The Relationship Between Design Outcomes and Mental States During Ideation

Using electroencephalography (EEG) to predict design outcomes could be used in many applications as it facilitates the correlation of engagement and cognitive workload with ideation effectiveness. It also establishes a basis for the connection between EEG measurements and common constructs in engineering design research. In this paper, we propose a support vector machine (SVM)-based prediction model for design outcomes using EEG metrics and some demographic factors as predictors. We trained and validated the model with more than 100 concepts, and then evaluated the relationship between EEG data and concept-level measures of novelty, quality, and elaboration. The results characterize the combination of engagement and workload that is correlated with good design outcomes. Findings also suggest that EEG technologies can be used to partially replace or augment traditional ideation metrics and to improve the efficacy of ideation research. [DOI: 10.1115/1.4036131]

1 Introduction

Almost all measures of early design processes approximate cognition in some way by using qualitative or quantitative measures of participant behavior [1]. Examples include self-efficacy surveys, protocol analyses, think-aloud protocols, among others. These methods vary in their level of intrusiveness, ability to accurately measure the desired construct, and when they are conducted relative to the actual task to be completed by the participant [2].

The ideal research method for the evaluation of design cognition would be able to directly measure thoughts in real time without intrusiveness or bias. Because no such method exists, current best practice triangulates a desired construct with multiple research methods or uses a commonly accepted method with an established construct. When new research methods become available, it is necessary to triangulate them with existing methods in order to make proper comparisons.

Some promising additions to the existing research methods are psychophysiological methods. These methods record physiological responses in real-time and are correlated with known psychological constructs. While different psychophysiological methods have different levels of intrusiveness, they all produce much richer, more objective datasets in shorter time frames than comparable qualitative methods. While there are also definite drawbacks to using these methods, they are considered promising in terms of augmenting existing methods for design research.

This paper specifically explores how electroencephalograms (EEG) correspond to traditional ideation outcome metrics such as novelty, quality, and elaboration. EEGs measure the electrical potentials that naturally occur during thought. These potentials change with time based on the activity of the brain at a given moment. This activity is strongly correlated with specific thinking or mental states. We evaluate the relationship between EEG data and design outcomes in order to identify a model of ideation effectiveness that relies on EEG measurements. In general, we attempt to answer the following research questions:

(1) What is the relationship between EEG metrics and traditional ideation metrics used in assessing design outcomes?

(2) How do designers’ mental states (as measured by EEG metrics) influence design outcomes?

While there is promise for psychophysiological methods, they have been tested mostly in divergent-thinking ideation tasks [3] and rarely with problem-solving ideation tasks. This difference is important because the construct for creativity and measure for success differ for the two types of tasks [4]. It is unknown how psychophysiological methods correlate with typical ideation metrics used in design research. To the best of our knowledge, this is the first paper that investigates the relationship between EEG data and design outcomes in a problem-solving ideation task. The results of this paper will facilitate the future use and interpretation of EEG data in design task studies.

In Sec. 2, we will briefly review traditional ideation research methods and some of the relationships between them. We will also review psychophysiological research methods, some initial findings on ideation, and typical EEG metrics. In Sec. 3, we will describe the experimental setup, the design task, and the participants. The EEG metrics used in this paper, the development of evaluation metrics, and surveys will be described in Sec. 4, while in Sec. 5 we will illustrate how we gathered and processed the EEG data. SVM classifiers will be trained and validated in Sec. 6 in order to identify the correlation between EEG measurements and ideation metrics. In Sec. 7, we will summarize the findings of this study.

2 Background

First we will review the traditional approaches to studying ideation and describe the metrics we used for the study. Second, we will review the use of psychophysiological methods in design and describe how they may be applied to ideation research. Last, we will introduce the study approach and the development of the research hypotheses.

2.1 Traditional Ideation Research Methods and Metrics.

Traditionally, engineering studies use external behavior to measure cognition during ideation [1]. Most of these approaches are qualitative, although a few are quantitative. Qualitative analyses usually occur after the ideation session, though some occur during the actual session. Post hoc methods include judging the creativity of ideas using a purely subjective rating [5,6], the function–behavior–structure (FBS) protocol [7], and linkography...
Examples of methods performed during the ideation task include think-aloud protocols, task load indexes [9], and design flow protocols [10]. Quantitative measures are rare, but usually involve simple counts of design outcomes. The most common of these is fluency, which counts the number of ideas [11].

A very common set of metrics for ideation are fluency, elaboration, novelty, and quality [1]; they are derived using alternative uses [12] and Torrance [13] tests. Dean et al. examined 90 studies and found that novelty is the most commonly used metric, but quality and creativity are also common constructs [14]. These metrics were popularized in the design research community by Shah et al. [11], who recommended using fluency, novelty, quality, and variety. A key feature of Shah’s approach is that in order to calculate novelty and variety, it requires identifying the functions embedded in each concept. Shah’s metrics have been updated several times by other researchers. For example, Nelson et al. proposed changes in how to calculate novelty and combined variety and fluency into a single metric called the “quality of design space exploration” [15]. For the purpose of this paper, we define creativity in problem solving as both novelty and quality. We define a successful concept as one that is both novel and meets the requirements of the problem. We will measure elaboration as well because it is known to decrease novelty and quality for divergent-thinking tasks [16].

Variations of the Torrance and other metrics exist and have been studied in detail. Examples of these metrics include the Creative Product Semantic Scale [17], the Consensual Assessment Technique [18], the Student Product Assessment Form [19], and scoring formulas that combine measures into a single score (e.g., see Oman and Tumer [20]). While many of these scales use a measure not proposed by Torrance, everyone uses either novelty or quality in some form [21]. Some work has also been done to combine existing constructs into a universal set of measures [14].

