ATLANTA

120th ASEE Annual Conference & Exposition
FRANKLY, WE DO GIVE A D*MN
June 23-26, 2013

Paper ID #6178

# Development and Initial Validation of an Innovation Assessment

**Dr. Geoff Wright, Brigham Young University**

Dr. Geoffrey A. Wright is an assistant professor of Technology and Engineering Education in the College of Technology and Engineering at Brigham Young University.

**Mr. Paul T Skaggs, Brigham Young University**
**Mr. Jacob Dean Wheadon, Purdue University**
**Dr. Clifton B. Farnsworth, Brigham Young University**

Clifton Farnsworth received B.S. and M.S. degrees in civil engineering from Brigham Young University and a Ph.D. in civil engineering from the University of Utah. He worked as a geotechnical engineer for eight years with the Utah Department of Transportation, spent three years as an Assistant Professor of civil engineering at The University of Texas at Tyler, and has a current appointment as an Assistant Professor of construction management at Brigham Young University.

# Development and Initial Validation of an Innovation Assessment

XXXX
XXXX University
United States of America
XXX@XXX.edu

## Introduction

*The Need for Innovation*

In industry and education, there is an increasing push for organizations and individuals to be more innovative (Wagner, 2010; Fagerberg, 1999). Rapid technological change has created the need for organizations and individuals to adapt quickly (Christensen and Eyring, 2011). Christensen (1997) described how disruptive innovations fundamentally change markets and require new ways of thinking for organizations to adapt and survive. He described how individuals in organizations need to think differently in order to compete in today's marketplace. Because of the rapid rate of technological change that is occurring today, disruptive innovations are changing markets even faster than in the past. This has led to a greater need for people to cultivate innovation skills.

Innovation skills are also needed to create job growth. Drucker (1985) showed that innovation has been the leading source of job creation in the United States over the last century. He called for organizations and individuals to focus their efforts on creating new value in society, both for their own good, and for the good of society in general. These calls have been echoed by politicians (Obama, 2011), economists (Friedman and Mandelbaum, 2011), and educators (Wagner, 2010).

In order to keep up with the demand for innovation education, educators at a private university in the western United States have developed a course focused on teaching innovation. The course, titled the Innovation Bootcamp teaches technology and engineering students many behaviors and processes of innovation that have been identified in past literature (Howell et al., 2011). At the Innovation Bootcamp, students learn tools that help them work through the five parts of the innovation model (as defined by the Innovation Bootcamp curriculum): idea finding, idea shaping, idea defining, idea refining, and idea communicating.

*The Need to Assess Innovation Teaching*

Using this model, educators have taught the Innovation Bootcamp since 2008. They have performed preliminary studies (Howell et al., 2011; Wright et al., 2010) and feel confident that the course is having a positive impact on the innovation skills of the students, even though they did not have a test to evaluate the impact the course was having on students' ability to innovate. Consequently, they, and other innovation educators, need an assessment of students' innovation skills that can be used as a pre- and post- test to see if a student is more innovative as a result of participating in the Innovation Bootcamp. Having an innovation test would be very useful for improving teaching in this particular course, and it is hoped that such a test will have value for anyone seeking to teach innovation.

*Current Innovation Assessments*

In an attempt to address this need, Lewis (2011) reviewed existing innovation and creativity tests and relevant literature. His study found that existing test instruments were lacking in two major areas. The first is that existing tests do not cover the whole process of innovation – focusing only on either creativity or implementation. He found that creativitycentric tests measure divergent thinking, while existing innovation tests focus primarily on convergent thinking.  Lewis states that this is problematic because innovation involves both divergent and convergent thinking.  He also suggested that the other issue of the innovation tests was that they only measured the performance of a product, team, or organization, and did not account for, or measure, the abilities of an individual.  This does not allow educators to see how their instruction changes a student's ability to innovate.  In order to meet the needs of the Innovation Bootcamp and other innovation educators, a test that measures an individual's ability to do activities across a greater part of the innovation process is needed.

### *Purpose Statement*

The purpose of this project was to develop an innovation test instrument and perform an initial validation.  The test needed to cover a broader range of innovation skills defined by the Innovation Bootcamp curriculum and needed to evaluate individual students' abilities at performing each of the tasks outlined therein.  This paper describes the development of the test, including analysis of the content domain, identification of the learning outcomes, item creation, testing of the test, and initial validation.

