Proceedings of the ASME 2010 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2010
August 15-18, 2010, Montreal, Quebec, Canada

DETC2010-28675

POTENTIAL LIMITATIONS OF VERBAL PROTOCOLS IN DESIGN EXPERIMENTS

I. Chiu
Department of Mechanical & Industrial Engineering
Ryerson University
350 Victoria Street
Toronto, ON, M5B 2K3, Canada
ivey.chiu@ryerson.ca

L.H. Shu*
Department of Mechanical & Industrial Engineering
University of Toronto
5 King's College Road
Toronto, ON, M5S 3G8, Canada
*Corresponding author email:
shu@mie.utoronto.ca

ABSTRACT

Contradictory results of a recent design stimulation and creativity experiment prompted us to re-examine our chosen methodological approach, namely the use of verbal protocols. We used verbal protocols to study design cognition associated with stimulus use. Our results showed that use of stimuli did not increase concept creativity, contradicting much of the design literature. After eliminating other possible errors, we re-examined the experimental methodology to identify potential design-specific limitations associated with verbal protocols.

Many researchers have used verbal protocol experiments, also known as talk-out-loud experiments, to study cognitive processes, as there are few other methods to study internal cognition. While verbal protocols are a widely debated method, research has been done to validate them, and precautions can be taken to mitigate associated risks. Based on reviewing the literature and our own experiences, we have developed design-specific guidelines for the use of verbal protocols. We also outline future work required to explore and understand the suitability of verbal protocols for design studies. Despite potential limitations, verbal protocols remain a valuable and practical tool for studying design cognition and therefore should not be discarded.

Keywords: Design cognition, design stimulation and creativity, verbal protocols, experimental methodology.

1 INTRODUCTION

Since the 1960's, there has been an interest in systematically studying and understanding the design process and underlying design cognition (Simon, 1969; Rittel & Weber, 1984). As design is integral to engineering and greatly influences subsequent processes in the product realization cycle, e.g., manufacture and product use, there is a clear need to understand design so that it can be supported. One of the difficulties of studying design is that it involves studying the human designer. While it is possible to observe and analyze the inputs and outputs of design, it is difficult to observe internal cognitive processes. Many methods have been applied to understanding design cognition including observational studies (Brereton & McGarry, 2000); interviews (Segers, 2004); analysis of design conversations (Dong, 2006); pen-and-paper studies (Yang & Cham, 2007); and physiological studies including use of body movements (Tang & Zeng, 2009) and functional magnetic resonance imaging (fMRI) (Alexiou et al., 2009). While physiological studies are promising, results are preliminary at this point (Zeng, 2009; Alexiou et al., 2009).

Although debated, another common method for studying internal cognitive processes is through the use of verbal protocols, also known as think-out-loud or talk-out-loud studies. In this method, participants are instructed to verbalize all thoughts as they simultaneously complete a task. Verbal protocols have been used to study various processes including human-machine interactions (Bainbridge et al., 1968; Obtata et al., 1993) and decision-making in medicine (Lutfey et al., 2008). Other methods also make use of speech or verbalizations, e.g., interviews, focus groups, etc., and protocol analysis can be conducted in a group setting such as in the Delft Protocols Workshop (Cross et al., 1996). In this paper, we focus on verbal protocols used to elicit a single participant's immediate thought processes while completing a design task.

2 MOTIVATION

We are motivated to re-examine the application of verbal protocols to design experiments because of surprising results obtained in a recent design stimulation and creativity

experiment. Originally, we were investigating the effects of opposite- and similar-stimulus words on concept creativity using verbal protocol studies. Comparing concepts generated with stimuli to those generated without, we observed no differences in creativity. This surprising finding prompted us to re-examine our results and methodology to explore the suitability of verbal protocols in design studies. In this paper, the comparison and discussion of our original results, along with results found from the literature, will form the basis of our investigation into the potential limitations of verbal protocol experiments in design.

Others have shown that random stimuli appear to increase concept creativity. This is likely because the designer is forced to reconcile differences between non-obvious stimuli and the problem-at-hand to arrive at a new perspective (de Bono, 1970; Thomas & Carroll, 1984). To us, it appeared that opposite-stimulus words would have the same advantage of "random" stimuli, in being non-obvious. At the same time, opposite-stimulus words can be systematically generated from antonymy relationships such as those found in a thesaurus. Because both opposite and similar words are represented in the antonymy/synonymy relationship, we decided to use both types of words as stimuli and hypothesized that opposite-stimulus words would increase concept creativity. Additionally, we sought to determine how designers were using the different types of stimuli and if differently related stimulus words would elicit different designer behaviors.

In our studies, which comprised a total of four experiments, we used a combination of pen-and-paper and verbal protocol experiments (Chiu & Shu, 2008a, 2008b). In pen-and-paper experiments, participants indicate their design concepts on worksheets. While this is a fairly concise and efficient method of collecting data from a large sample size, concepts are usually brief, with no explanation or insight to participants' design cognition. As we were also interested in *how* participants used stimuli, we incorporated verbal protocol experiments into our investigations. In verbal protocol experiments, participants verbalize all thoughts as they designed in addition to indicating concepts on worksheets. Using a combination of these two methods, we found that independent raters scored opposite-stimulus concepts as more creative than similar-stimulus concepts. Using verbal protocol experiments, we also gained insight into the effects of language stimuli on designer cognition and behavior. For example, we showed that stimulus words used as verbs introduced significantly more new concept elements into the concept generation process (Chiu, 2010).

However, surprisingly, when we introduced a no-stimulus condition, or a control condition, in our verbal protocol experiments, we found no advantages of using stimuli, i.e., either similar or opposite stimuli, in terms of creativity. No-stimulus concepts were found to be equally as creative as opposite-stimulus concepts. In fact, when we compared no-stimulus concepts with similar-stimulus concepts, we found that no-stimulus concepts were significantly more creative than similar-stimulus concepts. The experimental results are summarized in Table 1.

Table 1: Summary of experimental results.