One important difference between many of these competing metrics is whether they are evaluated at the concept level or whether they are broken down into features and evaluated at a feature level. Although evaluating concepts at the feature level is much more expensive than concept-level metrics, it has the advantage of a higher inter-rater reliability [22]. On the other hand, feature-level metrics also strongly depend on the interpretation of features by raters. Since creativity is subjective, feature-level metrics may constrain raters to a set of rules that prevent them from accurately measuring creativity [23]. Most work focused on comparing the metrics in the form of literature reviews, and only few studies have compared feature-level and concept-level metrics. These studies have found that the two approaches have similar results [24,25], suggesting that both approaches work. Therefore, in our opinion the less expensive approach is to be preferred.

2.2 Psychophysiological Research Methods. Traditional research methods for the investigation of ideation include qualitative methods, such as protocol analysis, and some quantitative methods such as surveys. Only recently have psychophysiological methods been used to study ideation based on the fact that the quantitative results can be correlated with qualitative constructs. Electroencephalograms (EEG) and functional magnetic resonance imaging (fMRI) are two common and well-established psychophysiological methods used in research. EEG is used to measure brain waves that are associated with different mental/cognitive states. Researchers can record them at very fine temporal resolutions in order to detect changes in brain activity caused by different events or activities. Functional magnetic resonance imaging is used to identify the part of the brain that is activated during a task, and can also be used to identify cognitive states.

There are also a number of less-invasive psychophysiological methods. Galvanic skin response (GSR, closely related to electrodermal activity orEDA) measures skin conductance and approximates emotional arousal. Heart rate, facial recognition software, or other methods can be used to identify emotional valence (i.e., positive or negative). Galvanic skin response can also be paired with these other methods to infer emotional states.

One of the disadvantages of traditional research methods is that they are external to the mind of the participant because the participant or the researcher has to attribute an artifact or event to a cognitive state. Additionally, the most popular methods are conducted after the ideation session has taken place. Although these post hoc methods avoid interrupting the thinking of a participant, they require either making broad interpretations of the session or guessing what the participant was thinking at a particular moment, which reduces reliability [2]. This limitation cannot be overcome, even when participants watch a video of the session and describe what they believe they were thinking at a particular moment, because research shows that it is often difficult for individuals to accurately recall their higher-order cognitive processes [26]. Other methods, such as think-aloud protocols or pausing the session to take a questionnaire, interrupt the thinking process and could negatively impact the results [2].

Psychophysiological methods may help address the disadvantage of traditional ideation research methods. Galvanic skin response is only mildly intrusive and can approximate emotion or cognition at each moment during the task. Another advantage of using fMRI and EEG is the fact that they can detect physiological responses that correspond to specific mental states [27], while traditional methods cannot detect a pattern prior to a moment of insight. One possible limitation associated with using EEG or fMRI is that the apparatus can be bulky or inhibiting. This may cause the participant to focus too much on the instrumentation and not enough on the task. To the best of our knowledge, this potential effect has neither been studied yet, nor has it been observed during studies conducted in our labs.

2.3 The Cognition of Creativity. Cognitive activities, such as visual perception and working memory, are important for creative processes. Visual perception is a dialectic process during which the designer “talks” with his or her sketch to generate new ideas from externalized memory [28,29]. Working memory is a core cognitive function that facilitates the transfer of memory from one cognitive process to another [30,31], and is traditionally measured using the backward digit span and forward digit span tests [32]. With regard to creativity, working memory is particularly important because it stores relevant details extracted from visual perception of design artifacts. The designer then compares these details with categorical information stored in long-term memory. This process chain is the source of categorical flexibility when generating concepts [3]. An extensive review of cognitive theories of creativity can be found in Ref. [3]. The links between working memory, visual perception, and creativity provide a foundation for some of the EEG measurements that will be described in Sec. 4.1. That section will also introduce the EEG metrics associated with the system we used and the constructs associated with them.

2.4 Psychophysiological Insights Into Ideation. Martindale reviewed research relevant to the psychophysiological responses of creativity [33] and found that many researchers reported an increase in alpha waves during a creative event. In fact, more creative individuals had stronger alpha waves while at rest. Martindale and others theorized that these results indicate that creativity is a “disinhibition” syndrome, closely related to characteristics of psychosis, or a disconnect from reality. That is to say that creativity is the result of making incongruous relationships between thoughts and then choosing to keep those connections. This adds weight to the widespread view that creativity is combinatorial [34,35].

While the studies reviewed by Martindale are useful for understanding creativity in general, his findings may not map directly to a problem-solving ideation task. Problem-solving tasks are not only judged by the diversity of ideas, as in a divergent-thinking
3 Experiment

The goal of this study is to identify relationships between EEG responses and design outcomes. We will particularly examine the individual concepts that are generated during the design task. The data used for the analysis presented in Sec. 6 were the result of an earlier experimental study with different treatment groups [46]. Because an analysis of variance (ANOVA) showed that the treatment groups had no effect on design outcomes, we combined all three groups into a larger dataset for the prediction model used in this paper.

Although the treatment group was not a significant factor in the analysis presented in this paper, we describe the entire process so that future researchers can see how the study was conducted.

3.1 Design Task. The choice of design task was governed by the fact that we wanted to recruit participants from multiple engineering disciplines. Consequently, we chose a problem that required minimal technical skills to solve and asked participants to generate concepts only. It is important to note that the choice of the design problem and how it is worded can affect the design outcome as measured by novelty, quality, fluency, and other parameters [48]. It is outside the scope of this work to measure how different design problems will affect the results; therefore, we kept this variable constant between all participants.

Participants were given the following design task: “Generate as many concepts as possible for a device that will aid a student athlete with a leg injury. The athlete needs to be able to do normal campus activities such as go to class, get food, or use the restroom.” Participants were then given the opportunity to get answers to any questions they had. If they asked for more details (e.g., the injury type), we instructed them to define details for themselves. Participants were also informed that they could use as many sheets of paper as they desired.