### *Innovation Models and Processes*

Because of the need to assess a person's skill at specific parts of the innovation process, it is important to describe the innovation processes and models used by leading innovation educators and consultants.  Although the different practitioners use varying language to describe their processes, there were many common elements and similarities across the different processes.  These common elements are found in the Innovation Bootcamp model.  Because the different groups use similar models and processes, future studies should be done to see if this instrument could be used more generally in innovation education.

### *The Innovation Bootcamp Model*

The five parts of the Innovation Bootcamp model (see Figure 1) are: Idea finding, idea shaping, idea defining, idea refining, and idea communicating.

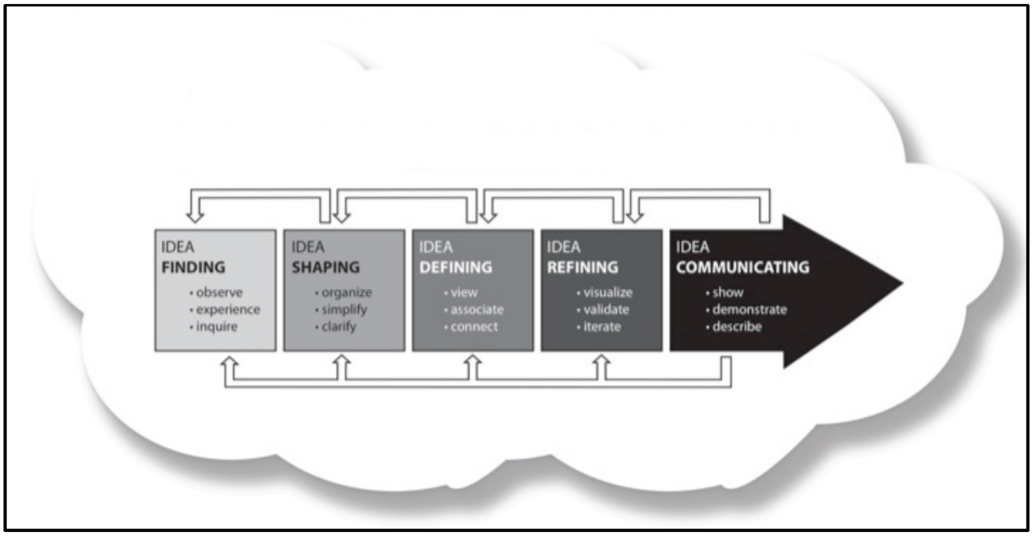

**Figure 1: Innovation Bootcamp Model**

Idea finding involves teaching students to see opportunities for innovation in the world around them. Students in the Bootcamp are taught to take on the role of anthropologist as they observe people. They are taught to actively experience what others are experiencing as to find issues that can be improved upon. Kelley (2005) suggests this approach to innovation and explains how it is used at IDEO. This ties closely to the empathize step in the Stanford D-School innovation process (Stanford, 2010) and also to the behavior of observation described by Dyer et al. (2011). When students actively observe the situations and people around them, they learn to identify opportunities for innovation.

The second part of the Innovation Bootcamp model is idea shaping. In idea shaping, students refine their observations from the idea finding phase. This relates to the define phase of the Stanford D-School model (Stanford, 2010). Stanford describes this as a time to "develop a deep understanding of your users and the design space." The major behavior of this phase is questioning (Dyer et al., 2011) The goal of this phase is to develop a clear, actionable problem statement. This problem statement guides the rest of the innovation process and gives focus to the participants.

The first 2 parts of the Innovation Bootcamp model comprise what Runco (2006) calls problem-finding. He says that there are multiple problem-finding skills, two of which are "problem discovery" and "problem definition." He cites Getzels' (1975) claim that "the quality of a problem determines the quality of a solution." Problem finding is a critical part of innovation and creates a foundation for the rest of the process.

Idea defining is the third part of the Innovation Bootcamp model. This phase begins the creation of solutions to the problem defined in earlier phases. In this phase, students learn about various methods of ideation and are encouraged to generate a large number of diverse ideas, which Runco (2006) calls fluency. They are taught that they are more likely to have good ideas if they generate many ideas.

Different practitioners use different tools and activities to help ideate. Many of the processes focus on associative thinking (Dyer et al., 2011; Runco, 2006; Mednick, 1962), combining different ideas (often from different fields) to solve the given problem. Dyer et al. (2011) consider associative thinking to be the crux of innovation. They claim that the other behaviors (observation, questioning, idea networking, and experimenting) either lead to creating associations, or are used to develop associations that have been made.

