requirements are captured and specified, and these differences could be confounded with other predictor variables of interest (e.g., conflict handling, which can influence accuracy of requirements, in addition to quality of solutions). Adding the extra step of screening the final set of requirements helps to ensure that the final measure of innovation validly measures the extent to which the design adds value over existing/prior designs.

2.2.2 The instructor-defined design problem case

For capstone projects where the instructor forms the design problem, requirements are generated *a priori*. The instructor can generate importance weightings *a priori* as well, since the instructor is, in effect, the client.

2.3 The rating scales

We define two separate scales for rating the degree to which requirements are met, even with relatively qualitative requirements. In the general case, the scale consists of 4-points as follows: 0 — Did not come close to meeting requirement; 1 — Fell just short of meeting requirement; 2 — Met requirement, but did not exceed significantly; 3 — Significantly exceeded the requirement.

In design, requirements can sometimes include a specification of both minimal and ideal values, where minimal values describe an outcome case with which stakeholders would be satisfied, and ideal values typically describe extremely high-quality thresholds that are often not possible without significant compromises on other sub-systems or significant breakthrough innovation. For example, a market analysis may find that a sufficient number of users would pay $20 for a product (minimal value), but a much larger number of users would pay $10 for the product (ideal value). In this more specialized case, a 5-point scale is employed, as follows: 0 — Did not come close to minimal; 1 — Fell just short of minimal; 2 — Met minimal but did not meet ideal; 3 — Met ideal; 4 — Significantly exceeded ideal value. This 5-point scale allows for a measure of design success that goes above and beyond normal standards of excellence. Each point on this scale below 4 corresponds to its matching point on the general 4-point scale. Table 2 shows the mapping between the scale variations.

General		If minimal or ideal specified
Did not come close	0	Did not come close to minimal
Fell just short	1	Fell just short of minimum
Met, but did not exceed significantly	2	Met minimal, but not ideal
Significantly exceeded	3	Met ideal
	4	Significantly exceeded ideal

Table 2. Unified rating scale for general requirements and requirements for which minimal or ideal values are specified

Why use these categories rather than simply directly using the metrics upon which the categories are based? First, the categories normalize the measure across many very diverse metrics of performance that are on completely different dimensions (e.g., cost, usability, strength). That is, one cannot meaningfully directly average dollars, usability ratings, and tensile strength measurements. Second, the approach allows for the inclusion of more qualitative dimensions that do not have an underlying quantitative scale (e.g., inclusion of a certain design esthetic). Third, the categories take into account the satisficing nature of design, in which the quality of a design changes categorically in the minds of users when thresholds are met (Bansal, et al., 2009). For example, a pen that is 120% of what the intended user is willing to pay is not so different from a pen that is 150% of what the intended user is willing to pay (i.e., neither pen is purchased); similarly, 20% and 50% of target costs are close to equally good. Fourth, the categories likely reflect the realistic precision on many of the metrics—while some dimensions like strength can be measured precisely to many decimals, dimensions like usability, attractiveness, and manufacturing costs can only be roughly estimated during the design process, so more precision on the innovation scale is not warranted.

2.4 Selecting appropriate judges

2.4.1 The client-defined design problem case

For each client-defined design problem case, an appropriate domain expert or stakeholder can serve as the judge.

2.4.2 The instructor-defined design problem case

In the instructor-defined design problem case, a faculty or staff member, e.g. graduate student, with relevant content expertise can act as an appropriate judge.

2.5 Final innovation measure

The final innovation measure for a given design is a weighted combination of the sub-scale ratings for that design, as given by equation 1:

$$\frac{\sum_{i=1}^{i} w_i \times r_i}{\sum_{i=1}^{i} w_i \times r_{max}} \times 100$$

where w_i is the importance weighting for the i[th] sub-scale to the overall project, e.g. how important is meeting the ideal cost relative to meeting the ideal tensile strength, r_i is the observed rating for the i[th] sub-scale, and r_{max} is the maximum possible rating for general requirements, set to 3. This equation yields a normalized score on a 0 to 100 scale. Intuitively, the innovation score indexes the percentage of available "innovation points" the overall design earned. Notice that, because r_{max} is set to a constant of 3, it is in principle possible to obtain a score that exceeds the maximum possible score (if the project includes requirements with minimal/ideal values). The reason the combination function allows for this is that a design aspect that significantly exceeds ideal values, i.e., a score of 4, it should be treated as an outstanding design effort that goes above and beyond excellence. This approach also puts the lower performance levels on equal levels, e.g., falling just short is always 33, even if an ideal is possible.