3.2 Experimental Procedure. Figure 1 shows the design of experiment for this study. Each experimental session began with the participant signing an informed consent form. All participants were connected to the B-Alert X10 system (Advanced Brain Monitoring, Inc., Carlsbad, CA) on their heads. We also video recorded the sessions, but informed subjects that the video frame would contain only the work area with their papers and materials. An example of the workstation is shown in Fig. 2. After setting up the equipment, participants completed a calibration exercise using the iMotions Attention Tool (iMotions, Inc., Boston, MA). Subsequently, participants took a short, 2-min rest. After this point, group 1 immediately began generating concepts. Groups 2 and 3 did their respective warm-up tasks, took a short survey asking about the tasks, and then generated concepts. Participants were told that they had 15 min to generate concepts and were given a timer with an alarm to keep track of their time. After the concept generation period, all participants took a postsurvey which will be described in more detail in Sec. 4.5.

Because the participants in the art activities group were required to use multiple media for the warm up, we left the supplies at the workstation (Fig. 2) and instructed all groups that they could use any media they wished. The types of media provided were pencils, pens, crayons, markers, and finger paints.

3.3 Participants. Participants were recruited using fliers, email lists, referrals, and by word of mouth. All participants were compensated at a rate of $10/h. The sample included 42 adults between 18 and 44 yr in age, from a single Midwestern university. Of the 42 adults, nine were females and 33 were males.
participants were in an engineering or technology program. Fifteen participants had previously experienced leg injuries that required the use of a medical device. According to the curricula across the engineering and technology programs at this university, it was assumed that all undergraduate participants had completed at least one design course. It was also assumed that all graduate participants had completed at least a senior design course.

4 Evaluation Metrics

In this section, we will discuss the evaluation approach for the design outcomes and mental states during ideation. We will also present the procedure for mapping cognitive activities to each concept.

4.1 EEG Metrics. There are several metrics associated with EEG measurements. The traditional outputs of an EEG are electrical potentials at different frequencies and at different powers. Power spectral densities of the output identify the intensity of a wave. They are further divided into frequency bands: delta, theta, alpha, beta, and others that vary depending on the definitions by researchers. The alpha band is most commonly associated with creativity [33] and studies measuring problem-solving ideation using EEG and self-report data also confirm that the alpha wave is relevant to measuring creativity [4,43].

However, it is important to note that there is some disagreement as to whether the alpha wave measures creativity or not [33]. Hocevar’s hypothesis may explain the dissent [50]. First, research on creativity has adopted a wide range of definitions of the term “creativity.” Second, most metrics for creativity fail to measure creativity itself but rather measure something else such as behavior correlated with creativity. Therefore, discrepancies in definitions and theories provide possible explanations for the conflicting results of different studies.

In addition to the traditional theta, alpha, and beta waves, the B-Alert X-10 device, used in this study, also provides additional metrics: forward and backward digit span (FBDS) and backward digit span (BDS) workload, high engagement, low engagement, and distraction. These are defined below (see also Figs. 3 and 4 for the relative relationships between metrics of the same family). All these metrics resulted from a variety of predictive models that used traditional EEG bands to predict qualitative constructs [51–55].

4.1.1 Workload Metric Family. These metrics measure how hard a participant is thinking. Generally, these are measures of working memory and indicate to what extent this memory is being used. These measures are derived from cognitive tests such as the forward and backward digit span tests. In these tests, a participant is shown a series of numbers and asked to recall them in either a forward or reverse sequence. These metrics were validated and explained in more detail in several studies [51–53]:

(1) **FBDS Workload** (working memory, in this paper): The extent of working memory being used at a given moment. It is a combination of several mental tasks such as forward and backward digit spans. These are associated with working memory, planning, and recall.

(2) **BDS Workload** (mental manipulation, in this paper): It is defined by processes involved during backward digit span tasks, which are mainly associated with the storage and reordering of the given information [56,57].
4.1.2 Engagement Metric Family. These metrics measure the levels of a participant’s attentiveness and focus. Generally, they are related to the processes of information gathering, sustained attention, and visual scanning. These metrics are validated and explained in more detail in prior work [53–55]:

(1) **High Engagement** (active attention, in this paper): This metric is related to the process of visual scanning and the sustained attention a participant is giving to their external environment. It is the kind of engagement a driver feels when making a turn or stopping at a light. New drivers are nearly always in this state.

(2) **Low Engagement** (passive attention, in this paper): This metric is roughly equivalent to the degree to which a participant is reflecting on an idea, and it may be associated with mental visualization. This kind of engagement is what a driver feels when cruising on a freeway for an extended period. The driver is aware of the road, but is only passively scanning for information.

(3) **Distraction**: The rate at which the participant is shifting between mental tasks. Martindale also describes distraction, or “defocused attention,” as being a key feature of creativity [33].

4.2 EEG Measurements. The nine-channel B-Alert X10 wireless EEG headset system was used to measure brain waves from nine balanced sites on the scalp, including Fz, F3, F4, Cz, C3, C4, POz, P3, and P4 at a sampling rate of 256 Hz. Foam sensors and highly conductive electrode cream were attached to these sites, providing electrically conductive interfaces between the headset and the scalp. A pair of reference electrodes was placed behind the ear just above the mastoid process on the temporal bone. The collected data were wirelessly transmitted to a laptop with a Windows 7 operating system through a bluetooth–USB dongle. The transmission range was up to 10 m, allowing participants to freely move their heads and bodies for comfort. The room temperature was maintained at 72–74°F to reduce possible effects of environmental temperature.

4.3 Procedure for Analyzing Individual Concepts. We began the analysis by calculating elaboration, novelty, and quality of each concept generated by the participants. Although fluency is a very common metric, we chose not to include it in the analyses. Fluency evaluates concepts at an aggregated level, which is beyond the scope of this paper as our objective is to understand the correlation between mental states and an individual concept. Two expert raters evaluated the concepts with 30% overlap between the raters [58], instead of the typical 10% [59,60].

