Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2018
August 26-29, 2018, Quebec City, Canada

DETC2018-85698

HUMAN VALIDATION OF COMPUTER VS HUMAN GENERATED DESIGN SKETCHES

Christian E. Lopez B.[1]
[1]Department of Industrial and Manufacturing Engineering
The Pennsylvania State University,
State College, Pennsylvania, 16802
Email: cql5441@psu.edu

Scarlett R. Miller[1, 2]
[2]School of Engineering Design, Technology and Professional Programs
The Pennsylvania State University,
State College, Pennsylvania, 16802
Email: shm13@psu.edu

Conrad S. Tucker[1,2]
[2]School of Engineering Design, Technology and Professional Programs
The Pennsylvania State University
University Park, PA 16802
Email: ctucker4@psu.edu

ABSTRACT

The objective of this work is to explore the perceived visual and functional characteristics of computer generated sketches, compared to human created sketches. In addition, this work explores the possible biases that humans may have towards the perceived functionality of computer generated sketches. Recent advancements in deep generative design methods have allowed designers to implement computational tools to automatically generate large pools of new design ideas. However, if computational tools are to co-create ideas and solutions alongside designers, their ability to generate not only novel but also functional ideas, needs to be explored. Moreover, since decision-makers need to select those creative ideas for further development to ensure innovation, their possible biases towards computer generated ideas need to be explored. In this study, 619 human participants were recruited to analyze the perceived visual and functional characteristics of 50 human created 2D sketches, and 50 2D sketches generated by a deep learning generative model (i.e., computer generated). The results indicate that participants perceived the computer generated sketches as more functional than the human generated sketches. This perceived functionality was not biased by the presence of labels that explicitly presented the sketches as either human or computer generated. Moreover, the results reveal that participants were not able to classify the 2D sketches as human or computer generated with accuracies greater than random chance. The results provide evidence that supports the capabilities of deep learning generative design tools and their potential to assist designers in creative tasks such as ideation.

Keywords: deep generative model, crowdsourcing, functionality, bias, sketches.

INTRODUCTION

"*Creativity… is an indispensable quality for engineering and given the growing scope of the challenges ahead and the complexity and diversity of the technologies of the 21st century, creativity will grow in importance*" [1]

Creative ideas can be extremely successful on the market, resulting in significant payoffs to sponsoring organizations and stakeholders [2]. Hence, a great deal of effort has been given to developing methods that promote the generation and selection of creative and innovative ideas. Thanks to the recent advancements in generative design, topology optimization, and deep learning algorithms, designers are increasingly benefiting from integrating computational tools into the design process [3]. Researchers argue that as these computational tools become more efficient at creating novel and functional ideas, they will foster designers' creativity. Hence, both computers and designers will co-create solutions that surpass each of their independently created ideas [4].

Deep learning algorithms (e.g., Generative Adversarial Networks [5] and Recurrent Neural Networks [6]) are being implemented to automatically generate new design ideas [7,8]. Similarly, Mass-Collaborative Product Development (MCPD) is gaining popularity within the design community as a method to develop new and creative products [3,9]. MCPD takes advantage of crowdsourcing methods to generate a larger pool of new and novel ideas [10,11]. Though an idea needs to be new and novel in order to be considered creative, it also has to meet its intended functionality and be useful [12,13]. During the latter stages of the design process, designers create CAD models and implement advanced numerical methods to test the functionality of their design ideas. However, during the early stages of the design process, designers use their experience and domain knowledge to

ensure their new ideas are relevant to the design problem at hand. In the literature, experts have been used to evaluate and screen crowdsourced ideas (i.e., human generated ideas) [10]. Likewise, crowdsourcing methods have been implemented to assess the ability of generative computational tools to produce new design ideas [7,8]. However, the functionality of 2D sketch ideas produced by computational tools has not been explored. During the early stages of the design process, rough 2D sketches are typically the primary communication source of ideas [14]. Hence, if computational tools are to co-create new products and solutions alongside designers, their ability to produce not only novel, but also functional ideas, needs to be explored.

The ability to generate creative ideas is an insufficient condition for innovation [15] because decision-makers need to not only generate, but also select creative ideas for innovation to occur. However, decision-makers can be biased, which can have a direct impact on the screening and selection of ideas [16,17]. As designers are increasingly integrating computational tools into the design process, their possible bias towards computer generated ideas needs to be explored. Similarly, their ability to accurately decipher between human created sketches and computer generated sketches needs to be studied. In light of this, the authors of this work present a crowdsourcing method to explore the perceived visual and functional characteristics of 2D design sketches generated by a deep learning generative model. Moreover, this method also allows the authors to explore the ability of raters to distinguish between human and computer generated sketches and their possible bias toward them.

LITERATURE REVIEW

To motivate the current work, the authors explored previous research regarding generative design and crowdsourcing validation methods.

Generative design

Generative design methods have captured the interest of both the design research and industry communities [18–20]. In Chandrasegaran et al. [3], the authors present a review of some of the challenges and future direction for computational support tools used in the product design process [3]. Designers already have several generative design tools in their portfolio (e.g., Siemen's Frustum[1], Autodesk's Nastran[2]), which allows them to iterate through several design combinations and attain feasible solutions for a given problem [14]. These tools take advantage of a mathematical approach that optimizes the layout of a material distribution within a given design domain, known as topology optimization [21–23]. Similarly, data-mining methods have been proposed to evaluate and generate new product ideas [24,25]. Other techniques such as genetic algorithms and procedural modeling have been proposed in the literature as well [26–30]. For example, Huang et al. [31] proposed a method to automatically compute the parameters of procedural models from hand-drawn 2D sketches, which allows designers to explore modifications of their ideas.

Recently, designers have started to integrate deep learning models into their generative design methods. Deep learning models are a class of hierarchical statistical models composed of multiple interconnected layers of nonlinear functions [32]. Design researchers have gained increased interest in deep learning models after studies have shown their potential in image recognition tasks [33,34]. Designers have gained a particular interest in Recurrent Neural Networks (RNN) [35–38] and Generative Adversarial Networks (GAN) [7,8,39]. RNNs are deep learning models that contain multiple interconnected hidden layers. The hidden layers in an RNN are able to use information from their previous state via a recurrent weight layer, which allows them to have a recollection of their previous states [6]. This property makes them suitable for autocorrelated time series data in handwriting and speech recognition tasks [40,41]. GANs are deep learning generative models composed of a generator and a discriminator. The generator is trained to generate new images that are still similar enough to the ground truth images, and that cannot be distinguished by the discriminator. In contrast, the discriminator is trained to discriminate the generated images from the ground truth data [5]. Due to the interaction between the generator and discriminator, these deep generative models are capable of generating designs that are different from the human training dataset (i.e., unique at a pixel level), while still maintaining some degree of similarity. For more details on RNNs and GANs see [5,6].