3. An example implementation

3.1 Research context and participants

In this section, we describe an example implementation of our metric to investigate:

- Can the proposed metric be used as an effective assessment in the context of an educational setting?
- Does the proposed metric provide a valid and reliable measure of success for a capstone design course with diverse set of design outcomes?

We apply our metric to a sample of engineering projects in the context of multidisciplinary engineering student design teams working enrolled in a product realization course at a large research university in the North-Eastern United States. Courses employing a product-based learning pedagogical approach and advised or sponsored by an outside client are common in US schools of engineering (Hotaling, et al., 2012). The course also has many features in common with capstone design courses required of all engineering undergraduates in the US, except perhaps being more multi-disciplinary and employing a more structured design process.

Our sample consists of 57 teams across 7 semester-long implementations of the course (from Spring 2007 to Summer 2009). In each implementation, multidisciplinary teams of 3 to 5 students took products from concept to functional prototype. Up to US$2,500 in funding was made available for student teams to make conceptual prototypes as part of their products. An industry mentor was made available for each project to assist the team in making design decisions.

Each team worked on a different project (typically 6 different teams/projects per semester), and each semester there were rarely repeated projects from prior semesters. Application domains were diverse, ranging from running shoe cushion monitoring devices to computerized pill minders for dispensing medication to Radio Frequency Identification (RFID) personnel badge systems. Examples of the final designs in Figure 1 illustrate the diversity in project outcomes and applications. Approximately 20% of design teams in these courses produced products that are later patented, and teams appeared to vary greatly in terms of overall innovation. This sample of teams was collected as part of a larger research project examining process antecedents of engineering team innovation, which motivated the need for a measure of innovation that is reliable and valid across this heterogeneous set of projects.

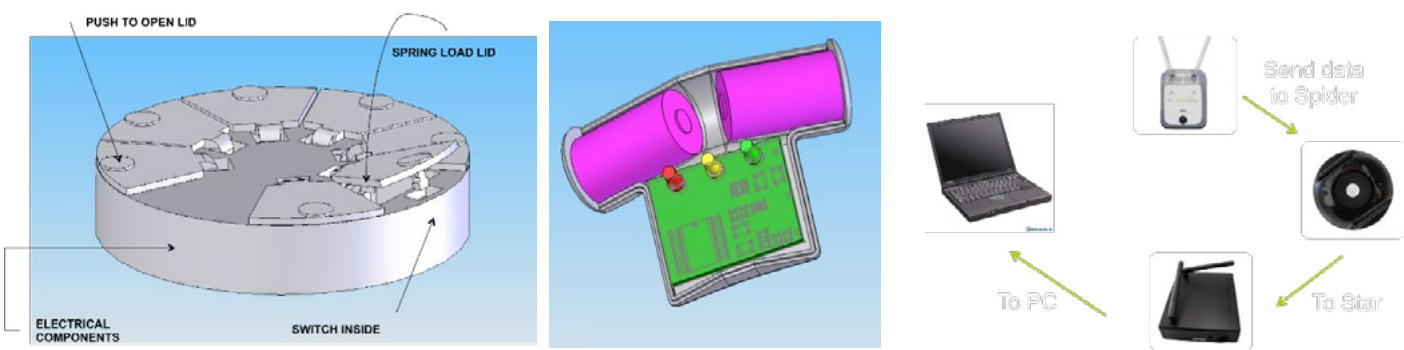

Figure 1. Examples of team final designs

3.2 Deriving the sub-scales

As part of the guided design process used in the course, teams developed requirements by specifying the needed functions, behaviors, and attributes of their proposed design solution. Teams also rated each requirement on an importance scale of 1 (weakly important to the development of the design) to 5 (absolutely important to the development of the design), and provided minimal and ideal values for requirements, where applicable. We extracted the specified requirements, levels, and importance ratings from their design artifacts, such as design notebooks, project write-ups, and intermediate project presentations.