The chosen raters were two researchers in design theory and methodology. Both have judged projects for design classes multiple times and have a background in mechanical engineering. Neither is a medical professional, but both are slightly more knowledgeable than average about medical care. The raters were trained using concepts that were removed from the dataset because of poor EEG signals or video data. When the raters disagreed, they compared responses to the training data, established a common understanding, and repeated the training with another subset of removed data. No formal definitions were provided to the raters. We used the intraclass correlation coefficient (ICC), which measures whether inter-rater agreement and reliability, and is therefore more conservative than other measures [61]. The ICC values were 0.809 for elaboration, 0.616 for novelty, and 0.721 for quality. The metrics are defined below:

**Elaboration**: We assessed novelty by using a decision tree. The approach we used to calculate this method is closely modeled after the decision trees used by Kershaw et al. [62], which is adapted from the creative engineering design assessment (CEDA) approach [63]. We favored this approach over feature-level measures of novelty because many concepts simply did not give enough information to be broken down into features. The decision tree we used is depicted in Fig. 5. The raters considered a concept to be radically different if it was categorically different from existing solutions.

**Novelty**: Shah argues that any method can be used [11] to measure quality, but recommends a weighted decision matrix. Because many concepts could not be broken down into features, we again measured quality at the concept-level by adapting the decision trees used by Kershaw et al. [62]. The decision tree we used is shown in Fig. 6. Elaboration: We measured elaboration subjectively. Elaboration first appears for the Torrance test [13] and is measured by counting the number of lines that participants add to the triangles used for the test. The alternative uses test counts the number of words [16]. Since our experiment allowed participants to sketch and write, it was unclear how to combine these two measures in a meaningful way. Consequently, we counted the number of “chunks” in each concept. A chunk consisted of a subjectively discrete idea embodied in the representation of the concept, an approach analogous to the concept of memory chunking. Where a concept comprised a complete, readily available system, this was defined to be a single chunk. For example, concept AK06 (Fig. 7, left) is a motorized wheelchair. This is a commonly known concept that can be readily purchased as a complete system, so it constitutes one chunk. Concept AK07 (Fig. 7, right) shows an actuated knee brace. The knee brace counts as one chunk and the actuator counts as another, giving this concept an elaboration score of 2.
4.4 Procedure for Mapping EEG Outputs to Each Concept. The ideation session for each participant was video-recorded, and the recording was used to determine when each concept was initiated on paper by the participant. The EEG output assigned to each concept was defined as the interval between the end of the concept being recorded and the end of the prior concept being recorded. Figure 8 is a visual representation of concept intervals. In this example, the participant jumped into recording concept 2 immediately after the end of recording concept 1, and pauses after the end of recording concept 2. We assume this pause was used to think about concept 3. Because this procedure was not qualitative, no inter-rater reliability was calculated.

4.5 Survey. The postsurvey consisted of task load index (TLX) questions [45], standard demographic questions, whether the participant has ever had a leg injury before, attitudes and self-efficacy toward sketching, and opinions about the experience in general. Only the demographic and injury experience information were used in this study (Table 1). Table 2 summarizes the information collected from all participants that were analyzed in this study (30 participants in total). The data processing principle will be discussed in Sec. 5.1.

5 EEG Data Analysis

We first verified the quality of EEG data for each concept, removed bad data (details described in Sec. 5.1), and then investigated whether the EEG data could be used to predict the design outcomes. We selected different sets of predictors for different types of outcome metrics (i.e., novelty, quality, and elaboration) based on SVM classification. Predictors are subsets of features, which include the EEG metrics and demographic information collected from the postsurvey.

5.1 Process for Excluding Data. Before mapping the EEG data to the video time stamps, we removed invalid data that resulted from an imperfect data collection process. The noise/outlier removal process is frequently used to enhance data analysis [64,65]. This process left 123 concepts from 30 participants to analyze. There were several criteria for removing data. If the interval could not be clearly defined, the concept was removed from the data set (see Fig. 8 for definition of concept interval). Some participants worked on a concept, moved to another, then returned to the original idea. Because our method for defining when a participant generated an idea could not clearly account for this behavior, these data were removed. Concepts were also removed if the EEG data was of poor quality or missing during the relevant interval. Finally, concepts were removed if the video data was poor during the relevant interval.

Table 1 Selected survey questions distributed after the idea-generation task

<table>
<thead>
<tr>
<th>Examples of survey questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>What year in school are you?</td>
</tr>
<tr>
<td>What is your gender?</td>
</tr>
<tr>
<td>What is your major?</td>
</tr>
<tr>
<td>Have you ever had an injury that required you to use a device to move around?</td>
</tr>
</tbody>
</table>

Table 2 Information about the participants that were analyzed in this study

<table>
<thead>
<tr>
<th>Subgroup</th>
<th># Participants</th>
<th>Subgroup</th>
<th># Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>1</td>
<td>Male</td>
<td>24</td>
</tr>
<tr>
<td>Sophomore</td>
<td>1</td>
<td>Female</td>
<td>6</td>
</tr>
<tr>
<td>Junior</td>
<td>1</td>
<td>Mech. Eng. (ME)</td>
<td>12</td>
</tr>
<tr>
<td>Senior</td>
<td>6</td>
<td>Non-ME</td>
<td>18</td>
</tr>
<tr>
<td>Graduate 1st and 2nd year</td>
<td>15</td>
<td>Had an injury</td>
<td>10</td>
</tr>
<tr>
<td>Graduate 3rd and 4th year</td>
<td>5</td>
<td>Have not had an injury</td>
<td>20</td>
</tr>
<tr>
<td>Graduate 5+ year</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.2 Data Processing. We further inspected the included EEG data for bad data points. Bad points may have been a result of artifacts or unstable Bluetooth connection and were marked as −99,999 in the Advanced Brain Monitoring (ABM) exported file. All EEG metrics were recorded at the rate of one data point per second. We replaced the bad data with the value of the previous second by assuming there were no abrupt changes in participants’ mental states between two adjacent moments while they were generating a concept. A total of 321 data points (1.7%) were identified as bad data and replaced. We then calculated the mean value of each EEG metric for each concept. Ultimately, we established a matrix of ten features (i.e., potential predictors), which included five EEG metrics, four demographic/experience factors, and the time spent (noted as TimeSpan afterward) for each concept (see Table 4 for list).