Deep generative methods have been used to help in the representation of the design space. For example, Burnap et al. [7], trained a deep generative model with a dataset of automotive designs and was able to generate new design ideas that morphed different body types and brands of vehicles. This allowed them to visualize new design ideas and explore the design space. Dosovitskiy et al. [42] trained a deep generative model to generate new 2D images of chairs. Kazi et al. [14] implemented deep generative models into their *DreamSketch* tool. The *DreamSketch* tool takes as input a rough 2D sketch and generates multiple augmented solutions in 3D. Similarly, Lun et al. [43] were able to implement deep learning algorithms to reconstruct 3D shapes from rough 2D sketches. Recently, Chen et al. [36] presented a modification of Ha and Eck's *Sketch-RNN* model [35] that was capable of recognizing and generating 2D sketches from multiple classes. As the authors highlighted, this model has the potential to help in creative tasks [36]. Deep generative methods have also been implemented to increase the veracity of big-data pipelines by generating new images [8]. However, an inherent challenge of these generative methods is that their objective to create new design ideas that still maintain a degree of similarity with the training data used are conflicting and challenging to evaluate. While studies have implemented pixel-level Euclidean distance and structured similarity indices to evaluate these methods, in many cases, these scores do not correlate to visual quality scores given by human raters [44].

[1] https://www.frustum.com
[2] https://www.autodesk.com/products/nastran/overview

Copyright © 2018 by ASME

Crowdsourcing and generative design validation

As a result of the current limitations in the evaluation metrics of generative models, researchers are starting to integrate crowdsourcing methods to evaluate their models. For example, Burnap et al. [7] used a crowdsourcing method to recruit 69 participants and assess the ability of their deep generative model to generate realistic designs. Their results showed that their model was able to generate realistic designs while exploring the design space. Chen et al. [36] conducted a Turing test to compare the capability of 61 human raters and four deep learning models to recognize human vs. computer generated sketches. In their experiment, they tested sketches of object commonly found in nature (i.e., cat, pig, and rabbit). Their results revealed that some of the deep learning models outperformed the human raters in accurately distinguishing between human vs. computer generated sketches. Dering and Tucker [8] used 252 human raters to evaluate the capability of their proposed method to generate new 2D sketches that were recognized to belong to a specific class. Their results indicated that human raters were able to accurately recognize the sketches of certain classes (e.g., bottle, hammer). These studies have analyzed the accuracy of human raters in classifying new images and sketches into specific classes, and not necessary evaluating the sketches' functionality.

Another product design approach that takes advantage of crowdsourcing methods is Mass-Collaborative Product Development (MCPD). MCPD is gaining popularity within the design community as a new design paradigm that decentralizes the product development process [3,9]. MCPD implements crowdsourcing to assist with the generation of new ideas [11]. Research indicates that crowdsourcing methods might constitute a promising paradigm for the product design process [10]. Table 1 shows a summary of existing literature related to deep generative design tools and the implementation of crowdsourcing methods used to evaluate them. Most of the current works focus on evaluating the capability of deep generative models to create new sketches or images that can be classified to belong to a specific category. Though an idea needs to be new and novel in order to be considered creative, it also has to meet its intended functionality and be useful [12,13].

TABLE 1. SUMMARY OF EXISTING WORKS

Reference	*Object Classification evaluation*	*Functionality evaluation*	*Crowdsourcing method*
[14][42]	X		
[7][8] [36][45]	X		X
[46]		X	
This work	X	X	X

During the later stages of the design process, designers create CAD models and implement advanced numerical methods, such as finite element analysis [47], to test the functionality of their design ideas. However, these methods are time-consuming and complex to implement. Researchers have started to explore how deep learning algorithms can be implemented to predict the ability of a 3D artifact to perform a function [46]. During the early stages of the design process, detailed 3D models are not widely available as compared to rough 2D sketches. Free-hand, low-fidelity 2D sketches are typically the primary communication source of ideas, especially in the early phases of the design process [14,48]. During these stages, designers use their experience and domain knowledge to ensure that generated ideas are relevant to the design problem. Hence, experts have been used to evaluate and screen crowdsourced ideas (i.e., human generated ideas) [10]. Similarly, crowds have been used to evaluate the perceptual attributes of new designs [45]. However, the functionality of 2D sketch ideas produced by computational tools has not been explored. If computational tools are to co-create new products and solutions alongside designers, their capability to produce not only novel, but also functional ideas, needs to be explored.

As stated by Rietzschel et al. : "*idea generation is only part of the innovative process, and the availability of creative ideas is a necessary, but insufficient condition for innovation*" [49]. This indicates that even if creative design ideas are generated, they will not advance innovation if they are not selected for further development. Unfortunately, human bias can have a direct impact on the screening and selection of ideas [50]. Studies indicate that decision-makers can experience ownership [51], framing [52], complexity [53], and even creativity biases [16]. Similarly, the human-computer interaction community has recognized that individuals can be biased towards automated systems [54]. This is known as *Automation bias* [55]. One of the factors that contribute to *Automation bias* is the trust given to automated support systems. This trust is the product of the humans' perception of these systems as having superior analytical capabilities [56]. For example, the results by Dzindolet et al. [57] indicate that participants expected an automated support system to outperform the human system in a visual detection task. These studies on *Automation bias* focus on safety and automation aids, and not directly on decision-makers bias towards sketches generated by deep generative design tools. Hence, as designers are increasingly integrating computational tools into the design process, their possible bias towards computer generated ideas and ability to accurately decipher between human and computer generated ideas need to be explored.

As a result of existing knowledge gaps, the authors of this work present a crowdsourcing method to recruit human rater and explore the perceived functionality of 2D design sketches generated via a deep learning generative model. The perceived functionality (i.e., the perception of how likely design sketches will perform a given function) of these computer generated

Copyright © 2018 by ASME

sketches is compared against the perceived functionality of human generated sketches. Additionally, the ability of raters to distinguish between human and computer generated sketches, and their possible bias is explored. In this work, the term 'sketch' is used to mean a low-fidelity, rough 2D drawing representation of an idea.