After the idea defining phase, the solutions need to be prioritized and refined. Idea refining is the next phase. It is similar to the idea shaping phase, but rather than narrowing down observations to a problem statement, students narrow down solution ideas to a single solution. This is done by choosing the best ideas among those generated in the idea defining phase and by prototyping and testing them. Experimentation is one of the behaviors identified by Dyer et al. (2011) and fits with the prototype and test steps in the Stanford D-School (2010) process. This phase is important because when many potential solutions are generated in previous steps, the best solutions need to be chosen and refined. By testing the solutions, innovators can see which ones work and how to improve them.

The last phase in the Innovation Bootcamp model is the idea communicating phase. Lewis (2011) found that this is the one part of the process that is unique to this particular model. Although this phase is not explicitly mentioned in the other processes studied, it is implied in all of them. All other practitioners of innovation communicate, share, or sell their innovations to others. Rogers (2003) described how innovations diffuse through a community and showed that communication is central to that diffusion. He described how innovations are adopted early by a small part of the population and are diffused to others in the population by various modes of communication.

This review describes the lack of adequate innovation assessments to evaluate the innovation skills of individuals. Current tests either do not test subjects on the whole process of innovation, or do not test an individual's skills. This review also describes the process of innovation that the test will cover. It serves as an explanation of the content domain of the assessment, which is necessary to creating a table of specifications and developing the test.

**Methods**

The creators of the Innovation Bootcamp curriculum identified five major phases in the process of innovation. These are described in detail in the literature review of this paper. These phases are used to identify the learning outcomes for the Innovation Bootcamp.

The first learning outcome is that students identify opportunities for innovation from a given context. This outcome combines the first two phases of the Innovation Bootcamp model under the umbrella of problem finding. Runco (2006) describes how there are various tools and techniques that fall under problem finding. By focusing on the broader outcomes rather than the particular tools and skills, students can use whatever techniques they want. This outcome focuses on seeing if a student can identify opportunities for innovation, without concern for how they do it.

The second learning outcome is that students create many and varied ideas to solve problems. This outcome tests an individual's fluency, the ability to come up with many and varied ideas. Fluency has long been an indicator of an individual's level of creativity. By using fluency in a specific context (problem solving, in this case), this outcome targets an individual's ability to create ideas that are useful in the given context.

The third outcome for the course is that students evaluate the ideas based on originality, usefulness, and feasibility. In the Innovation Bootcamp curriculum, innovation is defined as original and useful ideas, implemented successfully. Students should be able to decide whether the ideas they have had fit that definition in order to know which ideas to focus on as they refine and experiment with the ideas.

The fourth outcome is that students can effectively communicate their ideas to others. They should be able to clearly and concisely explain the need for their innovation and the benefit of it. They need to convince readers of the value and impact their innovation will add to those who use it.

*Item Development*

After creating the table of specifications, items for each outcome were created. This assessment had four item types. The first item type corresponded to the first learning outcome and tested students' ability to identify problems from a photograph. The students were graded on their ability to identify as many problems as possible in the photograph. Answers were given higher weight if they are less common.

The second type of item is similar to the first. It corresponds to the second learning outcome. Students were given a problem statement and asked to generate as many solutions as possible. They were also given more points for answers that come up less frequently than others. This grading scheme is used in other instruments to measure creativity (Torrance, 1969). Torrance uses shapes that students identify and elaborate on and they are awarded for having many answers and unique answers. The difference between the items for this innovation assessment and the items in Torrance is that these items are focused on problems that people have with their products or environments. So where Torrance items show an abstract shape, these items show an actual problem that could be solved.

The first two items types for this test were designed to be easily changed for future tests. In the first, a subject generates problem statements from a given photograph, and in the second, a subject generates solutions from a given problem statement. It was expected that it would be difficult to achieve equivalent item difficulties for these items on the first attempt. Subjects would likely find it easier or harder to find problems (or solutions) based on the given stimuli. For this reason the items have been designed to be easily modified for future testing. With this item design, photographs (or problem statements) can be easily switched out and tested until equivalently difficult stimuli can be found. In this study, the current items were be tested to see how equivalent they were. Future studies can then easily modify the items to get better equivalence, if needed.

The third type of question tests the students' ability to evaluate ideas. In the innovation process, students come up with many ideas to solve a certain problem. After they have generated those ideas, they have to decide which ideas to pursue and refine. The ability to decide which ideas will be best is what is tested in the third type of items. In this item type students were given a problem statement and four possible solutions. They were asked to rank the solutions according to the definition of innovation used by the Innovation Bootcamp: Original and useful ideas that can be implemented successfully. Their rankings were matched against the rankings that the University's Industrial Design faculty made.