5.3 Support Vector Machine (SVM). Support vector machine (SVM) is a machine learning method that is associated with analyzing data used for classification and regression analysis. There are some advantages to using SVM compared to common statistical methods. The relation between the features and the design outcomes was too complex and thus was not linearly separable. For instance, there was no simple linear regression model that could make the whole set of high-quality concepts linearly separable from the whole set of low-quality concepts. SVM classifiers are able to handle nonlinear classification problems. They map the input space into a high-dimensional feature space through a kernel function (e.g., polynomial or Gaussian), and then find the best hyperplane that separates all data points of one class from those of the other class. In addition, SVM classifiers not only have a regularization parameter that helps avoiding over-fitting, but they are also robust. An increase or decrease in the number of samples that are not support vectors (i.e., not on the margins) does not influence the SVM model. These advantages led to superior performance of SVM in several EEG studies, such as synchronous brain–computer interface, emotion recognition, and eye events, compared to other machine learning methods [66–68]. Therefore, SVM classifiers are extensively used in EEG-based research to facilitate predictions based on brain wave features, including ABM metrics [69–72].

The first step of training a SVM classifier is to label each concept together with the corresponding row of feature data with the design outcomes (classes). SVMs are binary classification methods, so we labeled the concepts into two groups for each outcome metric according to the criteria listed below. As a result, we had 61 high- and 62 low-quality concepts, 57 high- and 66 low-novelty concepts, and 57 high- and 66 low-elaboration concepts. Table 3 shows the number of concepts in each demographic group by the ideation metrics. The percentages indicate what proportion of the total number of concepts they represent. For example, undergraduates had 24 high-quality concepts out of 44, which equates to 54.5% of the high quality concepts. The differences in percentages are not necessarily statistically significant.

<table>
<thead>
<tr>
<th>Features</th>
<th>Novelty</th>
<th>Quality</th>
<th>Elaboration</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distraction</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>ABM-Distraction</td>
</tr>
<tr>
<td>Passive attention</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ABM-Low engagement</td>
</tr>
<tr>
<td>Active attention</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ABM-High engagement</td>
</tr>
<tr>
<td>Working memory</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ABM-FBDS Workload</td>
</tr>
<tr>
<td>Mental manipulation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ABM-BDS Workload</td>
</tr>
<tr>
<td>Year in school</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>Undergraduate or graduate</td>
</tr>
<tr>
<td>Gender</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Male or female</td>
</tr>
<tr>
<td>Major</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>ME or not</td>
</tr>
<tr>
<td>Injury experience</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>Had or not had</td>
</tr>
<tr>
<td>TimeSpan</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>Seconds spent on the task</td>
</tr>
</tbody>
</table>

Table 4 The ten features used to predict design novelty, quality, and elaboration. Included features are marked with a checkmark, and features to be excluded are marked with a dash.

We passed the data to the Statistics and Machine Learning Toolbox™ in MATLAB R2016a (The MathWorks, Inc., Natick, MA) to train SVM classifiers. There were ten features (see Table 4) from which we selected predictors that were best able to...
discriminate between the two classes (e.g., high- and low-novelty). We added one feature at a time to trained classifiers and compared the accuracy between these classifiers via five-fold validation [73] to select the set of predictors.

When training an SVM, it is important to use balanced data, where a similar number of samples describe each condition. This ensures that both the positive and negative cases have approximately the same accuracy value. With an unbalanced dataset, such as 90% of the samples in one group and 10% in another group, the accuracy value would be higher, but the result would be less reliable and misleading. In such cases, other machine learning methods would be required. Another potential problem is large differences in the accuracy results for the two conditions. For example, if high-novelty could be predicted with 80% accuracy, but low-novelty was predicted with only 60% accuracy, the result would not be as robust as a different classification that was able to predict results with 71% and 72% accuracy for the two conditions. Consequently, it is best practice to discuss the sample size effect and report the confusion matrix.

6 Results and Discussion
The analysis using the SVM prediction comprised three main parts. First, we randomly partitioned the samples (i.e., 123 concepts) into five equally sized subsamples. Four of the subsamples were used as training data, and the remaining subsample was used as the validation data to verify the accuracy of the classifier. This process was repeated five times (five-fold) across each subsample, and the average error was calculated. The five-fold validation was then repeated for each predicted measure: novelty, quality, and elaboration. The validation was completed separately for each measure because each one required a different set of predictors. Second, a pseudo-dataset was created. This pseudo-dataset comprised every possible combination and had the same range as the original dataset. The classifier was then applied to the pseudo-dataset to predict behavior when no data were originally available.

6.1 Classifiers and SVM Prediction Model Validation. Table 4 lists the predictors that have a higher probability of accurately classifying a design outcome based on its novelty, quality, and elaboration. For example, significant features for predicting novelty are a designer’s distraction, passive attention, active attention, working memory, mental manipulation, year in school, and their major. Some interesting findings include: distraction was a predictor for novelty only, and gender was not a significant predictor in any of the outcome metrics. Figure 9 shows the prediction accuracy (true positive rates) for each metric. The SVM classifiers were able to accurately predict if a concept was novel (74%) or was of high quality (70.7%), based on EEG data and demographic metrics. The prediction of elaboration was also available (64.2%), but was less accurate than the prediction of novelty and quality.