RESEARCH QUESTIONS

The objective of this work is to explore the perceived visual and functional characteristics of computer generated sketches, compared to human created sketches. In addition, this work explores the possible biases that humans may have towards the functionality of computer generated sketches. Specifically, the hypotheses and research questions (RQ) this work aims to test and address are presented next:

RQ1: How does the perceived functionality of 2D computer generated sketches compare to the functionality of human generated sketches?

RQ2: Are individuals' perceived functionality of 2D sketches biased towards computer generated sketches*?*

RQ3*:* Are individuals capable of accurately distinguishing between 2D human generated sketches and computer generated sketches?

The authors hypothesize that (h_1): the perceived functionality of 2D sketches generated by a deep learning generative model is not significantly different from the perceived functionality of the 2D human generated sketches used to train the model. This hypothesis is founded on research that reveals that generative models can produce new 2D sketches that still maintain a degree of similarity with the 2D sketches used to train the model [7]. Testing this hypothesis will allow the authors to address ***RQ1***. The hypothesis can be mathematically expressed as:

(h_1) h_o: $\overline{PF}_C = \overline{PF}_H$ vs h_a: $\overline{PF}_C \neq \overline{PF}_H$

Where,

$\overline{PF}_C$ is the average perceived functionality of the 2D sketches generated by a deep learning generative model.

$\overline{PF}_H$ is the average perceived functionality of the 2D human generated sketches used to train the deep learning generative model.

Moreover, following ***RQ2***, the authors hypothesize that (h_2): individuals' perceived functionality of 2D computer generated sketches is biased in comparison to their perceived functionality of human generated sketches. That is, individuals will perceive the functionality of the 2D sketches as greater when the sketches are explicitly presented as computer generated (i.e., with a label that says: "*computer generated*," see Fig. 1). The research that indicates that humans perceive automated systems as having superior capabilities [56,57] supports this hypothesis, which can be mathematically expressed as:

(h_2) h_o: $\overline{PF}_C = \overline{PF}_{C*}$ vs h_a: $\overline{PF}_C < \overline{PF}_{C*}$

Where,

$\overline{PF}_{C*}$ is the average perceived functionality of the 2D sketches explicitly presented as computer generated

Finally, since deep learning generative models have been shown to generate human readable 2D sketches that were recognized to belong to a specific category [8], the authors hypothesize that (h_3): individuals would not accurately distinguish between 2D sketches generated by a deep learning generative model and the 2D human generated sketches used to train the model. Testing this hypothesis will allow the authors to answer ***RQ3***. This hypothesis can be mathematically expressed as:

(h_3) h_o: $\overline{PC} = 0.5$ vs h_a: $\overline{PC} \neq 0.5$

Where,

$\overline{PC}$ represents the average accuracy of the raters when classifying the 2D sketches as human or computer generated.

CASE STUDY

In order to address the previous research questions and test the proposed hypotheses, a case study in which 2D boat sketches generated by humans and a deep generative model were presented to raters recruited via a crowdsourcing platform.

Dataset of 2D sketches

For this case study, the *Quick, Draw!* dataset was implemented [58]. This dataset was acquired by Google via the *Quick, Draw!* game[1]. In this game, individuals are asked to draw a specific object within 20 seconds. The objects include but are not limited to: alarm clocks, bats, jackets, rabbits, and boats. For this case study, a total of 132,270 boat sketches were used as a training dataset for the *Sketch-RNN*[2] algorithm presented by Ha and Eck [35]. The algorithm generates new 2D sketches by implementing a Recurrent Neural Network (RNN) based on a Variational AutoEncoder (VAE) framework [59]. Fig. 1 shows some of the human and computer generated boat sketches used in this work.

[1] https://quickdraw.withgoogle.com
[2] https://magenta.tensorflow.org/sketch-rnn-demo

Copyright © 2018 by ASME

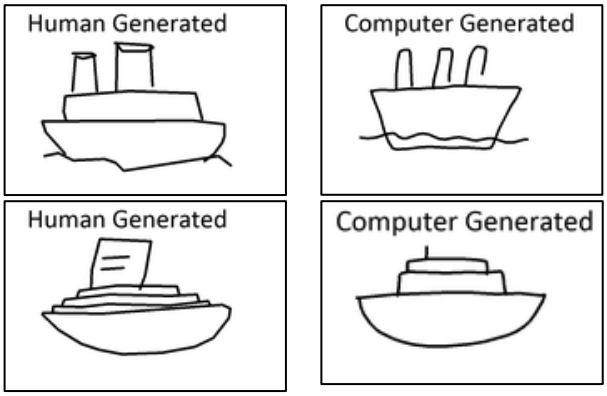

FIGURE 1. EXAMPLE OF HUMAN AND COMPUTER GENERATED BOAT SKETCHES

Crowdsourcing

In this work, Amazon Mechanical Turk[1] (AMT) was used as the crowdsourcing platform to recruit raters. AMT has been previously used to evaluate the output of deep generative models [7,8]. Moreover, AMT has established itself as a valuable tool for behavioral research since studies have found no significant differences in the response consistency between internet users and laboratory participants [60,61]. Compared to other crowdsourcing platforms, AMT provides the benefits of (i) low cost, (ii) large rater pool access, and (iii) large rater pool diversity [61,62]. In this work, a total of 983 raters were recruited to evaluate a set of 100 boat sketches. This set of randomly selected sketches was composed of 50 human, and 50 computer generated boat sketches. The raters were compensated $0.20 for their participation. In average, the raters spend 869.1 seconds to complete the experiment. Only raters with a 90% satisfaction rate (i.e., 90% of the questionnaires completed by the rater have been accepted) were allowed to participate in this experiment. Similarly, participants were only allowed to take the questionnaires once. These constraints were set following the guidelines for conducting research through AMT [61]. In addition, other quality assurance controls were set in place (e.g., reading time, control questions), which are explained in the following sections.