In order to create a key for the innovation ranking items, five Industrial Design professors were polled using the items from the assessment, which include the criteria for ranking the innovations. The key was made by giving points to the innovations that the professors ranked highly. With the totaled scores, an overall ranking could be created that combined the rankings of all the professors. Then the students' rankings could be compared to overall rankings when the tests were scored.

The fourth item type tests the students' abilities to communicate their ideas to others. In this item they are asked to create a pitch for the innovation that they ranked first in the previous ranking item. The pitches need to be concise, persuasive, and need to communicate the value of the innovation. In the test, the students are limited to 700 characters in order to maintain conciseness and are graded on persuasiveness and ability to communicate the value of their innovation.

In order to grade this item, two raters were used. Raters followed a provided rubric (see Table 1). Raters were trained on how to use the rubric and then graded five questions and discussed any areas that they disagreed upon. After the first five responses and their discussion, the raters graded five more responses and discussed the scores until raters achieved a correlation greater than 0.75, which is considered an "excellent" level of inter-rater reliability (Cicchetti, 1994).

**Table 1: Rubric for Communicate Items**

<table>
<tr><td colspan="3"><b>Explain problems</b>: How well does this explain the problem?</td></tr>
<tr><td>Fails to explain the problem<br>0</td><td>Alludes to the problem<br>1</td><td>Clearly explains the problem<br>2</td></tr>
<tr><td colspan="3"></td></tr>
<tr><td colspan="3"><b>Explain solutions</b>: How well does this explain how the solution works or solves the problem?</td></tr>
<tr><td>Fails to explain the solution<br>0</td><td>Explains, but not clearly<br>1</td><td>Clearly explains the solution<br>2</td></tr>
<tr><td colspan="3"></td></tr>
</table>

| **Persuasiveness**: How well does this convince you of the benefit of the innovation (overall score)? | | |
|---|---|---|
| This does not convince me<br>0 | This makes me interested<br>1 | This convinces me<br>2 |

*Testing Procedures*

In order to collect initial evidence of validity and form equivalence of the instruments, the test was administered to the students of the Innovation Bootcamp from winter semester 2012. During this semester there were five sections of the Bootcamp with 20 students in each section. As a preliminary check, the first three sections received the test. After they responded, the results were analyzed and revisions were made to the test. The revised test was then given to all 100 students from all sections of the Bootcamp from winter semester. For the full test, students were instructed that the test would be a contest. The students competed for prize money that would be awarded to the students with the highest scores on the test. This was done in order to raise the stakes for the test enough to prompt maximum performance. Then the full test results were analyzed, and suggestions for future studies were made.

In this study, various types of validity evidence were gathered. Content-related evidence was gathered as part of the review of the literature, the comparison of the Innovation Bootcamp with other innovation models, and the description of the alignment between the Bootcamp curriculum and the ITI. Construct-related evidence was addressed in the revisions that were made between the two rounds of testing, and the description of the methods could be used as initial evidence that could support construct validity once other studies have been performed. Some evidence of face validity was observed through students' enthusiasm for the test and curiosity about the test and how it was graded. Criterion-related evidence was gathered indirectly, with informal observations that connected high test performance to high performance in the Bootcamp. Because the results of this test will have no impact on the students taking it, consequence-related evidence was not a major issue in this study.

*Revisions to the ITI After Initial Test*

After the first round of testing, the results were analyzed and revisions to the ITI were made in order to improve the test. These revisions were made to address three major issues: Lack of high performance, lack of variation in some responses, and problems with the communicate items.

*Lack of High Performance*

The biggest problem with the initial test was that the subjects did not achieve high performance. Many students failed to finish, and the open-ended responses were often too short to evaluate the students' skills. It was hypothesized that this was the result of test fatigue based on a comparison of the mean scores of the two forms for each group (see Table 2). When group one took form one and then form two, the mean dropped from 45 to 31. Group two took the tests with the form order reversed, and their mean dropped from 52 to 46. This order effect was remedied by making the test shorter.