6.2 Relationships Between EEG Measurements and Design Outcomes. We noticed that the mental states corresponding to “good” design outcomes may be different between demographic subgroups (e.g., ME students and non-ME students). Even if a demographic predictor (e.g., major) is not significant, that does not necessarily mean it has no effect on the design outcomes. However, it implies that the relationship between EEG data and design outcomes is similar between the two demographic subgroups. For example, consider the EEG metric active attention and the design outcome of quality. If the active attention during ideation of graduate students is higher as compared to undergraduate students, and graduate students generate more high-quality concepts, the information we can extract from “year in school” and “active attention” are overlapped. Therefore, including both predictors will not enhance the classification accuracy. The primary effect of these demographic predictors is outside the scope of this work which is to discover the relationship between EEG data and design outcomes, and is subject to future studies. Below, we will detail the predicted relationship between EEG data and each design outcome metric by demographic subgroups.

We quantified low, middle, and high active attention as “0.0–0.3, 0.3–0.6, and 0.6–0.9 high engagement,” respectively. The same criterion was also applied to passive attention (low engagement) and distraction. Moreover, we quantified middle, mid-high, and high working memory as “0.5–0.67, 0.67–0.83, and 0.83–1 FBDS Workload.” The same quantification was used for mental manipulation (BDS Workload). To facilitate a discussion of the effects of the complete workload metric family, we further defined “overall workload” as the integrated effect of working memory and mental manipulation. For example, both high working memory combined with middle mental manipulation and middle working memory combined with high manipulation indicate mid-high overall workload. Figures 10–17 show the SVM prediction results. In these figures, the x-, y-, and z-axis represent the level of low engagement, high engagement, and BDS Workload, respectively. Other features that cannot be accommodated in the axes will be represented categorically in subplots (i.e., in rows and columns). More details about the figures can be found in the Appendix.

6.2.1 Relationship Between EEG Measurements and Novelty of Concepts. The pseudo-data set for novelty consisted of 37,400 data points. Figures 10–13 show the complete set of findings derived from the SVM prediction results of design novelty based on distraction, passive attention, active attention, working memory, and mental manipulation for ME undergraduate students, non-ME undergraduate students, ME graduate students, and non-ME graduate students, respectively. The mental states predicted to be most supportive for the generation of novel concepts varied
Fig. 10 Predicted concept novelty for ME undergraduate students. The $x$, $y$, and $z$-axes represent low engagement (passive attention), high engagement (active attention), and BDS Workload (mental manipulation), respectively. The left to right columns represent low (0–0.3), middle (0.3–0.6), and high (0.6–0.9) distraction. The top to bottom rows represent mid (0.5–0.67), mid-high (0.67–0.83), and high (0.83–1) FBDS Workload (working memory).

Fig. 11 Predicted concept novelty for non-ME undergraduate students. The $x$, $y$, and $z$-axes represent low engagement (passive attention), high engagement (active attention), and BDS Workload (mental manipulation), respectively. The left to right columns represent low (0–0.3), middle (0.3–0.6), and high (0.6–0.9) distraction. The top to bottom rows represent mid (0.5–0.67), mid-high (0.67–0.83), and high (0.83–1) FBDS Workload (working memory).
Fig. 12 Predicted concept novelty for ME graduate students. The x-, y-, and z-axes represent low engagement (passive attention), high engagement (active attention), and BDS Workload (mental manipulation), respectively. The left to right columns represent low (0–0.3), middle (0.3–0.6), and high (0.6–0.9) distraction. The top to bottom rows represent mid (0.5–0.67), mid-high (0.67–0.83), and high (0.83–1) FBDS Workload (working memory).

Fig. 13 Predicted concept novelty for non-ME graduate students. The x-, y-, and z-axes represent low engagement (passive attention), high engagement (active attention), and BDS Workload (mental manipulation), respectively. The left to right columns represent low (0–0.3), middle (0.3–0.6), and high (0.6–0.9) distraction. The top to bottom rows represent mid (0.5–0.67), mid-high (0.67–0.83), and high (0.83–1) FBDS Workload (working memory).
Fig. 14 Predicted concept quality for ME students. The $x$, $y$, and $z$-axes represent low engagement (passive attention), high engagement (active attention), and BDS Workload (mental manipulation), respectively. The left to right columns represent short (0–300 s), middle (300–600 s), and long (600–900 s) TimeSpan. The top to bottom rows represent mid (0.5–0.67), mid-high (0.67–0.83), and high (0.83–1) FBDS Workload (working memory).

Fig. 15 Predicted concept quality for non-ME students. The $x$, $y$, and $z$-axes represent low engagement (passive attention), high engagement (active attention), and BDS Workload (mental manipulation), respectively. The left to right columns represent short (0–300 s), middle (300–600 s), and long (600–900 s) TimeSpan. The top to bottom rows represent mid (0.5–0.67), mid-high (0.67–0.83), and high (0.83–1) FBDS Workload (working memory).
across demographic groups. The figures show that in mid to mid-high working memory mental states, undergraduate students were able to generate high-novelty concepts in a wider range of mental states than graduate students (see rows 1 and 2 of Figs. 10 and 12 and Figs. 11 and 13). This may imply that the undergraduate students’ mental states were more free or flexible during ideation. On the other hand, graduate students generated more novel concepts when they relied on high working memory mental states than undergraduates (see row 3 of Figs. 10 and 12 and Figs. 11 and 13). In general, we concluded that the mental states correlated with high-novelty were: (1) undergraduate students in states of high distraction (i.e., often shifting between mental tasks) and/or mid-high overall workload (see Figs. 10 and 11); (2) graduate students in states of high distraction, and/or relying on working memory rather than on mental manipulation (see rows 1, 3 and column 3 of Figs. 12 and 13). Moreover, graduate students with an increased level of active attention generated less novel concepts, while active attention did not have a similar effect on undergraduate students (see row 1 of Figs. 10–13).