Exp. Condition	Result
No stimulus (control)	Concepts scored equally creative as opposite-stimulus concepts, p-values > 0.05
Similar stimulus	Concepts scored significantly less creative than opposite-stimulus and no-stimulus concepts, p-values ~ 0.05.
Opposite stimulus	Concepts scored equally creative as no-stimulus concepts, p-values > 0.05

The results described above and in Table 1 contradict our intuition and much of the design literature advocating use of stimuli to increase design concept creativity, e.g., synectics, (Gordon, 1961), random stimuli (de Bono, 1970; Thomas & Carroll, 1984), and TRIZ (Altshuller & Shulyak, 1996). Design stimulation experiments such those conducted by Thomas and Carroll (1984), and Tseng et al., (2008) have found that the use of stimuli increased concept creativity measures. Specifically, Thomas & Carroll (1984) found that participants provided with semi-random stimulus words generated more creative concepts than those not provided with stimulus words. Tseng et al., (2008), found that stimulus participants generated a larger number of concepts and more novel concepts than no-stimulus participants. It should be noted that most design stimulation studies, and the above specifically, are pen-and-paper experiments.

The differences between our results and those of others lead us to theorize that our unexpected results are due to experimental methodology; namely the use of verbal protocols rather than pen-and-paper. We theorize that the use of verbal protocols may have negatively affected participant performance. It is known that task overload is detrimental to performance, e.g., multitasking while driving, and our experiment may have demonstrated this within a design context. Specifically, using stimulus *and* designing *and* verbalizing concurrently may have increased the participants' cognitive workload to the point of deteriorated performance.

As verbal protocols appear a practical method for studying design cognition, we feel that it is valuable to further understand this method in a design study context, and to provide guidelines for its use, rather than to recommend discarding verbal protocols. In the rest of this paper, we will first review the literature associated with verbal protocols and also describe some other design experiments using verbal protocols. Next, we will summarize our experimental results and provide guidelines for verbalization experiments. Finally, we will outline future work required to further understand and quantify the limitations of verbal protocols in design experiments.

3 BACKGROUND

In this section, we present background information related to the development and use of verbal protocol experiments, including the risks associated with verbal protocols.

3.1 Development of verbal protocols

Multiple philosophical approaches have been developed to study human cognition, with many of them using verbal protocols as a method to study internal cognitive processes. These approaches include Gestalt Psychology, Behaviorism and Cognitivism. None of these approaches are mutually exclusive, but rather were developed in an attempt to explain gaps in other approaches.

Gestalt Psychology takes a holistic approach to thought and originated as a study of perception. Gestalt Psychology acknowledges that thought is holistic, parallel, analog, and that a correlation exists between cognitive processes and conscious experiences. While the descriptions of cognition offered by Gestalt Psychologists are valuable, the Gestalt approach was often criticized for being merely descriptive (Köhler, 1959).

Behaviorism, in contrast, often dealt with behavioral conditioning in addition to descriptions, and does not recognize "introspection". Behaviorism is based on studying cognition only through observable phenomena, e.g., stimulus and response. A goal of Behaviorism was to formalize psychology as an objective branch of science similar to physics or chemistry (Deubel, 2003).

Cognitivism asserts that internal cognitive processes occur, and these internal processes mediate the response to stimulus. Unlike Gestalt Psychology, Cognitivism takes a reductionist approach, reducing human cognition to the smallest steps possible, similar to how a computer operates. Cognitivism is the predecessor to the Human Information Processing model of cognition (Deubel, 2003), which is commonly accepted in psychology today (Wickens et al., 1997).

Advocates of all three approaches have used verbal protocols to study human cognition (Ericsson, 2002). For example, since Behaviorists hold that stimulus directly affects response, then verbalizations translated directly into thoughts, and vice versa. In Cognitivism, it is held that cognitive processes can be inferred from behaviors such as verbalizations, but not that there is a direct mapping. This is because there are intermediate cognitive processes, e.g., memory, attention, etc., that mediate stimulus and response (Ericsson & Simon, 1993).

3.2 Risks associated with verbal protocols

While verbal protocols and verbalizations in design studies are generally accepted, and regarded as relatively objective (Hubka & Eder, 1996; Cross et al., 1996; Gero & McNeill, 1998; Atman et al., 2004; Visser, 2006), it is not without controversy. Some of the widely debated risks include time and resource intensiveness; data validity; and study of tasks not conducive to verbalization. These risks are discussed below.

Time and resource intensiveness: Verbalization experiments are regarded as requiring more time and resources than other methods, e.g., pen-and-paper experiments. Experiment sessions must be conducted individually; sessions must be recorded and subsequently transcribed; and transcripts must be coded, or marked up, before the main analysis. To prevent bias, independent transcriptionists and coders are hired, adding to the expense and lead-time involved with conducting verbal protocol experiments.

A direct consequence of increased time requirements is that verbalization experiments often have a smaller sample size. A survey of design and problem solving studies using verbal protocol experiments shows typical sample sizes range from one to 20 participants, although larger samples sizes are also possible, e.g., 93 (Atman et al., 2004) and 244 (Lutfey et al., 2008). It should be noted that a recent brain imaging design study involved 18 participants (Alexiou et al., 2009).

Data validity: Another debate is whether verbal reports accurately reflect the events being reported. Nisbett & Wilson (1977) compared verbal reports and actual recordings of the events reported upon and found they do not necessarily match. There are also concerns that talking about the task will change the task. However, these concerns can be avoided if verbal reports occur "on-line", that is, immediately. According to Ericsson & Simon, (1993) immediate verbalization accurately describes the task and does not alter the task being studied because such verbalizations are strictly drawn from short-term memory. Short-term memory, also known as working memory, is regarded as the "workbench" of consciousness where cognitive processes occur, e.g., comparison, evaluation and transformation of representations, and therefore these cognitive processes are accessible to verbalizations, providing clues to cognition (Wickens et al., 1997).

Since verbal protocols do not require reconstruction or retrieval from long-term memory, there is no risk that memories and facts can be altered. Altered memories due to reconstruction and retrieval are often a problem in after-the-fact reporting, e.g., eyewitness accounts of accidents (Wickens & Hollands, 2000).

Task not conducive to verbalization: In some cases, tasks cannot be verbalized accurately because of parallelism and automaticity (Rasmussen & Jensen, 1974; Gordon, 1992). Because verbalization is serial, verbalizing may encourage participants to change a parallel task to a serial task so that it can be verbalized sequentially. Automaticity of a task occurs as expertise is gained through rehearsal and knowledge is moved to long-term memory, making the automated process inaccessible to verbalization (Wickens & Hollands, 2000).