These team-provided requirements and importance ratings were treated as a first draft of the requirements officially used to evaluate innovation. Prior to assessing team performance for each subscale on our innovation metric, the course instructor assigned an importance rating to each requirement, evaluated the requirement set as a whole to identify if requirements necessary to the success of the design solution were omitted during the requirement identification stage, and identified unnecessary or extraneous requirements. Instructors identified requirements that should have been included and these "omitted requirements" were added (along with their respective weights) to the evaluation set. Unnecessary or extraneous requirements were also removed, based on the instructor's evaluation.

3.3 Selecting appropriate experts

In this implementation, the course instructor for the semester served as the expert to judge each team's innovation measure. In semesters where there were multiple instructors, supervision of teams tended to be divided among the instructors based on their relative expertise within each team's project domain. Thus, one expert rated each team's requirements, i.e., either the overall course instructor (for semesters where there was only one instructor) or the instructor who was most familiar with and had expertise in their project (for semesters where there were multiple instructors). As is often the case with instructors of such courses in schools of engineering, the course instructors had extensive experience in product realization, including numerous patents, startup company experience, and industry consulting, and also had extensive content knowledge about material sciences, mechanical engineering, and electrical engineering, the key elements of projects selected for analysis.

3.4 Ratings and final innovation measures

To give a flavor of the ratings and importance-weighted combination final measures in this dataset, we present examples of ratings and final innovation measures from two teams. The first team's project was to design a prototype pill minder device to help patients who are on multiple medications to take the proper pills on the proper day. Table 3 shows the requirements, instructor-checked importance weightings, and final ratings for the performance of the team's final solution.

Requirement	Importance	Performance Rating
Daily Pill Keeper Format (e.g. visible daily format)	3	1
Good Ergonomics (e.g. elderly or mild arthritis can still open)	5	2
Convenient format for data (e.g. separately recorded events)	3	1
Portable and Durable (e.g. Minimized Size)	5	1
Provide data and time of openings	5	2
Exportable data files (e.g. user-friendly data transfer via USB port)	3	2
Unique Device Identifier	3	1
Stable Data Recording (bumps or drops not likely to damage the recorded data)	3	1

Table 3. Requirements, importance weights, and performance ratings for Team 1

Following equation 1, the final innovation score for this team was:

$$\frac{3\times 1)+(5\times 2)+(3\times 1)+(5\times 1)+(5\times 2)+(3\times 2)+(3\times 1)+(3\times 1)}{3\times 3)+(5\times 3)+(3\times 3)+(5\times 3)+(5\times 3)+(3\times 3)+(3\times 3)+(3\times 3)}$$

$$=\frac{43}{90}\times 100=48$$

which is a very *low* score among teams in this dataset.

The second team's project was to develop a low-cost, portable implementation of a blood pathogen detector using Loop-mediated isothermal Amplification technology. Table 4 shows the requirements, importance weightings, and ratings for this team.

Requirement	Importance	Performance Rating
Handheld device; size of a cigarette pack	3	2
Able to heat samples at 60±5 °C for 60 minutes; heating coil or exothermic reaction	5	3
Portable; Use batteries or reaction; Not to be plugged into wall socket	4	2
Have a blue light to check samples as positive or negative; Contain a blue LED / Black Light / UV light that shines on test	4	3
Come as a complete kit, including everything that is needed; Contain grid with all features, a pipette, test tubes with reactant	4	3
Affordable; Cost between $50-$100	4	3

Table 4. Requirements, importance weights, and performance ratings for Team 2

Following equation 1, the final innovation score for this team was:

$$\frac{(3\times 2)+(5\times 3)+(4\times 2)+(4\times 3)+(4\times 3)+(4\times 3)}{(3\times 3)+(5\times 3)+(4\times 3)+(4\times 3)+(4\times 3)+(4\times 3)}=\frac{65}{72}\times 100=90$$

which is a very *high* score among teams in this dataset.