The observed differences between undergraduate and graduate students agreed with existing literature reporting that age, knowledge/training, and quality of thinking are significant factors in creative thinking [74–76]. Because the majority of the participants were senior undergraduate students (six out of nine) and graduate students in their first or second years (15 out of 21), the participants were very likely to be in the same age group. This implies that knowledge, work experience, and other training activities may have contributed to the differences in undergraduate and graduate students’ creativity more than age did. Future work could explore the effects of these factors on cognitive activities during ideation.

On closer inspection of the effects of each feature, we determined that for ME undergraduate students, highly distracted states of mind with middle mental manipulation were helpful in generating novel concepts, regardless of the passive and active attention levels (see column 3 of Fig. 10). It can also be seen in column 3 that overall working memory did not significantly affect the novelty either. In the case of low to middle distraction levels, ME undergraduate students needed an approximately middle/mid-high overall workload (see columns 1 and 2 of Fig. 10, especially rows 1 and 3). When middle distraction and high working memory were present, low passive attention levels were helpful (see subplot (3,2) in Fig. 10). These relationships between EEG data and concept novelty were also observed in predictions for non-ME undergraduate students, but the range of mental states correlated with generating novel concepts was narrower than for ME undergraduate students (see Fig. 11).

For ME graduate students, highly distracted states of mind were generally helpful in generating novel concepts, preferably with high working memory and middle mental manipulation (see column 3 of Fig. 12). For low to middle distraction levels, high working memory or mental manipulation levels were helpful (see columns 1 and 2 of Fig. 12). In addition, given middle or mid-high working memory states, ME graduate students needed decreased active attention levels (see rows 1 and 2 of Fig. 12). This implies that decreased or disengaged sensory scanning may have benefited them. Similar to results for undergraduate students, the observed relation between EEG data and concept novelty for ME and non-ME graduate students were comparable. However, the range of mental states correlated with generating novel concepts was narrower for non-ME graduate students (see Fig. 13).

6.2.2 Relationship Between EEG Measurements and Quality of Concepts. Similar to the approach in Sec. 6.2.1, we established a test data set consisting of 187,200 data points, and the results are shown in Figs. 14 and 15. These figures show SVM predictions of design quality based on low engagement (passive attention), high engagement (active attention), FBDS Workload (working memory), BDS Workload (mental manipulation), and TimeSpan for ME and non-ME students, respectively. Both the time spent on generating a concept and the engagement level had a significant influence on the results. A student was more likely to generate a high-quality concept if he/she spent a relatively long time generating it, and was in the state of high workload (see column 3 of Figs. 14 and 15). An increased level of either active or passive attention was also helpful (see column 2, 3 and subplot (3,1) of Figs. 14 and 15). This implies that distraction had a negative effect on concept quality. While distraction was correlated with high novelty, this result suggests that the conditions facilitating concept generation processes resulting in high novelty and high quality may be incompatible. This supports the general observation that novel concepts are often immature and less practical. In addition, high-quality concepts are positively correlated with a longer ideation time. This implies that designers generate fewer new ideas during ideation (i.e., low overall quantity), and instead, focus more on evaluating and reassembling ideas and/or refining requirements.

These results can be partly explained by existing studies. Sobek and Jain reported that the time spent on idea generation on a system-level is positively correlated with design quality; however, more time spent on concept-level design was associated negatively with design quality [77]. If we apply the ideation loop proposed by Chusilp and Jin [78], “system-level design” involves three stages of iteration: problem redefinition (PR), idea stimulation (IS), and concept reuse (CR), while “concept-level design” specifically involves idea simulation. The number of iterations in each loop has a positive correlation with design quality, but only the IS loop facilitates idea novelty. The PR loop, on the contrary, has a negative impact on idea novelty [79]. The IS loop associates memory retrieval and creative property perception by visual or other input. Based on EEG metrics and the design outcome prediction model proposed in this paper, the cognition in the IS loop iteration is similar to distraction (shifting between mental tasks), which was positively correlated with concept novelty, but may be negatively correlated with concept quality. The latter finding contradicts the ideation loop study, but agrees with Sobek and Jain’s results on concept-level idea generation and quality. One major difference between our and these studies is the duration of the task (15 min versus approximately 1 h and a semester). Future studies should investigate the relationship between cognitive activity, task duration, iterations, and concept- or system-level approach.

The detailed findings of our study also indicate that high active attention levels combined with mid-high overall workload can help ME students generate high-quality concepts (see column 2, subplots (1,3) and (3,1) of Fig. 14), while states of middle mental manipulation or increased passive attention are more helpful for non-ME students (see rows 2, 3 and column 3 of Fig. 15). Focusing on the ideation task also benefits concept quality. When participants spent 5–10 min to generate one concept, increased active attention was correlated with high design quality (see column 2 of Figs. 14 and 15). High working memory demand promoted concept quality when 5–10 min were spent on concept generation. The positive effects of higher workload (both working memory and mental manipulation) become more recognizable with a longer TimeSpan (see subplots (2,2), (2,3), (3,2), and (3,3) of Figs. 14 and 15).

6.2.3 Relationship Between EEG Measurements and Elaboration of Concepts. The test data set for understanding the relationship between EEG measurements and elaboration consisted of 62,400 data points. Figures 16 and 17 show SVM predictions of the design elaboration based on low engagement (passive attention), high engagement (active attention), BDS Workload (mental manipulation), TimeSpan, and the experience of prior leg injury for graduate and undergraduate students, respectively. The overall results show that participants with personal experience regarding the design problem were able to elaborate more in less time (see column 1 in Figs. 16 and 17). Nonetheless, when a participant without injury experience spent a relatively long time and was
actively attentive, he/she was more likely to embed more chunks in a concept (see column 3 in Figs. 16 and 17). We assume that participants with personal and contextual experience with the task may have been more fluent in requirement elicitation, and thus capable of providing more details in their final concepts. It is intriguing to further explore whether this kind of contextual experience can help designers address more requirements and propose better systems.