Design is an example of a non-sequential activity that is linked to expertise (e.g., Dieter, 2000; Ullman, 2003), making design appear non-compatible with verbal protocols. However, what *can be* verbalized may pinpoint interesting phenomena for further study (Visser, 2006), and elicited differences in expert and novice designers can be valuable for training and educating engineers and designers.

3.3 Measures to mitigate associated risks

To mitigate the risks discussed above, the following precautionary measures are recommended for implementation in verbal protocol experiments.

Training*:* As "talking to yourself" can feel unnatural, participants should be trained to verbalize all thoughts prior to the main experiment through use of "warm up" exercises.

These exercises can be simple, e.g., arithmetic and word problems or practice problems similar to the experimental task (Atman et al., 2004). The goal of training is to habituate participants to verbalization (Ericsson & Simon, 1993).

Prompting: Despite training, participants may still forget to verbalize. If there are long silences, participants should be prompted to "keep talking" (Ericsson & Simon, 1993).

Encouraging free reporting: Participants should be encouraged to report thoughts as they occur to them, and not to plan their verbalizations nor to judge their thoughts (Bainbridge, 1991; Wickens et al., 1997).

Discouraging conversation: Participants should be discouraged from conversing with the investigators during the session, as conversations do not report cognition. This can be done by placing the investigator out of the participant's sight, e.g., in a different room or seated behind the participant.

3.4 Verbal protocols in design studies

Many design researchers acknowledge that verbalization studies are necessary and appropriate (Cross, 2006; Visser, 2006). Table 2 below summarizes some of the problem solving and engineering design studies using verbal protocol experiments.

Table 2: Survey of verbal protocol experiments.

Authors and Experiment Summary
Rasmussen & Jensen (1974) - Determined troubleshooting strategies of electronic technicians. Study involved 6 individuals on 8 different types of equipment.
Bhaskar & Simon (1977) - Modeled one individual's problem solving process in semantically rich domains, e.g., thermodynamics**.**
Atman & Bursic (1996) - Determined effects of reading a design text on design. Study involved 10 individuals: 5 who read a short design text and 5 who had not, prior to designing.
Benami & Jin (2002) - Modeled design cognition and investigated creative stimulation in design. Modeling study involved 4 engineering students designing for 30 minutes each. Creativity study involved 10 senior and graduate students.
Atman et al. (2004) - Compared freshman and senior students' design strategies**.** Study involved 32 freshmen & 61 seniors solving two short problems.
Nagai & Taura (2006) - Modeled the design synthesis process and investigated stimulation for creative design. Study involved 3 individuals performing 2 tasks for 10 minutes each.
Kim, Jin & Lee (2006) - Determined effects of personality on design creativity**.** Study involved 8 individuals: 4 experts and 4 students, designing for 60 minutes each.
Srinivasan & Chakrabarti (2009) – Evaluated concept novelty when designing within a framework. Study involved 8 individuals: 4 experts and 4 novices, using the framework.

We had originally used verbal protocol experiments to compare effects of opposite stimuli and similar stimuli on conceptual design in a small-scale experiment consisting of six participants designing for one problem (Chiu & Shu, 2008b). Our recent expanded study involved 14 participants. We describe this recent study in the next section.

4 EXPERIMENTAL WORK

This section summarizes our recent verbal protocol study and describes participants, experimental procedure and design, analysis and results.

4.1 Participants

This recent experiment consisted of 14 participants. All were fluent English speakers recruited from the Department of Mechanical and Industrial Engineering at the University of Toronto. Participants consisted of 13 males and one female, ranging from fourth-year undergraduate students to second-year Ph.D. students. Participants were paid $12 CAD upon completion of the hour-long experimental session.

4.2 Experimental Procedure and Design

Participants first completed three training problems to habituate them to verbalizing. The training problems included an arithmetic problem, a word problem and a problem similar to the main experiment problems. Then, participants were instructed to verbalize all thoughts as they completed a series of three design problems. Fifteen minutes were allotted for each problem.

Ten of the participants were provided with stimulus words, while four were not provided with stimulus words. Of the 10 stimulus participants, five switched stimulus type between problems. Table 3 summarizes the experimental conditions applied to each participant.

Table 3: Summary of experimental conditions applied to participants.

	Stimulus Participants										Control Participants			
Prob.#	TH	JS	VT	SW	JL	DRO	DR	UG	MM	DH	DL	JM	AF	AP
1	S	S	S	S	S	S	O	O	O	O	N	N	N	N
2	O	O	O	S	S	S	S	O	S	O	N	N	N	N
3	S	S	S	S	S	S	O	O	O	O	N	N	N	N

Problem#: 1 = Bushing, 2= Snow, 3 = Coal
S = Similar stimuli, O= Opposite stimuli, N= No stimuli

In our experiment, we were careful to incorporate the precautionary measures and to operate within parameters of other verbal protocol experiments as described above in Sections 3.3 and 3.4.

4.2.1 Problems and Stimulus Sets

The three problems provided to the participants were the bushing-and-pin problem, the snow insulation problem and the coal storage problem. The bushing-and-pin problem is a problem that should be familiar to most students within mechanical and industrial engineering. The snow insulation and the coal storage problem are general domain problems and not specific to any engineering discipline. The problems are summarized below.

Bushing problem: Parts that are automatically mated, e.g., a bushing and a pin, must be positioned so that their axes coincide. Using chamfers on mating parts does not solve the alignment problem. Develop a concept to center mating parts that does not require high positioning accuracy (Kosse, 2004).

Snow problem: In Canada, snow is readily available in the winters and has good insulating qualities due to the amount of air in it. However, if the snow is packed to the point it becomes ice, it is less insulating due to the loss of air. Come up with a concept to enable snow to be used as an additional layer of insulation for houses in the winter.

Coal problem: Clean coal and clean coal combustion technologies make it possible to generate cleaner electricity. That, combined with the increasing cost of oil and natural gas, power plant operators may consider converting or reconverting their power plants from oil or natural gas back to coal. However, there may not be enough land area near the plant that can be used for on-the-ground coal storage. Propose alternative solutions to a conventional coal pile (adapted from Dieter, 2000).