4. Assessing the validity of the measure

4.1 Methods

To examine the validity of the measure, we examined the relationship between teams' final innovation scores and whether their sponsor implemented their resulting design to some degree. To obtain this measure of implementation status, we queried sponsors at the end of the semester whether they planned to implement or were implementing the teams' final design (or at least some aspect of it) at the company. On the basis of the responses from each team's sponsor's responses, we constructed a three-level (yes, maybe, no) implementation status measure, where "yes" indexed responses that indicated at least some aspect of the team's design was currently being implemented, "maybe" indexed response that indicated at least some aspect of the design was being considered for future implementation by the sponsor in some fashion, and "no" indexed unambiguous responses that indicated no part of the team's design was currently being implemented or considered for future implementation by the sponsor. Our sample for this analysis consisted of 47 teams across 7 semesters of the course; 10 of the 57 teams in the full dataset were excluded because we were either unable to obtain the implementation information from the sponsor, or the sponsor had not yet decided on implementation during the duration of our data collection period.

4.2 Results

Obtained innovation scores in our sample ranged from a minimum of 0 (only one team obtained this score) to a maximum of 100 (only one team obtained this score), with a mean of 63.9 and a standard deviation of 18.9. That is, the proposed measure has significant variation across teams of this sort (i.e., the full range of the measure is obtained, and the mean performance is somewhere near the middle of the scale). There were 18 teams with a "yes" response for implementation, 16 teams with "maybe", and 13 teams with "no".

Figure 2 shows the relationship between teams' innovation score and implementation status. To statistically explore the association between the measures, we conducted a one-way analysis of variance (ANOVA) with implementation status as the between-subjects factor. The ANOVA revealed a significantmain effect of implementation status on innovation score, $F(2, 44) = 8.22$, $p < 0.01$. Post-hoc pairwise Tukey tests revealed that teams with "yes" implementation status had significantly higher innovation scores ($M = 73.3$, $SD = 13.2$) than teams with "no" implementation status ($M = 49.2$, $SD = 23.8$), Cohen's $d = 1.36$, $p < .001$. Teams with "maybe" implementation status also had significantly higher innovation scores ($M = 67.8$, $SD = 12.3$) than teams with "no" implementation status, Cohen's $d = 1.07$, $p < 0.05$. "Yes" teams also had higher scores than "maybe" teams, $d = 0.52$, but the difference was not statistically significant, $p = 0.49$. While one might expect "yes" and "maybe" teams to be significantly different, we believe the small difference observed here reflects the complexity of sponsors' decision processes for implementing teams' projects. It is likely the case that many project outcomes were good enough to implement, but various internal/contingent factors (e.g., budget constraints, intellectual property considerations) might have prevented immediate implementation.

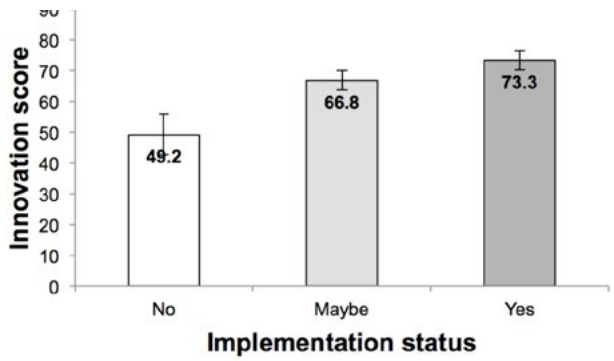

Figure 2. Team innovation score by implementation status. Error bars are +/-1 standard error

5. Conclusions

In this paper we have motivated the need for a robust and efficient measure of design innovation, suitable for design outcome assessment in authentic project/problem-based design courses. We have presented an advance on Shah and colleagues' ideation metrics system (Shah, Kulkarni and Vargas-Hernandez, 2000; Shah, Vargas-Hernandez and Smith, 2003), adapting their quality metric into an improved innovation measure that meets this need, and addressing key issues in adapting this measure to the design education context (see Table 5 for a summary of how our measure addresses each key issue). Finally, we have demonstrated evidence of this measure's convergent validity in terms of a significant association with the probability of a final design concept being implemented by its client/sponsor, in the context of a product realization course.