The observed correlation between EEG measurements and Time Span with respect to concept elaboration was consistent across various combinations of mental states. The difference between participants who were familiar with the design problem and those who were not was much larger for concepts that were created in a short time frame (see column 1 in Figs. 16 and 17). This difference is less obvious, but still observable for a medium Time Span (see column 2 in Figs. 16 and 17), and not discernible for a long Time Span (see column 3 in Figs. 16 and 17). In addition, graduate students were more likely to express high elaboration than undergraduate students (see subplots (1,1) and (1,2) in Figs. 16 and 17). Finally, moderate mental manipulation benefited elaboration irrespective of the status of engagement in most cases (observable in each subplot except (1,1) and (1,2) in Fig. 17).

6.3 General Discussion. The research questions that guided this work were: (1) What is the relationship between EEG metrics and traditional ideation metrics used in assessing design outcomes? (2) How do designers’ mental states (as measured by EEG metrics) influence design outcomes? We found that the EEG metrics (distraction, passive attention, active attention, working memory, mental manipulation) are associated with the traditional ideation metrics novelty, quality, and...
elaboration in different ways. For example, distraction is associated with novelty and working memory is associated with novelty and quality (see Table 4). Our findings suggest the following about designers’ mental states and their influence on design outcomes: (1) passive attention, active attention, and mental manipulation are significant predictors for all three ideation metrics; (2) distraction is positively correlated with design novelty and may have indirect negative effects on design quality; (3) highly active attention is correlated with good design quality. Table 5 summarizes EEG factors that have the strongest effect on the design outcome based on the SVM-predicted results. In addition, we discovered that the set of EEG predictors may differ for different demographic groups. For example, designers with a mechanical engineering background can generate novel concepts in a wider range of mental states.

These results suggest that EEG metrics are indicative of good ideation and are capable of predicting the novelty, quality, and elaboration of concepts while considering designer demographics.

### Table 5 Qualitative description of the mental states that are strongly correlated with design ideation metrics. Each column represents one set of attributes that leads to the result shown in the header. The time spent on concept generation is also included (L, low; M, medium; MH, medium-high; and H, high).

<table>
<thead>
<tr>
<th>Features</th>
<th>High-novelty</th>
<th>High-quality</th>
<th>High-elaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distraction</td>
<td>H</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Passive attention</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Active attention</td>
<td>—</td>
<td>—</td>
<td>H</td>
</tr>
<tr>
<td>Working memory</td>
<td>M–H</td>
<td>MH–H</td>
<td>MH–H</td>
</tr>
<tr>
<td>Mental manipulation</td>
<td>M</td>
<td>M–MH</td>
<td>MH–H</td>
</tr>
<tr>
<td>TimeSpan</td>
<td>—</td>
<td>—</td>
<td>M–H</td>
</tr>
</tbody>
</table>

7 Conclusion

The results of this study suggest that EEG and designers’ demographics can be used to partially replace the traditional ideation metrics, and it has an accuracy of 70%. This means that EEG data can be paired with factors like year in school or major to obtain meaningful information, without the cost of rating individual concepts. In addition, the results of this study could be used to analyze optimal mental states of designers during ideation tasks in real time. Some possible applications of our results include the ability to cue designers when they have reached a good break point or for designers to analyze whether they are in the right state of mind or whether they need to do warm ups. Studies of this kind may provide fundamental knowledge as our society moves toward the development of cyber-physical-social systems [80] and cyber-learning environments [81]. Having access to methods to capture real-time cognitive responses during design activities can open up new avenues for research in these areas.

Fig. 18 An example prediction result of concept novelty with three 2D projection views
The results of this study revealed the importance of considering demographic factors and their role in ideation studies. The results showed that demographics such as participants’ year in school and major are significant predictors in SVM prediction models. They imply that designers’ knowledge or training affect the ideation process, which aligns with previous findings in the literature [74–76].

Many EEG features and demographic data were needed to predict quality and novelty, indicating that the EEG metrics we used do not have a one to one mapping relationship to novelty or quality. This supports the research objective to triangulate EEG measurements with established ideation metrics. Future work may include comparing EEG responses to other approaches to ideation research.

One limitation of our work is that we used a single design problem for all participants. Exploratory work has shown that several characteristics, including the scale and scope of the design problem for all participants. Exploratory work has shown that several characteristics, including the scale and scope of the design problem, can affect the design outcomes as measured by novelty, quality, and fluency [48]. An important element of future work will be to investigate the effect of changing the design problem on EEG responses.

In addition, significant research on creativity involves examining the role of design fixation [82,83], where designers rely on familiar or present examples to generate ideas. In our work, example solutions were not presented; however, it is possible that participants relied on mental examples from past experiences. Future work may involve assessing the degree to which participants are relying on examples from past experience to guide their design process.

Appendix: Illustration of Result Figures

Figure 18 is an example prediction result of concept novelty, which can be found in subplot (1,1) in Fig. 10 which presents results from ME undergraduate students. Five features: passive attention (low engagement, z-active), active attention (high engagement, y-active), mental manipulation (BDS Workload, z-active), working memory (FBDS Workload, catego-

References

[22] Srivatsavai, R., Genco, N., Holttia-Otto, K., and Seepersad, C. C., 2010, “Study of Many EEG Features and Demographic Data Were Needed to Predict Quality and Novelty, Indicating that the EEG Metrics We Used Do Not Have a One to One Mapping Relationship to Novelty or Quality. This Supports the Research Objective to Triangulate EEG Measurements with Established Ideation Metrics.”

Journal of Mechanical Design

MAY 2017, Vol. 139 / 051011-15

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/pdfaccess.ashx?url=/data/journals/jmdedb/936131/ on 04/03/2017 Terms of Use: http://www.asme.org