The stimulus sets for the opposite- and similar-stimulus conditions were generated using a combination of a thesaurus (Merriam-Webster.com) and WordNet (WordNet, 3.0). Some keywords did not have natural antonyms, e.g., "to insulate", so opposite stimuli were generated based on opposition to the problem itself, e.g., "to pack", as the problem stated that "packing" of snow is undesired. Hypernyms/hyponyms, or superordinate/subordinate words, of keywords were also used. In using the hypernym/hyponym hierarchy, it is common to encounter both synonyms and antonyms. For example, "rise" and "fall" are a synonym/antonym pair, but are also both hyponyms of "move", describing specific ways of moving. As generating opposite and similar verbs was not possible for all keywords, and antonyms/synonyms are sparse for verbs to start with, we used a combination of antonyms/synonyms and hypernyms/hyponyms for all three problems when necessary. Table 4 summarizes the stimulus sets.

4.2.2 Session transcription and concept identification

A professional transcriptionist was hired to transcribe design session recordings verbatim. The following is a transcript excerpt representing approximately 30 seconds of an experiment session. Transcript lines are numbered for referencing purposes.

1. *I think the obvious...*
2. *The first thing that comes to mind is that you'd like blanket the house...uh…*
3. *Essentially blanket the house in a layer, in a thin layer of snow…um…*
4. *"If the snow is packed to the point that it becomes ice."*
5. *I guess you'd obviously try to figure out what amount of packing you'd have to do…*
6. *To restrict the snow from becoming ice due to over packing.*

Table 4: Stimulus words.

	Keywords	Similar stimuli	Opposite stimuli
Bushing	**Similar:** Align and insert **Opposite:** Opposite of align and insert	Inject, transplant, sandwich, connect, skew, mount, misalign, attach, join, reorient, adjust, modify, match.	Change, disorder, disarrange, scramble, randomize, misalign, tumble, skew, move, expel, pull, eject, evict.
Snow	**Similar:** Insulate and surround **Opposite:** Pack and compact	Blanket, control, cover, defend, enclose, immerse, pack, preserve, prevent, restrain, restrict, submerge, touch.	Arrange, bundle, change, compress, constrict, contract, force, impact, move, push, squeeze, tighten, wad.
Coal	**Similar:** Store **Opposite:** Opposite of store	Accumulate, collect, displace, distribute, feed, give, heap, keep, place, supply, transfer, withhold	Abandon, discard, discharge, dispense, disperse, dispose, distribute, export, remove, scatter, spread, waste

Finished transcripts were corrected for minor spelling errors, e.g., "chamfer" for "camphor", "pedal" for "petal", but were otherwise not annotated nor changed.

An independent reviewer was recruited and paid to identify concepts from the free-form transcripts. The reviewer identified and summarized concepts with the aid of both the participant worksheets and the transcripts. A typical worksheet and experimental set up is shown in Figure 1.

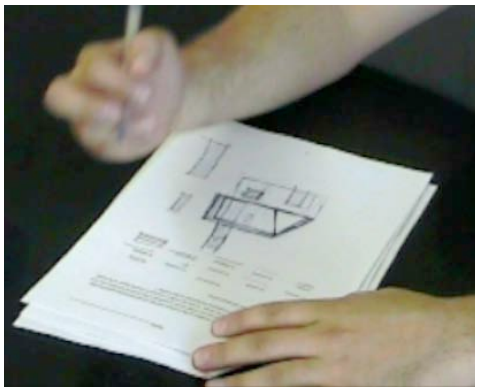

Figure 1: Typical experimental setup.

Concepts identified by the reviewer were compared with those identified by the investigators. A concept set consisting of 195 concepts was compiled for evaluation based on the union of the reviewer and investigator sets.

4.3 Concept metrics and evaluation

The original metric of interest was creativity and it was measured via three components: novelty, usefulness and cohesiveness. These creativity components were selected based

on review of creativity literature. Most creativity experts agree that creative artifacts are "novel" and "useful" (Torrance, 1974; Besemer & Treffinger, 1981; etc.). Novelty and functionality are generally agreed-upon engineering metrics (Shah et al., 2000; Brown, 2008). "Cohesiveness" was used to capture the idea that creative concepts should also be "elegant", "whole" and "detailed" (Torrance, 1974; Besemer & Treffinger, 1981).

Three independent raters were recruited to evaluate the concepts. The raters consisted of two males and one female, all familiar with conceptual design. Raters were not provided with identities of designers, nor the stimulus condition under which the concepts were generated. Concepts were presented to the raters in random order. Raters were trained with low anchor and high anchor concepts obtained from previous experiments and evaluated all concepts based on those anchors. Concepts were rated using the scale illustrated in Figure 2.

Low				Medium			High			
0	1	2	3	4	5	6	7	8	9	10
Not novel/useful/cohesive……Very novel/useful/cohesive										

Figure 2: Rating scale for scoring concepts.

The direct scaling method of obtaining human judgments is commonly used in psychophysics, a branch of psychology that deals with relating physical stimuli with cognitive phenomena (Engen, 1971). Such methods have been applied to evaluating the pleasantness of smells, emotions and beauty.

4.4 Analysis

Rater scores for each concept were averaged, and then all concepts from the same participant were aggregated to facilitate analysis. Aggregated scores were analyzed using a mixed-model ANOVA. Because five participants switched stimulus types between problems during the experiment (identified as TH, JS, VT, MM, DR in Table 3), pseudo-replicates were created to model these participants as independently contributing to each stimulus condition. This effectively increases the sample size from 14 to 19. This technique is used to deal with scenarios where not all participants contributed independently to only one experimental condition over multiple trials and results in a conservative estimate of differences (Duquette, 2009).

4.5 Results

Overall, opposite-stimulus concepts and no-stimulus concepts were judged to be more creative than similar-stimulus concepts. See Figures 3-5 for graphs of the results.