**Table 2: Summary of Overall Scores**

| | Overall totals | Total from 1 | Total from 2 |
|---|---|---|---|
| 1->2 Group | 158 | 78 | 80 |
| | 119 | 64 | 55 |
| | 109 | 59 | 50 |
| | 91 | 53 | 38 |
| | 76 | 44 | 32 |
| | 72 | 39 | 33 |

| | Overall totals | Total from 1 | Total from 2 |
|---|---|---|---|
| 2->1 group | 166 | 85 | 81 |
| | 162 | 77 | 85 |
| | 128 | 62 | 66 |
| | 118 | 52 | 66 |
| | 114 | 44 | 60 |
| | 104 | 57 | 57 |

| | | | |
|---|---|---|---|
| | 68 | 41 | 27 |
| | 67 | 49 | 18 |
| | 52 | 29 | 23 |
| | 41 | 29 | 12 |
| | 32 | 29 | 3 |
| | 25 | 25 | 0 |
| Mean | 75.83 | 44.92 | 30.92 |
| SD | 36.95 | 15.67 | 21.88 |
| Correlation | .93 | | |

| | | | |
|---|---|---|---|
| | 100 | 31 | 60 |
| | 91 | 42 | 58 |
| | 79 | 37 | 42 |
| | 55 | 34 | 20 |
| | 54 | 35 | 20 |
| | 7 | 0 | 7 |
| Mean | 98.17 | 46.33 | 51.83 |
| SD | 43.58 | 21.60 | 23.60 |
| Correlation | .86 | | |

The second hypothesized reason for inadequate performance was the lack of incentive for students to prompt high performance. In order to resolve this issue, the second round of testing was implemented as a competition. Prizes were offered to subjects who scored more highly on the tests. The highest-scoring subject would receive $100, the next two highest would receive $50, the next two received $25 and the next ten received vouchers for a free smoothie. This would presumably be enough of an incentive to prompt maximum performance.

*Lack of Variation in Responses to Problem-Finding Items*

Another problem in the initial test was the lack of varied responses in some of the items. In the first version of the test, photographs of problem scenarios were used in the problem-finding items as stimuli for the subjects to find problems. Students were prompted to identify as many problems in the photograph as possible. For example, subjects were given a picture of a person sleeping on one of the public couches on campus (see Figure 2). Student responses were limited to the identification of only a few problems.

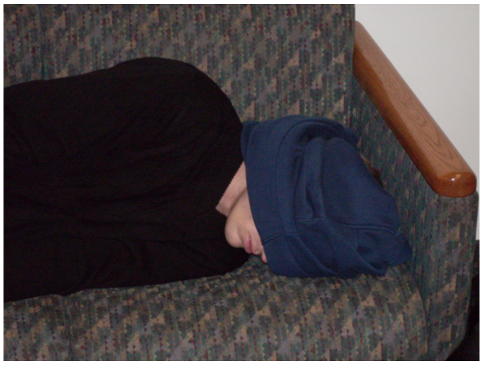

**Figure 2: Example of Problem-Finding Photograph**

In order to solve this problem, the new version of the test had wider-angle photographs of rooms (see Figure 3). This gave the subjects more opportunities to notice a larger number of problems. The hypothesis was that giving the subjects more to look at would allow for a greater variety of answers and give researchers a better idea of the subjects' ability to find problems. By fixing this problem, evidence of construct validity was strengthened because the item was more able to better target varying levels of the construct.

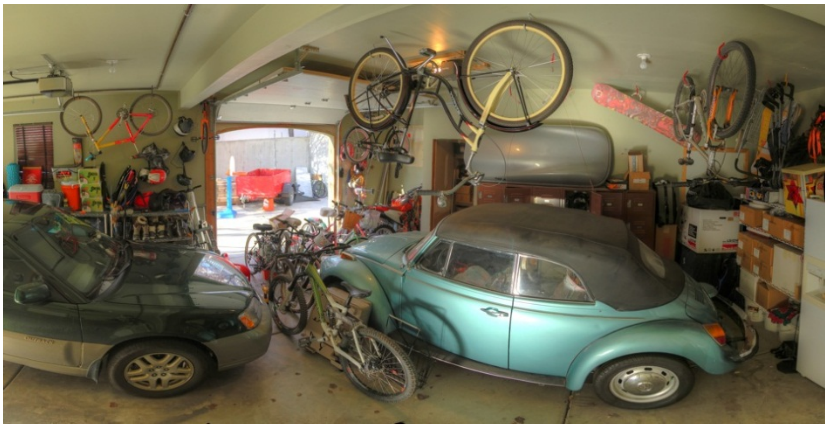

**Figure 3: Example of Revised Test Photograph**

*Communicate Items*

Another problem with the initial instrument was in the communication questions. It was evident from many of the answers that the students did not understand what was required. Many failed to describe the problem or solution well. They had a hard time describing what the problem was that they were trying to solve. They also did not realize that they needed to describe how the solution worked. This may have been because they were creating a pitch for a given solution. To fix this, the communicate items' instructions were revised and the items were moved to follow the solution-generating items. Rather than trying to pitch a solution that was given to them, the subjects were now asked to pitch their favorite of the solutions they came up with.