Questionnaire

For this work, a between-subject experiment was implemented to test the effects that labeling the sketches as either human or computer generated had on participants' response (i.e., ***RQ2***). Once the participants consented to be part of the experiment, they were randomly assigned to one of 25 conditions of the questionnaire. Each condition contained questions regarding a unique set of eight different 2D boat sketches. Each set of images was composed of: (i) 2 human generated and (ii) 2 computer generated sketches without a label, as well as (iii) 2 human generated and (iv) 2 computer generated sketches with a label (see Fig. 1). The instructions provided to the participants on how to complete the questionnaire are shown in Fig. 2. For quality control purposes, the response of participants that spent less than 10 seconds on the instruction page was not considered for analysis since it is assumed that they did not read the instructions carefully. Subsequently, participants were introduced to the five questions shown in Table 2. A 7-point liker scale was used for questions one, two, four, and five. Under questions two and five, the 2D environment shown in Fig. 2 was presented to participants. For questions one and four, the same 2D environment, without the boat sketch and force representation, was presented. Questions one and two allow the authors to address the research question ***RQ1***; while questions four and five address the research question ***RQ2***. Finally, question three addresses the research question ***RQ3***.

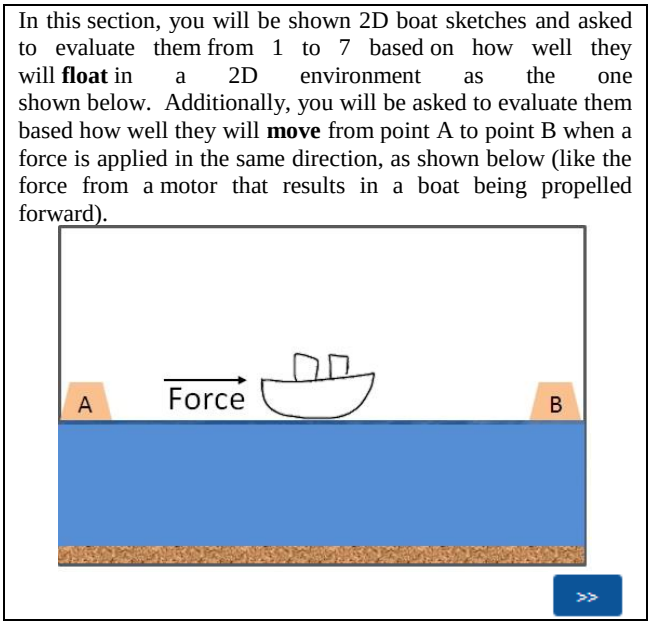

FIGURE 2. INSTRUCTION PAGE FROM QUESTIONNAIRE

TABLE 2. QUESTIONS PRESENTED TO PARTICIPANTS

Q1: Please evaluate the following boat sketches based on how well they will **float** in the 2D environment shown below.
Q2: Please evaluate the following boat sketches based on how well they will **move** from point A (left) to point B (right) when a force is applied in the 2D environment as shown below.
Q3: Please classify the following sketches as *human-generated* (drawn by a person) or *computer-generated* (drawn by a computer).
Q4: Please evaluate the following computer and human generated boat sketches based on how well they will **float** in the 2D environment shown below.
Q5: Please evaluate the following computer and human generated boat sketches based on how well they will **move** from point A (left) to point B (right) when a force is applied in the 2D environment as shown below.

[1] https://www.mturk.com

Copyright © 2018 by ASME

On questions four and five, participants were shown 2 human generated and 2 computer generated boat sketches with their respective labels as shown in Fig. 1. While for questions one, two, and three a different set of 2 human and 2 computer generated sketches without labels were presented. For each question one through five, the 2D boat sketches were presented in a random order. That is, the first sketch on question one might have been the fourth sketch on question two. Furthermore, in question three, there was an additional image that explicitly asked participants to select the "*human-generated*" option. This was for quality control purposes. Participants that did not correctly answer this control question were excluded from the analysis. This was done to filter out participants that "clicked through" the questionnaire.

RESULTS AND DISCUSSION

After filtering the participants based on their response to the quality control question and time spent reading the instructions, the responses of only 619 participants (49.6% females) were used in this work. The age of the participants ranged from 18 to 76 years of age (μ= 35.95, σ=11.44). In this work, an alpha level of 0.01 is used to test the statistical significance of the results. Table 3 shows the summary statistics for the participants' response to *Q1*, *Q2*, *Q4*, and *Q5*.

TABLE 3. SUMMARY STATISTICS FOR *Q1, Q2, Q4,* AND *Q5*

	Computer generated			Human generated		
	μ	***median***	***σ***	***μ***	***median***	***σ***
Q1	5.13	6	1.64	4.41	5	1.94
Q2	5.03	5	1.62	4.34	5	1.83
Q4	4.99	5	1.6	4.13	4	1.91
Q5	4.97	5	1.59	4.13	4	1.8

Inter-rater reliability

The inter-rater reliability of participants' response to the conditions of the questionnaire was assessed via Cronbach's alpha [63]. Fig. 3 shows the Cronbach's alpha distribution for the different conditions of the questionnaire (i.e., sets of images). The results indicate that on average, participants' responses had a Cronbach's alpha of 0.828 (σ= 0.053). The results reveal that participants' responses were more consistent when evaluating certain sets of images (i.e., conditions) (Cronbach's alpha range= [0.739-0.916]). Overall, the Cronbach's alpha indicates an acceptable inter-rater reliability (i.e., >0.7) [64]. This indicates that in general participants showed consensus in their responses.