For the metric novelty, a significant main effect (i.e., $p < 0.05$) was found for Stimulus Type $F(2, 27.58) = 7.09$, $p = 0.003$, $p < 0.05$. Planned contrasts comparing individual experimental conditions, e.g., opposite-stimulus concepts versus no-stimulus concepts, show no significant novelty difference (i.e., $p > 0.05$) between opposite-stimulus concepts and no-stimulus concepts, but show a significant difference between opposite-stimulus and similar-stimulus concepts, $t(27.65) = -3.02$, $p = 0.0025$, $p < 0.05$. See Figure 3 for a graph of novelty results and Table 5 for results of planned contrasts.

For usefulness and cohesiveness, planned contrasts show that opposite-stimulus concepts are borderline significantly more useful and cohesive than similar-stimulus concepts, $t(31.51) = -1.61$, $p = 0.059$, $p \sim 0.05$ and $t(29.42) = -1.62$, $p = 0.058$, $p \sim 0.05$, respectively. Planned contrasts show no significant difference between opposite-stimulus and no-stimulus concepts. Problem Order, the order in which the problems were completed, was found to have an effect on cohesiveness and was corrected for in the planned contrasts. See Figures 4 and 5 for graphs of usefulness and cohesiveness results, and Table 5 for results of planned contrasts.

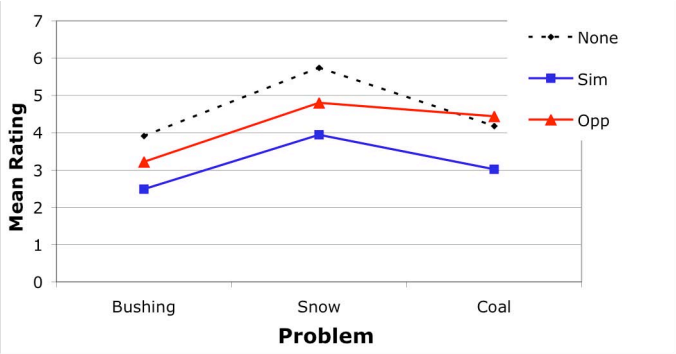

Figure 3: Mean novelty ratings for each problem.

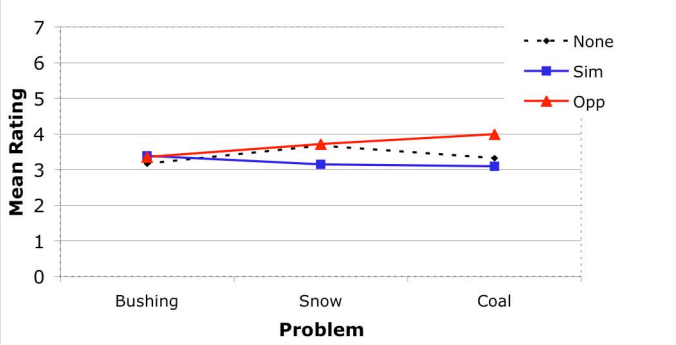

Figure 4: Mean usefulness ratings for each problem.

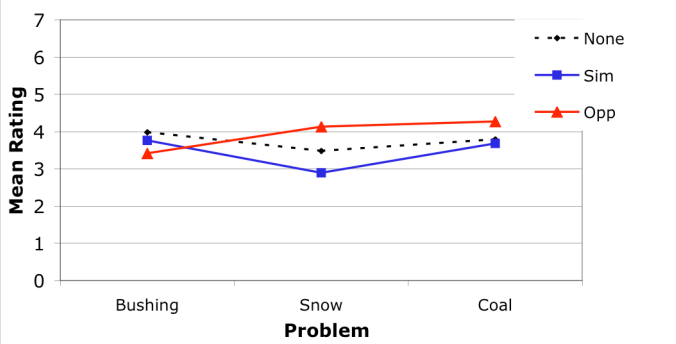

Figure 5: Mean cohesiveness ratings for each problem.

Table 5: Planned contrast results for concept scores.

	Estimated Mean Score (min. = 0, max. = 10)		Contrast t- and p-values
	Cond. 1 Mean Score	Cond. 2 Mean Score	
Novelty	**Opp**: 4.09	**None**: 4.26	t(28.456) = 0.32, p = 0.378, p > 0.05
	Opp: 4.09	**Sim**: 2.88	t(27.65) = -3.02, p = 0.0025, p < 0.05*
Use.	**Opp**: 3.75	**None**: 3.47	t(31.51) = -0.73, p = 0.24, p > 0.05.
	Opp: 3.75	**Sim**: 3.17	t(31.51)= -1.61, p = 0.059, p ~ 0.05*
Coh.	**Opp**: 4.09	**None**: 3.74	t(30.16) = -0.92**, p = 0.18, p > 0.05
	Opp: 4.09	**Sim**: 3.52	t(29.42)= -1.62**, p = 0.058, p ~ 0.05*

*Statistically significant ($p<0.05$), or borderline significant ($p\sim 0.05$). **Adjusted for effects of Problem Order.

Overall, the ANOVAs and planned contrasts support the original hypothesis that opposite-stimulus concepts are more novel, useful and cohesive, and therefore more creative, than similar-stimulus concepts. However, opposite-stimulus concepts and no-stimulus concepts were found to be equally creative.

5 DISCUSSION

In this section we will examine the literature and our experimental results with a focus on methodological issues related to verbal protocols. While we were able to support our original hypothesis that opposite stimuli would increase concept creativity compared to similar stimuli, other results of our experiment show there are no creativity advantages to using stimuli in concept generation. This appears to contradict much of the design literature and results from other design stimulation studies.

Possible explanations for our unanticipated results include random errors, errors that were unpredictable. Potential random errors include:

1. More creative individuals in the control group: This is unlikely as participants were randomly assigned to stimulus or no-stimulus conditions. However, it may be possible to control for individual creativity in future studies by administering a creativity assessment test, e.g., the Torrance Test of Creative Thinking (Torrance, 1974), prior to the main experiment.

2. Biased raters: Evaluating creativity is a difficult task and is based on the rater's own creativity and experience. It is possible that raters were biased in their judgments despite the training provided. However, all raters judged all concepts, and not just a subset of concepts, e.g., one rater did not only judge control concepts. Raters were also unaware of the identity of the participants; unaware of the experimental condition under which concepts were generated; and each scored concepts in a different, random, order.