Aspect	Question	Answer
Functional decomposition	How to decompose?	Use project-defined, expert/stakeholder-vetted decomposition
Novelty	What universe of ideas serves as baseline?	Collapse novelty into "value added" composite scale
Quality sub-scales	Who/how to define sub-scales?	Sub-scales are project-defined and expert/stakeholder-vetted based on functional decomposition
Quality rating scale	Size and anchors for scale?	Rating scale is standardized across sub-scales using qualitative anchors

Table 5. Summary of our answers to key methodological ambiguities in Shah and colleagues' ideation metrics system

6. Limitations

While the measure described and evaluated in this paper demonstrated a robust metric based on the functionality and requirement performance, we also identify potential limitations with this approach. First, a prerequisite of our innovation measure is that the designer or design team being measured goes through a process of explicit and iterative refinement of design requirements. While developing our measure based on this prerequisite makes our measure well suited for many realistic design contexts, it does mean that the measure may not be valid in settings where this assumption does not hold, for instance, in design courses where student teams do not follow a requirements-focused design process (e.g., in more esthetic-oriented design or when exploring new technologies more broadly without consideration of particular user groups' needs). Thus, it is possible that our measure cannot capture the full bottom range of design innovation performance. However, as we have shown in our example implementation, and in our validity study, our measure does have a useful range of variation, and is able to capture the performance differences of design teams whose final design concepts are of poor enough quality to preclude consideration for implementation.

Second, some evaluators might wonder about the reliability of our measure, given that we have not reported reliability analysis. The reason for the lack of reliability data was that, in our implementation of the measure described above, we were not able to get more than one qualified expert for most of the teams. In most semesters, there was only one instructor, and in semesters where there were multiple instructors, they focused separately on teams whose projects they were most familiar with; in that case, obtaining ratings from multiple instructors for all teams was inappropriate, as their level of expertise and familiarity with some teams was not sufficient. We note, however, that the convergent validity of the measure can be treated as indirect evidence for its satisfactory reliability. If the measure were generally unreliable, this statistical noise would obscure meaningful relationships, and the relationship between the innovation score and implementation status presented above would be unlikely to be detected. In our other work, the innovation measure proposed here has also proven useful for discriminating between better and worse design process characteristics, such as the amount and timing of appropriate design tool use (Jang and Schunn, 2012; Paletz, Chan, and Schunn, 2017).

From these two productive uses of the measure, we conclude that our measure has satisfactory reliability —at least for detecting effects of equivalent or greater size compared with the effects presented in this paper and in (Jang and Schunn, 2012; Paletz, Chan, and Schunn, 2017)— and is robust to the potential statistical noise from using one expert.

Third, we have provided an approach to measuring outcomes, but it does not provide insights into underlying educational problems. Measurement is helpful to formative assessment, but quantitative measures alone are not enough to guide student learning. It is particularly likely that feedback about design process quality will be needed to complement the feedback obtained on design outcomes from our measures.

Finally, some researchers may worry that our measure fails to capture the distinction between incremental and radical innovation, at least in part because novelty is only implicit in the measure. However, as we have argued, the theoretical basis of our measure of innovation construes *value-added* as the primary component of the construct of innovation. Nevertheless, we freely admit that a single ideal measure of innovation for all purposes is neither feasible nor desirable. Design instructors and/or researchers may find it useful to view our measure as being part of a suite of possible innovation measures to illuminate the various aspects or nuances of design innovation skill. We recommend future research investigating the capacity of a more absolute scale that incorporates weightings based on instructor or client perceptions of difficulty.

7. Suggestions for design educators

Capstone design instructors are faced with assessing the success of the course projects, which is particularly difficult when only one design outcome exists for a particular design prompt. The culmination of a capstone design project exhibits the final design outcomes as well as various design documents that support the process used to arrive at the final design solutions—including the translation of requirements into design solutions. While a successful design outcome is beneficial to the client as well as the design team, what we can derive from this finding as design educators is the impact of developing students' design process skills. In this study, requirement definition and its link to

successful design outcomes demonstrates the importance of instructing and facilitating students in defining and writing requirements that address client needs/ wants. Based on the results of our measure's implementation, we recommend that students iteratively work with their client(s) and possibly the instructor to define the applicable requirements as well as a method for meeting the requirements to the appropriate degree.