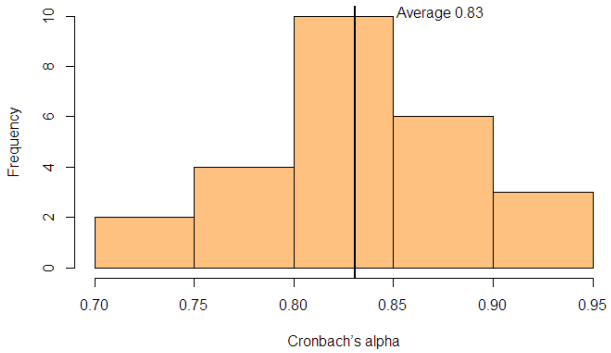

FIGURE 3. CRONBACH'S ALPHA DISTRIBUTION OF THE CONDITIONS OF THE QUESTIONNAIRE

RQ1: Perceived functionality of sketches

Figure 4 shows a plot of the participants' responses to questions one (Q1) and questions two (Q2). The results indicate that for Q1, the average response of the computer generated sketches was significantly greater than the average response of the human generated sketches (*t-value*: 10.02, *p-value*<0.001). Similarly, for Q2 the average response for the computer generated sketches was significantly greater than the average response of the human generated sketches (*t-value:* 9.88, *p-value*<0.001). This reveals that the participants perceived the sketches generated by the deep generative model as more likely to float and move compared to the human generated sketches. Hence, indicating that the perceived functionality of 2D computer generated sketches was greater than the functionality of human generated sketches. Fig. 5 shows the three human and computer generated sketches that on average were perceived as the most likely and less likely to move and float. These sketches do not show any major pattern that may provide evidence of their perceived functionally. Nonetheless, the test results provide enough evidence to reject the null hypothesis number one (h_1). Previous research has revealed that deep generative models can be implemented to automatically generate new design ideas. However, these new design ideas have to meet its intended functionality in order to be considered creative. The findings of this work show the potential of deep generative design tools and their ability to generate ideas that are perceived as functional. These results indicate that deep generative design tools could potentially assist in creative tasks such as ideation.

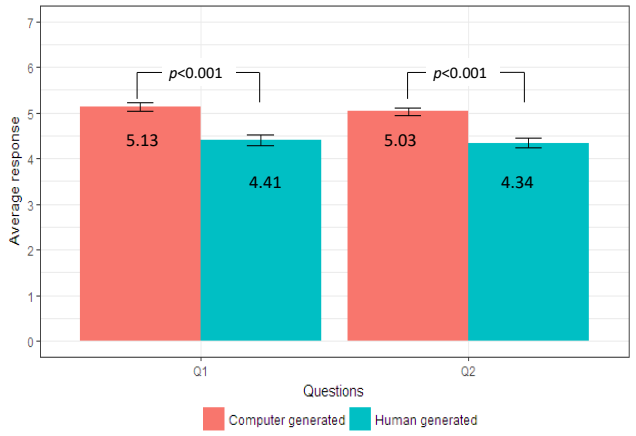

FIGURE 4. SUMMARY OF PARTICIPANTS' RESPONSE TO O1 AND O2

Copyright © 2018 by ASME

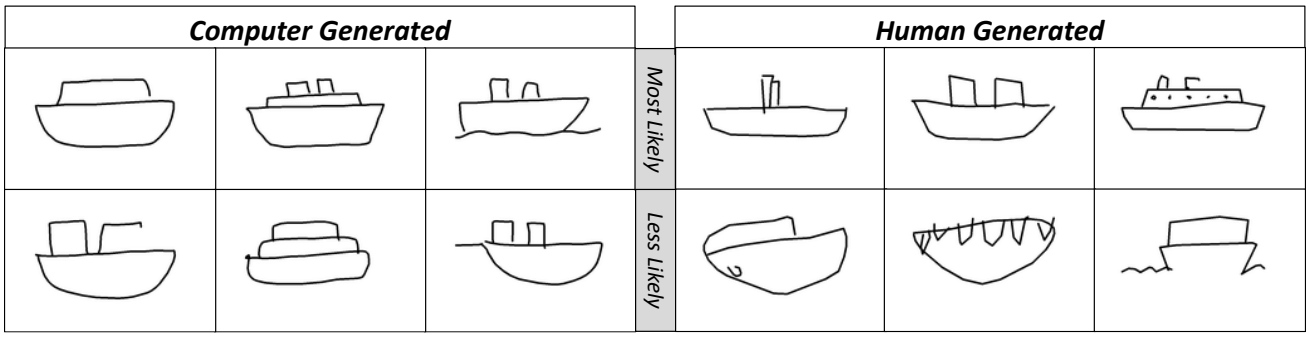

FIGURE 5. SKETCHES THAT ON AVERAGE WERE PERCEIVED AS MOST LIKELY AND LESS LIKELY TO MOVE AND FLOAT.

RQ2: Perceived functionality bias

Figure 6 shows a plot that compares the participants' responses to question one (Q1) and four (Q4), as well as question two (Q2) and five (Q5). As stated previously, Q1 and Q2 presented sketches without labels; while Q4 and Q5 presented sketches with their respective labels. The results indicate that the average response on Q1 and Q4 was not significantly different for the computer generated sketches (*t-value:* 2.017, *p-value*=0.043). Nonetheless, for the human generated sketches the average response on Q1 and Q4 was significantly different (*t-value:* 3.434, *p-value*<0.001). Similarly, for Q2 and Q5, the average response was not significantly different for the computer generated sketches (*t-value:* 1.015, *p-value*=0.31). The difference was only significant for the human generated sketches (*t-value:* 2.944, *p-value*=0.003). This reveals that the participants' response regarding the perceived functionality of the computer generated sketches did not change significantly when explicitly presented as computer generated. In contrast, participants' perceived functionality of human generated sketches did change significantly. These results do not provide enough evidence to reject the null hypothesis number two (h_2). They indicate that participants' perceived functionality of the computer generated sketches was not biased by explicitly presenting them as computer generated. However, that was not the case for the human generated sketches. While previous studies have shown that individuals can be subject to *Automation bias* [56,57], they did not explore the possible bias decision-makers may have towards the functionality of computer generated sketches. In this work, the results reveal that the perceived functionality of computer generated sketches was not biased by the fact that they were explicitly presented as computer generated. This indicates that during the evaluation and screening process of new design sketches, the individuals' perceived functionality of computer generated sketches would not be biased by the fact they were not generated by a human.

TABLE 4. CONFUSION MATRIX OF SKETCHES CLASSIFICATION

		Ground truth		
		Computer	*Human*	**Total**
Prediction	*Computer*	264	269	533 (22%)
	Human	974	969	1943 (78%)
	Total	1238 (50%)	1238 (50%)	**2476 (100%)**

RQ3: Distinguishing between human and computer generated sketches

Table 4 presented the confusion matrix for the participants' response to question three (i.e., classification of sketches as human or computer generated). The results indicate that participants were capable of achieving a classification accuracy of only 49.8% (95% CI= [47.81%-51.79%]). This accuracy was not significantly different from random chance (i.e., 50%) (*p-value*=0.588). The 2-sample proportion test indicates that the proportion of correctly classified and incorrectly classified sketches was not significantly different for both computer generated (proportion=0.213), and human generated sketches (proportion=0.217) (χ^2=0.038, *p-value*=0.844). These results do not provide enough evidence to reject the null hypothesis number three (h_3). They indicate that individuals' cannot accurately distinguish between 2D human generated and