Ruling out random errors, we turned our attention to other possible factors. As noted before, design experiments reporting increased concept creativity due to stimuli generally employed pen-and-paper methods. Re-visiting other design-specific verbal protocol experiments, such as those summarized in Table 2, we noticed that these verbal protocol experiments generally do not include a control condition. In some of the studies summarized in Table 2, an absolute control condition is either not necessary, or not possible. Studies used for exploratory and modeling purposes do not require a control, such as in the modeling studies of Rasmussen and Jensen (1974), Bhaskar and Simon (1977) and Benami and Jin (2002). In Kim et al. (2006), where they were examining design with respect to personality types, there is no "control personality" with which to compare results.

Other studies from Table 2 could be logically extended to include a control condition, such as the design stimulation studies (Benami & Jin, 2002; Nagai & Taura, 2006). However, the scope of these studies were defined such that effects of different stimulus types were compared relatively, i.e., stimulus *x* versus stimulus *y*, not absolutely, i.e., stimulus *x* versus no stimulus. For example, Benami and Jin (2002) modeled design cognition and then compared effects of different analogical stimuli (function, form, behavior and knowledge) on design outputs. Nagai and Taura (2006) modeled the design synthesis process and investigated the interpretation of closely related and distantly related noun-noun pairs on creativity within design synthesis.

Comparison of our experimental results and similar results of others, namely from pen-and-paper results, suggests that our unexpected results originated with the chosen experimental methodology. Specifically, we theorize there may be a higher cognitive workload for stimulus-condition participants who were required to design using stimuli while verbalizing. Thus, deteriorated performance was observed for stimulus-condition participants as compared to no-stimulus participants. Cognitive workload is concerned with measures such as how busy the operator is, task complexity and if the operator can handle additional tasks (Wickens & Hollands, 2000). It is known that increasing cognitive workload can deteriorate task performance. In our experiment, it is possible that stimulus-participant performance deteriorated due to the increased cognitive workload required to use stimulus words while designing *and* while verbalizing. To begin comparing the workload between the different experimental conditions, the number of aggregated tasks in each experimental condition can be compared. The tasks are enumerated in Table 6, where it can be seen that no-stimulus participants performed only two tasks as compared to stimulus participants who performed three tasks.

Table 6: Aggregated task comparison between no-stimulus and stimulus participants.

No-stimulus Tasks	Stimulus Tasks (Opposite or Similar)
1. Design	1. Design
2. Verbalize	2. Verbalize
-	3. Use stimuli

Increased demand on cognitive resources, such as increased reasoning and attention required to use stimulus words, can decrease performance, creativity in this case. Other examples of decreased performance due to increased demands on an operator include talking on the phone or texting while driving. In a recent study, it was found that drivers' reaction times increased by 9% when talking on the phone, and 30% when texting when compared to a driving-only control condition (Drews et al., 2009). Increased driver reaction times have an obvious detrimental effect on road safety.

Increased cognitive workload may be also non-ideal for creative concept generation, especially considering design is already a complex task. Tang and Zeng (2009) have been quantifying the cognitive stress of designers and they theorize there is an optimal cognitive stress level for designers, i.e., both too little stress and too much stress may be detrimental to design outcomes. Additional cognitive tasks imposed upon stimulus participants, i.e., verbalizing *and* using the stimulus words, may have increased these participants' stress level past the optimal level for creativity.

A review of the design literature shows another verbal protocol experiment with results similar to ours. In a design education experiment, Atman and Bursic (1996) were studying the effects of reading a short design text prior to the design task. In this experiment, five undergraduate students were provided with a short design text before a design task, and five were not provided with the text. Participants were then instructed to verbalize their thoughts as they designed. Atman and Bursic (1996) found that those provided with the design text prior to designing spent significantly more time on the problem, however, the "quality" of the design concepts, including fulfillment of functional requirements, were the same between the two groups.

While Atman and Bursic's (1996) study was not a design stimulation study per se, the design text could be comparable to stimulus, and increases in objective measures, e.g., time spent designing, should correlate to increases in concept quality. However, this was not the case as design quality was the same in both conditions. This other similar result supports our theory that the use of verbal protocols, rather than random errors, may have introduced limitations into our experiment.

It is unclear why the issue of cognitive overloading is not a generally discussed limitation of verbal protocols, unlike the issue of time and resource intensiveness. Our survey of design-specific verbal protocol experiments found that unequal cognitive workloads between experimental conditions generally do not exist, e.g., there is no control condition, so potential limitations due to unequal cognitive loading may not have been previously exposed. Other researchers with an interest in verbal protocols may not have encountered this potential limitation because design is a very complex, open-ended task; likely more complex than other tasks typically studied using verbal protocols, e.g., human-machine interaction (Bainbridge et al. 1968; Obata et a., 1993). It is also possible that previous contradictory results have not been reported.

6 RECOMMENDATIONS AND GUIDELINES

Despite the potential design-specific limitations of verbal protocols, we recommend the continued use of verbal protocols for studying design cognition. However, we also recommend careful attention be paid to experimental design. Below are guidelines we have compiled based on the literature and our own experimental experiences.

- Verbal protocols are appropriate for:
  - Exploratory studies and cognitive modeling;
  - Comparison of relative results between two or more equal workload conditions, e.g., opposite versus similar stimulus, closely related versus distantly related stimulus.

- Verbal protocols are *not* appropriate for:
  - Comparison of absolute results, e.g., opposite stimulus versus no stimulus;
  - Comparison of relative results between possibly unequal workload conditions, e.g., semantic versus pictorial stimuli.

- Experiments using verbal protocols should:
  - Incorporate precautionary measures listed in Section 3.3;
  - Be used in conjunction with other methods, e.g., pen-and-paper, observation, etc. (Visser, 2006).

More specific guidelines can be developed based on the results of future work suggested in the next section.

7 FUTURE WORK

Future work can be performed to further determine and quantify the extent of potential methodological limitations. Two approaches are suggested: 1) determine and compare cognitive workloads and 2) extend the current experimental design to include a "dummy" comparison condition. These approaches can be used separately or together.