Typically, one instructor is responsible for assessing multiple projects as part of the capstone design course, which makes it difficult to employ many types or different assessments for the elements of the design project. The satisfactory convergent validity (and hence implicitly the reliability) of our measure is encouraging for instructors who wish to employ our measure, but are also constrained by resource limitations, e.g., only 1 qualified expert per team (a situation that, we suspect, is not uncommon in capstone design courses). Nevertheless, we encourage instructors who wish to employ our measure to use at least two experts, if possible, and recommend explicit reliability analysis to future development work on the measure.

In the context of design education research and development, a results-focused approach to measuring student outcomes in project-based learning cannot and should not be viewed as a comprehensive measure of student progress and performance in developing innovation competencies. Important complements include knowledge and process-based assessments targeted specifically for creativity and innovation (Daly, Mosyjowski and Seifert, 2014). These types of assessments can also interface with more holistic assessment approaches of design outcome success, to create higher-quality input to the assessments of the design teams' performances.

Acknowledgements

This work was supported by grant SBE-0738071 from the National Science Foundation.

References

AMABILE, T. M. (1982). "Social psychology of creativity: a consensual assessment technique". *Journal of Personality and Social Psychology*, 43 (5): 997-1013.

AMABILE, T. M. (1983). *The social psychology of creativity.* New York: Springer-Verlag.

AMABILE, T. M. (1996). *Creativity in context: update to the social psychology of creativity.* Boulder: Westview.

ATMAN, C. J., et al. (2014) "Engineering design education: research, practice and examples that link the two". In: A. Johri and B. Olds (eds.), *Cambridge Handbook of Engineering Education Research*. Cambridge: Cambridge University Press, pp. 201-226.

BANSAL, M., et al. (2009). "Product design in a market with satisficing customers". In: S. Netessine and C. S. Tang (eds.), *Consumer-driven demand and operations management models: a systematic study of information-technology-enabled sales mechanisms*. New York: Springer, pp. 37-62.

BESEMER, S. P. (1998). "Creative product analysis matrix: testing the model structure and a comparison among products--three novel chairs". *Creativity Research Journal*, 11 (4): 333-346.

BESEMER, S. P. and O'QUIN, K. (1999). "Confirming the three-factor creative product analysis matrix model in an American sample". *Creativity Research Journal*, 12 (4): 287-296.

BESEMER, S. P. and TREFFINGER, D. J. (1981). "Analysis of creative products: review and synthesis". *The Journal of Creative Behavior*, 15 (3): 158-178.

BRANSFORD, J. (2007). "Preparing people for rapidly changing environments". *Journal of Engineering Education*, 96 (1): 1-3.

CHAN, J., et al. (2011). "On the benefits and pitfalls of analogies for innovative design: ideation performance based on analogical distance, commonness, and modality of examples". *Journal of Mechanical Design*, 133, 081004.

CHANDY, R. K. and TELLIS, G. J. (1998). "Organizing for radical product innovation: the overlooked role of willingness to cannibalize". *Journal of Marketing Research*, 35 (4): 474-487.

THE CPSSACADEMIC. (n.d.). *The cpssacademic*. Retrieved from: http://ideafusion.biz/home/creative-product-semantic-scale/the-cpss-academic

CRAWLEY, E., et al. (2007). *Rethinking engineering education: the CDIO approach*. New York: Springer.

CROPLEY, D. H. and CROPLEY, A. J. (2005). "Engineering creativity: a systems concept of functional creativity". In: J. C. Kaufman and J. Baer (eds.), *Creativity across domains: faces of the muse*. Mahwah: Lawrence Erlbaum, pp. 169-185.

DALY, S. R., MOSYJOWSKI, E. A. and SEIFERT, C. M. (2014). "Teaching creativity in engineering courses". *Journal of Engineering Education*, 103: 417-449.

DYM, C. L., et al. (2005). "Engineering design thinking, teaching, and learning". *Journal of Engineering Education*, 94 (1): 103-120.

GARCIA, R. and CALANTONE, R. (2003). "A critical look at technological innovation typology and innovativeness terminology: a literature review". *Journal of Product Innovation Management*, 19 (2): 110-132.

HOTALING, N., et al. (2012). "A quantitative analysis of the effects of a multidisciplinary engineering capstone design course". *Journal of Engineering Education*, 101: 630-656.