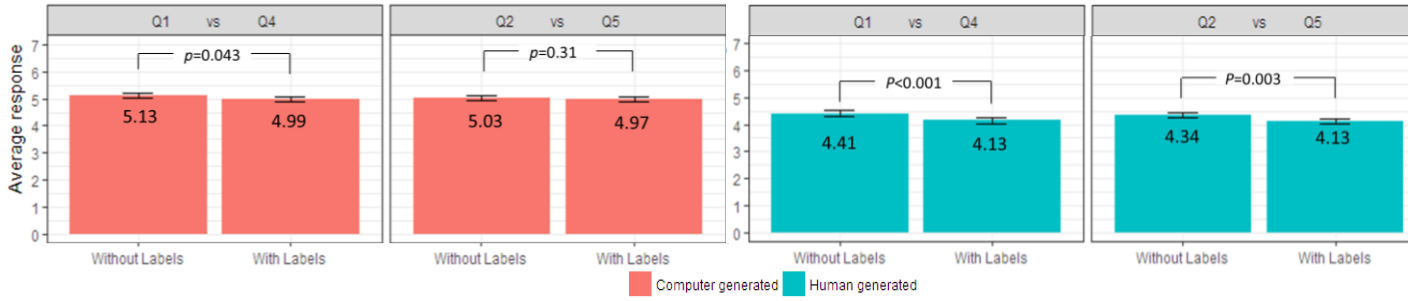

FIGURE 6. COMPARISON OF PARTICIPANTS' RESPONSE TO Q1 vs. Q4 AND Q2 vs.Q5

Copyright © 2018 by ASME

computer generated sketches. Moreover, the results show that participants tended to classify sketches as human generated more frequently than computer generated. Along with the previous findings (i.e., h_2), these results reveal that during the evaluation and screening process of new design sketches it may be helpful to avoid classifying the sketches as either human or computer. If decision-makers try to decipher the origin of a sketch, they will likely classify it as human generated, which may bias the evaluation and selection process.

CONCLUSION AND FUTURE WORK

Recent advancements in technology have allowed designers to implement computational tools to automatically generate large pools of new design ideas. Nonetheless, an idea needs to meet its intended functionality and be useful in order to be considered creative. Therefore, if computational tools are to co-create ideas and solutions alongside designers, their capability to produce not only novel, but functional ideas, needs to be explored. The ability to generate creative ideas is an insufficient condition for innovation because decision-makers need to not only generate, but also select creative ideas for innovation to occur. Hence, decision-makers' bias towards computer generated ideas and the ability to distinguish them need to be explored. In order to fill this knowledge gap, this work implemented a crowdsourcing method to explore the perceived functionality of 2D design sketches generated by a deep generative model. Additionally, the ability of raters to distinguish between human and computer generated sketches, and their bias toward them was explored. In summary, the results of this work indicated that:

1. Computer generated sketches were perceived as more functional than the human generated sketches.
2. The perceived functionality of computer generated sketches was not affected by explicitly presenting them as computer generated.
3. The perceived functionality of human generated sketches was affected by explicitly presenting them as human generated.
4. Individuals were not able to accurately distinguish between the human and computer generated sketches.

The results revealed that participants perceived the 2D boat sketches generated by the deep generative model as more likely to float and move than the human generated sketches. These findings provide evidence that deep generative design tools are able to generate ideas that are perceived as functional. As these tools become more efficient at creating novel and functional ideas, researchers argue they will foster the designers' creativity and help in creative tasks [4,36]. Moreover, the results of this work revealed that participants' perceived functionality of computer generated sketches was not biased by explicitly presenting them as computer generated. In contrast, the results showed that participants' perceived functionality of human generated sketches was significantly less on the sketches that were presented with labels than those presented without labels. The human-computer interaction community has recognized that *Automation bias* can affect individuals' perception of automated system's capabilities. However, the results of this work did not provide enough evidence to support participants' bias towards the functionality of 2D computer generated sketches. Finally, the results revealed that participants were not able to accurately classify the 2D sketches as either human or computer generated. These findings are in line with previous studies and reveal the capability of deep learning generative designs tools to generate new sketch ideas that are indistinguishable from human generated sketches [36]. However, the results also indicated that if decision-makers try to decipher the origin of a sketch, they will likely classify it as human generated. This inaccurate classification may impact the evaluation and selection process.

While this work provides evidence that supports the capabilities of deep generative designs tools and their potential to assist designers in creative tasks, several limitations exist. For example, although the results indicate that participants' perceived functionality of the computer generated sketches was statistically significantly greater than the human generated sketches, the practical significance of these differences (i.e., $Q1\Delta = 0.72$, $Q2\Delta = 0.69$) needs to be explored. Moreover, the effect of presenting the sketches with and without labels on participants' perceived functionality cannot be disentangled from a possible order effect. This is because all the questions that contained sketches with labels (i.e., Q4 and Q5) were presented after the questions that contained sketches without labels (i.e., Q1 and Q2). Future work should explore how an individual's attributes (e.g., gender, age, education level) impact his/her perceived functionality of computer generated ideas since studies have shown that the personality traits, risk attitudes, and gender of decision-makers can influence their selection of ideas [65–67]. Similarly, while the sketches in Fig. 5 do not show any major pattern that may provide evidence of their perceived functionally, future work should explore other methods to analyze the correlation between the visual characteristics and perceived functionally of the sketches.

ACKNOWLEDGMENT

This research is funded in part by DARPA HR0011-18-2-0008 and NSF NRI # 1527148. Any opinions, findings, or conclusions found in this paper are those of the authors and do not necessarily reflect the views of the sponsors. The authors would like to acknowledge Raj Desai and Albert Wilson for their contributions to this work.

REFERENCES

[1] National Academy of Engineering, 2004, The Engineer of 2020: Visions of Engineering in the New Century, Washington, DC.