In the first approach, the cognitive workload of different experimental conditions can be determined. Empirical subjective techniques can be used to collect data from participants while they are designing. A well-known method for evaluating cognitive workload involves using the Cooper-Harper scale that measures workload in a single dimension on a standardized scale (Charlton, 1996). Other established methods, such as the NASA Task-Loading Index (NASA TLX) and Subjective Workload Assessment Technique (SWAT)

measure subjective workload on multiple dimensions, e.g., physical, emotional, cognitive, etc. (Hart & Staveland, 1988; Charlton, 1996). In all cases, these subjective techniques are relatively easy to administer as participants can indicate their perceived workload on a survey during the task, after task segments, or even after the entire task. In terms of objective techniques, a timeline analysis can be used to generate a workload profile. This is obtained by summing the number of tasks at a specific point in time (Charlton, 1996).

In the second approach, a "dummy-stimulus" condition can be added to determine if performance decreases similar to those observed with similar stimuli occur for the dummy-stimulus condition. A dummy-stimulus condition can be added where stimulus words are generated randomly rather than based on opposite or similar relationships. If all stimulus concepts, i.e., concepts generated using opposite, similar and random stimuli, are scored as equally creative or less creative than no-stimulus concepts, this will further support our theory that verbalization can affect the results of design stimulation studies in some cases. However, if there is a significant difference between random-stimulus concepts and all other concepts, this may indicate issues with the specific type of stimuli used, i.e., related semantic stimuli.

Additionally, a comparative study could be undertaken to investigate differences between pen-and-paper and verbal protocol studies. While some differences between the two methodologies may appear intuitive, e.g., pen-and-paper to obtain statistically significant results and verbal protocol studies to elicit cognitive processes, quantitatively comparing the two methods may contribute to further development of appropriate methods to study and better understand design.

8 SUMMARY AND CONCLUDING REMARKS

Originally, we were using verbal protocols to study design cognition, specifically to determine how designers used opposite- and similar-stimulus words. Our hypothesis that opposite-stimulus concepts would be judged as more creative than similar-stimulus concepts was supported. However, we also found that no-stimulus concepts were judged to be equally as creative as opposite-stimulus concepts, and thus more creative than similar-stimulus concepts. This result contradicts our intuition and the design literature, which generally supports the use of design stimuli to increase design creativity.

This contradictory result prompted us to re-examine verbal protocols both in the literature and in our own experiments. Based on reported results and our experiences, we theorize that verbalizing while designing decreases designer performance under conditions where cognitive workload is increased, e.g., use of stimulus.

In light of potential limitations imposed by verbal protocols, we developed guidelines for more appropriate applications of verbal protocols in design experiments. We have also proposed future work to further identify and quantify design-specific limitations related to verbal protocols. Despite potential limitations, verbal protocols currently remain a practical and valuable method for furthering our understanding of design.

ACKNOWLEDGEMENTS

We wish to thank all the participants, raters and reviewers. We also wish to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support.

REFERENCES

Alexiou, K., Zamenopoulos, T., Johnson, J.H., 2009, Exploring the neurological basis of design cognition using brain imaging: some preliminary results, Design Studies 30:623-647.

Altshuller, G.S., Shulyak, L., 1996, *And Suddenly the Inventor Appeared: TRIZ, the Theory of Inventive Problem Solving, 2/e*, Technical Innovation Center, Worcester, MA.

Atman, C.J, Bursic K., 1996, "Teaching Engineering Design: Can Reading a Textbook Make a Difference?" *Research in Engineering Design*, 7/7:240-250.

Atman, C.J., Cardella, M.E., Turns, J., Adams, R., 2004, Comparing freshman and senior engineering design processes: an in-depth follow-up study, Design Studies, 26:325-357.

Bainbridge, L., Beishon, J., Hemming, J.H., Splaine, M., 1968, A Study of Real-Time Human Decision-Making Using a Plant Simulator, OR, Special Conference Issue: Decision-Making 19:91-106.

Bainbridge, L., 1991, Chapter 7: Verbal Protocol Analysis, in J.R. Wilson and E.N. Corlett's (eds.) *Evaluation of Human Work*, Taylor & Francis.

Benami, O., Jin, Y., 2002, Creative stimulation in conceptual design, ASME DETC/CIE, Montreal, Canada, DETC2002/DTM-34023.

Besemer S.P., Treffinger D.J., 1981, Analysis of Creative Products: Review and Synthesis, J. Creative Behavior, 15:158-178.

Bhaskar, R., Simon, H. A., 1977, Problem solving in semantically rich domains: An example from engineering thermodynamics, Cognitive Science, 1:193-215.

Brereton, M., McGarry, B., 2000, An Observational Study of How Objects Support Engineering Design Thinking and Communication: Implications for the design of tangible media, CHI, 2/1:217-224

Brown, D.C., 2008, Guiding Computational Design Creativity Research, in J. Gero's (ed.) Studying Design Creativity, Springer.

Charlton, S. G., 1996, Mental Workload Test and Evaluation, in T.G. O'Brien and S.G. Charlton's (eds.) *Handbook of Human Factors Testing and Evaluation*, Earlbaum.

Chiu, I., Shu, L.H., 2008a, Use of opposite-relation lexical stimuli in concept generation, Annals of the CIRP 57/1:149-152.

Chiu, I., Shu, L.H., 2008b, Effects of Dichotomous Lexical Stimuli in Concept Generation, ASME IDETC/CIE, New York City, NY, USA, August 3-6, 2008. DETC2008-49372 (DTM).

Chiu, I., 2010, Quantifying Oppositely and Similarly Related Semantic Stimuli on Concept Creativity, University of Toronto, Ph.D. Thesis.

Cross, N., Christiaans, H., Drost, K., 1996, Introduction: The Delft Protocols Workshop, in N. Cross, H. Christiaans, K. Drost's (eds.) Analysing Design Activity, John Wiley & Sons, West Sussex, UK.

Cross, N., 2006, *Designerly Ways of Knowing*, Springer-Verlag, London, UK.

De Bono, E.,1970, Lateral thinking: creativity step by step, Harper & Row.

Deubel, P., 2003, An investigation of behaviorist and cognitive approaches to instructional multimedia design, Journal of Educational Multimedia and Hypermedia 12/1:63-90.

Dieter, G.E., 2000, *Engineering Design: A Materials and Processing Approach, 3rd Edition*, McGraw-Hill, NY.