HUANG, G. (1996). *Design for X: Concurrent engineering imperatives*. London: Chapman & Hall.

JANG, J. and SCHUNN, C. D. (2012). "Physical design tools support and hinder innovative engineering design". *Journal of Mechanical Design*, 134 (4): 041001.

KAUFMAN, J. C., et al. (2008). "A comparison of expert and nonexpert raters using the consensual assessment technique". *Creativity Research Journal*, 20 (2): 171-178.

KAUFMAN, J. C., BAER, J. and COLE, J. C. (2009). "Expertise, domains, and the consensual assessment technique". *The Journal of Creative Behavior*, 43 (4): 223-233.

KURTOGLU, T., CAMPBELL, M. I. and LINSEY, J. S. (2009). "An experimental study on the effects of a computational design tool on concept generation". *Design Studies*, 30 (6), 676-703.

LINSEY, S., et al. (2011). "An experimental study of group idea generation techniques: understanding the roles of idea representation and viewing methods". *Journal of Mechanical Design*, 133 (3): 031008.

LINSEY, J. S., et al. (2010). "A study of design fixation, its mitigation and perception in engineering design faculty". *Journal of Mechanical Design*, 132 (4): 041003.

LITZINGER, T., et al. (2011). "Engineering education and the development of expertise". *Journal of Engineering Education*, 100 (1): 123-150.

MCKENZIE, L. J., et al. (2004). "Capstone design courses and assessment: a national study". Salt Lake City, paper presented at the American Society for Engineering Education Annual Conference and Exposition.

MILLS, J. E. and TREAGUST, D. F. (2003). "Engineering education: is problem-based or project-based learning the answer?" *Australasian Journal of Engineering Education*, 3 (2): 2-16.

NASSERSHARIF, B. and ROUSSEAU, C. E. (2010). "Best practices in assessing capstone design projects". Paper presented at the Proceedings of the 2010 Capstone Design Conference.

O'QUIN, K. and BESEMER, S. P. (1989). "The development, reliability, and validity of the revised creative product semantic scale". *Creativity Research Journal*, 2 (4): 267-278.

OTTO, K. N. and WOOD, K. L. (2000). *Product design: techniques in reverse engineering and new product development*. Upper Saddle River: Prentice Hall.

PAHL, G. and BEITZ, W. (1996). *Engineering design: a systematic approach*. London: Springer-Verlag.

PAHL, G., et al. (2007). *Engineering design: a systematic approach*. London: Springer.

PALETZ, S., CHAN, J. and SCHUNN, C. D. (2017). "Dynamics of micro-conflicts and uncertainty in successful and unsuccessful design teams". *Design Studies*, 50: 39-69.

PEMBRIDGE, J. and PARETTI, M. (2010). "The current state of capstone design pedagogy". Louisville, paper presented at the Proceedings of the 2010 American Society for Engineering Education Annual Conference and Exposition.

PUGH, S. (1991). *Total design: integrated methods for successful product engineering.* Reading: Addison-Wesley.

SHAH, J. J., KULKARNI, S. V. and VARGAS-HERNANDEZ, N. (2000). "Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments". *Journal of Mechanical Design*, 122 (4): 377-384.

SHAH, J. J., VARGAS-HERNANDEZ, N. and SMITH, S. M. (2003). "Metrics for measuring ideation effectiveness". *Design Studies*, 24 (2): 111-134.

SOBEK, D. K. and JAIN, V. K. (2004). "Two instruments for assessing design outcomes of capstone projects". Paper presented at the American Society for Engineering Education Conference Proceedings.

TORRANCE, E. P. (1962). *Guiding creative talent*. Englewood Cliffs: Prentice Hall.

TSENG, I., et al. (2008). "The role of timing and analogical similarity in the stimulation of idea generation in design". *Design Studies*, 29 (3): 203-221.

ULLMAN, D. (2002). *The mechanical design process*. New York: McGraw-Hill.

ULRICH, K. T. and EPPINGER, S. D. (2008). *Product design and development*. New York: McGraw-Hill

WILSON, J. O., et al. (2010). "The effects of biological examples in idea generation". *Design Studies*, 31 (2): 169-186.