Copyright © 2018 by ASME

[2] Pahl, G., Beitz, W., Feldhusen, J., and Grote, K.-H., 2007, *Engineering design: a systematic approach*, Springer, (2), p. 617.
[3] Chandrasegaran, S. K., Ramani, K., Sriram, R. D., Horváth, I., Bernard, A., Harik, R. F., and Gao, W., 2013, "The evolution, challenges, and future of knowledge representation in product design systems," *Comput. Des.*, **45**(2), pp. 204–228.
[4] Liapis, A., Yannakakis, G. N., Alexopoulos, C., and Lopes, P., 2016, "Can Computers Foster Human User's Creativity? Theory and Practice of Mixed-Initiative Co-Creativity," *Digit. Cult. Educ.*, **8**(2), pp. 136–153.
[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., 2014, "Generative Adversarial Nets," *Adv. Neural Inf. Process. Syst.* 27, pp. 2672–2680.
[6] Boden, M., 2001, "A guide to recurrent neural networks and backpropagation," *Electr. Eng.*, (2), pp. 1–10.
[7] Burnap, A., Lui, Y., Pan, Y., Lee, H., Gonzalez, R., and Papalambors, P., 2016, "Estimating and Exploring the Product Form Design Space Using Deep Generative Models," *Des. Autom. Conf.*, pp. 1–13.
[8] Dering, M. L., and Tucker, C. S., 2017, "Generative Adversarial Networks for Increasing the Veracity of Big Data," *IEEE Int. Conf. on Big Data* (BIGDATA), pp. 2513–2520.
[9] Le, Q., and Panchal, J. H., 2011, "Modeling the Effect of Product Architecture on Mass-Collaborative Processes," *J. Comput. Inf. Sci. Eng.*, **11**(1), p. 11003.
[10] Poetz, M. K., and Schreier, M., 2012, "The value of crowdsourcing: Can users really compete with professionals in generating new product ideas?," *J. Prod. Innov. Manag.*, **29**(2), pp. 245–256.
[11] Sun, L., Xiang, W., Chen, S., and Yang, Z., 2015, "Collaborative sketching in crowdsourcing design: a new method for idea generation," *Int. J. Technol. Des. Educ.*, **25**(3), pp. 409–427.
[12] Boden, M. A., 2003, *The creative mind: Myths and mechanisms*, Second edition.
[13] Shah, J., 2003, "Metrics for measuring ideation effectiveness," *Des. Stud.*, **24**(2), pp. 111–134.
[14] Kazi, R. H., Grossman, T., Cheong, H., Hashemi, A., and Fitzmaurice, G., 2017, "DreamSketch: Early Stage 3D Design Explorations with Sketching and Generative Design," *Proc. 30th Annu. ACM Symp. User Interface Softw. Technol.*, pp. 401–414.
[15] Rietzschel, E. F., Nijstad, B. A., and Stroebe, W., 2006, "Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection," *J. Exp. Soc. Psychol.*, **42**(2), pp. 244–251.
[16] Mueller, J. S., Melwani, S., and Goncalo, J. A., 2012, "The bias against creativity: Why people desire but reject creative ideas," *Psychol. Sci.*, **23**(1), pp. 13–17.
[17] Toh, C. A., Patel, A. H., Strohmetz, A. A., and Miller, S. R., 2015, "My Idea Is Best! Ownership Bias and its Influence on Engineering Concept Selection," P*roc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, p. V007T06A005-V007T06A005.
[18] Orsborn, S., Cagan, J., and Boatwright, P., 2009, "Quantifying Aesthetic Form Preference in a Utility Function," *J. Mech. Des.*, **131**(6), p. 61001.
[19] Reid, T. N., Gonzalez, R. D., and Papalambros, P. Y., 2010, "Quantification of Perceived Environmental Friendliness for Vehicle Silhouette Design," *J. Mech. Des.*, **132**(10), p. 101010.
[20] Kokai, I., Finger, J., Smith, R., Pawlicki, R., and Vetter, T., 2007, "Example-Based Conceptual Styling Framework for Automotive Shapes," P*roc. 4th Eurographics Workshop on Sketch based interfaces Model.*, **1**, pp. 37–44.
[21] Bendsøe, M. P., and Sigmund, O., 2003, Topology optimization: theory, methods, and applications.
[22] Wang, M. Y., Wang, X., and Guo, D., 2003, "A level set method for structural topology optimization," C*omput. Methods Appl. Mech. Eng.*, **192**(1–2), pp. 227–246.
[23] Ulu, N. G., and Kara, L. B., 2015, "DMS2015-33: Generative interface structure design for supporting existing objects," *J. Vis. Lang. Comput.*, **31**, pp. 171–183.
[24] Kang, S. W., and Tucker, C. S., 2015, "Automated Concept Generation Based On Function-Form Synthesis," *Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, pp. 1–11.
[25] Lopez, C., Zheng, X., and Miller, S. R. ., 2017, "Linking Creativity Measurements to Product Market Favorability: A Data-Mining Approach.," *Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, p. V02AT03A013-V02AT03A013.
[26] Yannou, B., Dihlmann, M., and Awedikian, R., 2008, "Evolutive design of car silhouettes.," *Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, pp. 15–24.
[27] Poirson, E., Petiot, J.-F., Boivin, L., and Blumenthal, D., 2013, "Eliciting User Perceptions Using Assessment Tests Based on an Interactive Genetic Algorithm," *J. Mech. Des.*, **135**(3), p. 31004.
[28] Chan, K. Y., Yuen, K. K. F., Palade, V., and Kwong, C. K., 2014, "Computational intelligence techniques for new product development," *Neurocomputing*, **142**, pp. 1–3.
[29] Parish, Y. I. H., and Müller, P., 2001, "Procedural modeling of cities," *Proc. of the 28th Annu. Conf. on Comp. Graphics and Interactive Tech.* , pp. 301–308.
[30] Schwarz, M., and Wonka, P., 2014, "Procedural Design of Exterior Lighting for Buildings with Complex Constraints," *ACM Trans. Graph.*, **33**(5), p.