Dong, A, 2006, Concept formation as knowledge accumulation: a computational linguistics study, Artificial Intelligence for Engineering Design Analysis & Manufacturing, 20/1:35-53.

Drews, F. A., Yazdani, H., Godfrey, C.N., Cooper, J.M., Strayner, D.L., 2009, Text Messaging During Simulated Driving, Human Factors: The Journal of the Human Factors and Ergonomics Society, accessed online at http://hfs.sagepub.com on January 17, 2010.

Duquette, L., 2009, Statistical Consultant, Department of Statistics, University of Toronto, Personal communication.

Engen, T., 1971, Psychophysics II. Scaling Methods. In J.W. King, L.A. Rigg, Woodforth & Schlossber's (eds.) *Experimental Psychology, 3rd ed.,* Holt Rinehart & Winston

Ericsson, K. A., Simon, H.A., 1993, *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge, MA.

Ericsson K. A., 2002, Protocol analysis and Verbal Reports on Thinking, An updated and extracted version from Ericsson, accessed online at http://www.psy.fsu.edu/faculty/ericsson/ericsson.proto.thnk.html

Gero, J.S., McNeill, T., 1998, An approach to the analysis of design protocols, Design Studies, 19:21-61.

Gordon, S.E., 1992, Implications of Cognitive Theory for Knowledge Acquisition, in R.R. Hoffman's (ed.) *The Psychology of Expertise, Cognitive Research and Empirical AI*, Springer-Verlag, New York, NY.

Gordon, W. J. J., 1961, *Synectics, The Development of Creative Capacity*, Harper & Row, New York, NY.

Hart, S.G, Staveland, L.E.,1988, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock, N. Meshkati (eds.) *Human Mental Workload*, 139-183. Elsevier Science: NorthHolland

Hubka, V., Eder W. E., 1996, *Design Science- Introduction to the Needs, Scope and Organization of Engineering Design Knowledge*, Springer-Verlag, London, UK.

Kim, Y.S., Jin, S.T., Lee, S.W., 2006, Design Activities and Personal Creativity Characteristics: A Case Study of Dual Protocol Analysis Using design Information and Process. ASME IDECT/CIE, Philadelphia, PA, DETC2006-99654(DTM).

Köhler, W., 1959, Gestalt Psychology Today, American Psychologist, 14:727-734.

Kosse, V., 2004, *Solving Problems With TRIZ: An Exercise Handbook, 2nd Edition*, Ideation Int'l Inc., Southfield, MI.

Lutfey, K.E., Campbell, S.M., Renfrew, M.R., Marceau, L.D., Roland, M., McKinlay, J.B., 2008, How are patient characteristics relevant for physicians' clinical decision making in diabetes? Social Science & Medicine, 67:1391–1399.

Merriam-webster.com, 2008, *Merriam-Webster Online Dictionary*, http://merriam-webster.com

Nagai, Y. and T. Taura, 2006, Formal Description of Concept-Synthesizing Process for Creative Design, in *Design Computing and Cognition '06*, Ed. J.S. Gero, 2006.

Nisbett, R., Wilson, T., 1977, Telling more than we can know: Verbal reports on mental processes, *Psychological Review*, 84:231-259.

Obtata, T., Daimon, T., Kawashima, H., 1993, A Cognitive Study of In-vehicle Navigation Systems : Applying Verbal Protocol Analysis to Usability Evaluation, Proceedings of IEEE – IEEE Vehicle Navigation & Information Systems Conference, Ottawa.

Rasmussen, J., Jensen, A., 1974, Mental procedures in real life tasks. A case study in electronics trouble shooting, Ergonomics, 17:293-30.

Rittel, H.W.J., Webber, M. M., 1984, "Planning Problems are Wicked Problems", in N. Cross' (ed.) *Developments in Design Methodology*, John Wiley & Sons, Bath, England, pp 135-144.

Segers, N., 2004, Computational representations of words & representations of words & associations in architectural design, development of a system support creative design, Bouwstenen 78, Technische Universiteit Eindhoven, Ph.D. Thesis.

Shah, J., Kulkarni, S., Vargas-Hernandez, N., 2000, Evaluation of Idea Generation Methods for Conceptual Design: Effectiveness Metrics & Design of Experiments, J. Mech. Des., 122:377-384.

Simon, H., 1969, *The Sciences of the Artificial,* MIT Press, Cambridge, MA.

Srinivasan, V., Chakrabarti, A., 2009, An Empirical Evaluation of Novelty-Sapphire Relationship, ASME IDETC/CIE, San Diego, CA, USA, August 30-September 2, 2009. DETC2009-86668 (DTM).

Tang, Y., and Zeng., Y., 2009, Quantifying Designer's Mental Stress in the Conceptual Design Process Using Kinesics Study, Proc of International Conference on Engineering Design, ICED'09, Stanford University, Stanford, CA, August 24-27, 2009.

Thomas, J.C., Carroll, J.M., 1984, The Psychological Study of Design, in N. Cross's (ed.) *Developments in Design Methodology*, pp 221-235.

Torrance, E.P.,1974, *Torrance Tests of Creative Thinking*, Scholastic Testing Service, Inc.

Tseng, I., Cagan, J., Moss, J. and Kotovsky, K., 2008, Overcoming Blocks in Conceptual Design: The Effects of Open Goals and Analogical Similarity on Idea Generation, ASME IDETC/CIE, Brooklyn, NY, Aug 3-6, DETC2008-49276(DTM).

Wickens, C.D., Gordon, S., Liu Y., 1997, Introduction to Human Factors Engineering, Longman, New York, NY.

Wickens, C.D., Hollands, J.G., 2000, Engineering Psychology and Human Performance, 3rd Edition, Prentice Hall, Upper Saddle River, NJ.

WordNet, 3.0, http://www.cogsci.princeton.edu/~wn.

Ullman, D., 2003, *The Mechanical Design Process, Third Edition*, McGraw-Hill, New York, NY.

Visser, W., 2006, *The Cognitive Artifacts of Designing*, Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

Yang, M.C., Cham, J.G., 2007, An analysis of sketching skill and its role in early stage engineering design, J. Mech. Des., 129:476-482.

Zeng, Y., 2009, Environmentally Based Design : A Method for Innovative Design, Chalmers Design Seminar, University of Toronto, March 18, 2009.