Copyright © 2018 by ASME

166:1--166:16.
[31] Huang, H., Kalogerakis, E., Yumer, E., and Mech, R., 2017, "Shape Synthesis from Sketches via Procedural Models and Convolutional Networks," *IEEE Trans. Vis. Comput. Graph.*, **23**(8), pp. 2003–2013.
[32] Schmidhuber, J., 2015, "Deep Learning in neural networks: An overview," *Neural Networks*, **61**, pp. 85–117.
[33] Krizhevsky, A., Sutskever, I., and Geoffrey E., H., 2012, "ImageNet Classification with Deep Convolutional Neural Networks," A*dv. Neural Inf. Process. Syst.* 25, pp. 1–9.
[34] Simonyan, K., and Zisserman, A., 2015, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Int. Conf. Learn. Represent.*, pp. 1–14.
[35] Ha, D., and Eck, D., 2017, "A neural representation of sketch drawings.," *arXiv*, preprint arXiv:1704.03477.
[36] Chen, Y., Tu, S., Yi, Y., and Xu, L., 2017, "Sketch-pix2seq: a Model to Generate Sketches of Multiple Categories.," *arXiv*, preprint arXiv:1709.04121.
[37] Junokas, M. J., Kohlburn, G., Kumar, S., Lane, B., Fu, W. T., and Lindgren, R., 2017, "What You Sketch Is What You Get: 3D Sketching using Multi-View Deep Volumetric Prediction," *CEUR Workshop Proc.*, **1828**(1), pp. 89–93.
[38] Choy, C. B., Xu, D., Gwak, J., Chen, K., and Savarese, S., 2016, "3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction," *in European Conf. on Computer Vision* , pp. 628–644.
[39] Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L., 2017, "Learning Representations and Generative Models for 3D Point Clouds," *arXiv* preprint arXiv:1707.02392.
[40] Graves, a, Mohamed, A., and Hinton, G., 2013, "Speech recognition with deep recurrent neural networks," *Icassp*, (3), pp. 6645–6649.
[41] Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J., 2009, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, **31**(5), pp. 855–868.
[42] Dosovitskiy, A., Springenberg, J. T., Tatarchenko, M., and Brox, T., 2017, "Learning to Generate Chairs, Tables and Cars with Convolutional Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**(4), pp. 692–705.
[43] Lun, Z., Gadelha, M., Kalogerakis, E., Maji, S., and Wang, R., 2017, "3D Shape Reconstruction from Sketches via Multi-view Convolutional Networks," *arXiv*, prepr. arXiv:1707.06375.
[44] Theis, L., Oord, A. V. D., & Bethge, M., 2016, "A note on the evaluation of generative models.," *Int. Conf. on Learning Representations*, pp. 1–9.
[45] Ren, Y., Burnap, A., Papalambros, P., 2013, "Quantification of perceptual design attributes using a crowd," *Proc. of the 19th Int. Conf. on Eng. Design*, pp. 19–22.
[46] Dering, M., and Tucker, C., 2017, "A Convolutional Neural Network Model for Predicting a Product?s Function, Given Its Form," *J. Mech. Des.*, pp. 1–29.
[47] Zeng, T., Wu, L., and Guo, L., 2004, "Mechanical analysis of 3D braided composites: a finite element model," *Compos. Struct.*, **64**(3–4), pp. 399–404.
[48] Buxton, B., 2007, *Sketching User Experiences: Getting the Design Right and the Right Design*.
[49] Rietzschel, E. F., Nijstad, B. A., and Stroebe, W., 2010, "The selection of creative ideas after individual idea generation: Choosing between creativity and impact," *Br. J. Psychol.*, **101**(1), pp. 47–68.
[50] Toh, C. A., Miele, L. M., and Miller, S. R., 2016, "Which One Should I Pick ? Concept Selection in Engineering Design Industry," *Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, pp. 1–10.
[51] Onarheim, B., and Christensen, B. T., 2012, "Distributed idea screening in stage-gate development processes," *J. Eng. Des.*, **23**(9), pp. 660–673.
[52] Levin, I. P., Schneider, S. L., and Gaeth, G. J., 1998, "All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects," *Organ. Behav. Hum. Decis. Process.*, **76**(2), pp. 149–188.
[53] Cox, D., and Cox, A. D., 2002, "Beyond first impressions: The effects of repeated exposure on consumer liking of visually complex and simple product designs," *J. Acad. Mark. Sci.*, **30**(2), pp. 119–130.
[54] Parasuraman, R., and Manzey, D. H., 2010, "Complacency and bias in human use of automation: An attentional integration," *Hum. Factors*, **52**(3), pp. 381–410.
[55] Mosier, K. L., and Skitka, L. J., 1996, "Human decision makers and automated decision aids: Made for each other?," *Parasuraman Raja Ed*, pp. 201–220.
[56] Lee, J. D., and See, K. A., 2004, "Trust in Automation: Designing for Appropriate Reliance," *Hum. Factors J. Hum. Factors Ergon. Soc.*, **46**(1), pp. 50–80.
[57] Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A., 2002, "The Perceived Utility of Human and Automated Aids in a Visual Detection Task," *Hum. Factors J. Hum. Factors Ergon. Soc.*, **44**(1), pp. 79–94.
[58] Jongejan, J., Rowley, H., Kawashima, T., Kim, J., and Fox-Gieg., N., 2016, "The Quick, Draw! - A.I. Experiment.," https://quickdraw.withgoogle.com/.
[59] Kingma, D. P., and Welling, M., 2013, "Auto-Encoding Variational Bayes," *arXiv*, Prepr. arXiv1312.6114.
[60] Buchanan, T., 2000, "Psychological Experiments on the Internet," *Psychol. Exp. Internet*, pp. 121–140.

Copyright © 2018 by ASME

[61] Mason, W., and Suri, S., 2012, "Conducting behavioral research on Amazon's Mechanical Turk," *Behav. Res. Methods*, **44**(1), pp. 1–23.
[62] Paolacci, G., Chandler, J., and Ipeirotis, P., 2010, "Running experiments on amazon mechanical turk," *Judgm. Decis. Mak.*, **5**(5), pp. 411–419.
[63] Bland, J. M., and Altman, D. ., 1997, "Statistics notes: Cronbach's alpha.," *Br. Med. J.*, **314**, p. 572.
[64] Cortina, J. M., 1993, "What is coefficient alpha? An examination of theory and applications.," *J. Appl. Psychol.*, **78**(1), pp. 98–104.
[65] Kichuk, S. L., and Wiesner, W. H., 1997, "The big five personality factors and team performance: implications for selecting successful product design teams," *J. Eng. Technol. Manag.*, **14**(3–4), pp. 195–221.
[66] Toh, C. A., Strohmetz, A. A., and Miller, S. R., 2016, "The Effects of Gender and Idea Goodness on Ownership Bias in Engineering Design Education," *J. Mech. Des.*, **138**(10), p. 101105.
[67] Toh, C. A., and Miller, S. R., 2014, "The role of individual risk attitudes on the selection of creative concepts in engineering design.," *Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, pp. 1–10.

Copyright © 2018 by ASME