## Results

This section describes the data collected from the second round of testing and gives explanation of the results. It describes the overall results and the results and analysis of each item type.

*Overall Results of the Second Test*

With the new revisions made based on the analysis of the initial test, the second version of the test could be administered to the group. Of the students who were given the opportunity to take the second version of the test, 39 responded. All students who responded completed all items on the test and many of the students spent more time on the second version of the test than on the first, even though the second test was half as long. The results of the second round of testing are shown in Table 3.

**Table 3: Summary of Second Test Scores**

| | Overall totals | Total from 1 | Total from 2 |
|---|---|---|---|
| Group C | 116 | 62 | 54 |
| Form 1->2 | 105 | 47 | 58 |
| | 84 | 39 | 45 |
| | 79 | 42 | 37 |
| | 79 | 50 | 29 |
| | 75 | 34 | 41 |
| | 73 | 38 | 35 |
| | 71 | 37 | 34 |

| | Overall totals | Total from 1 | Total from 2 |
|---|---|---|---|
| Group D | 142 | 61 | 81 |
| Form 2->1 | 95 | 54 | 41 |
| | 92 | 38 | 54 |
| | 89 | 42 | 47 |
| | 88 | 44 | 44 |
| | 83 | 42 | 41 |
| | 82 | 39 | 43 |
| | 72 | 33 | 39 |

| | | | |
|---|---|---|---|
| | 71 | 34 | 37 |
| | 71 | 38 | 33 |
| | 70 | 33 | 37 |
| | 66 | 28 | 38 |
| | 64 | 33 | 31 |
| | 61 | 36 | 25 |
| | 59 | 32 | 27 |
| | 59 | 29 | 30 |
| | 54 | 24 | 30 |
| | 50 | 21 | 29 |
| | 41 | 22 | 19 |
| | 41 | 24 | 17 |
| mean | 69.45 | 35.15 | 34.30 |
| st dev | 17.95 | 9.74 | 9.81 |
| correlation | .69 | | |

| | | | |
|---|---|---|---|
| | 70 | 29 | 41 |
| | 69 | 27 | 42 |
| | 64 | 29 | 35 |
| | 63 | 24 | 39 |
| | 61 | 28 | 33 |
| | 60 | 35 | 25 |
| | 58 | 36 | 22 |
| | 56 | 26 | 30 |
| | 56 | 31 | 25 |
| | 53 | 27 | 26 |
| | 48 | 25 | 23 |
| mean | 73.74 | 35.26 | 38.47 |
| st dev | 21.28 | 9.76 | 13.28 |
| correlation | .70 | | |

These data show that the order effect was greatly reduced from the initial test. The increased consistency of the scores made the comparisons between the items in the new test more meaningful than in the initial test.

*Results for Problem-Finding Items*

The problem-finding items on the second version of the test used the same format as the first, but with different pictures. The pictures used in the second version of the test are shown in Figure 4 and Figure 5. The response counts are shown in Table 4 and Table 5.

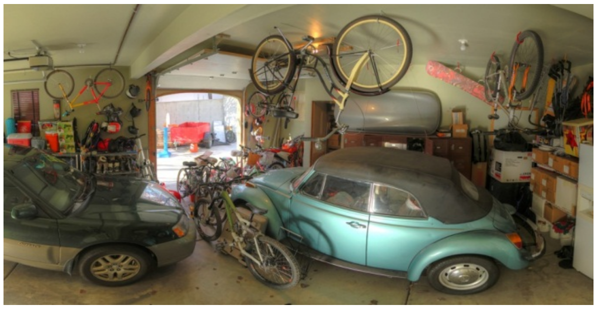

**Figure 4: Photograph from Garage Problem-Finding Item**

**Table 4: Response Counts for Garage Item**

| Response | Frequency | Score |
|---|---|---|
| Organization of bikes | 31 | 1 |
| general storage/org | 22 | |
| parking arrangements | 16 | |
| items inaccessible | 12 | |
| shelving | 10 | 2 |
| lack of space | 10 | |
| poor lighting | 10 | |
| oil stains/dirty floor | 10 | |
| hooks from ceiling | 10 | |
| too many bikes | 9 | |
| small door | 7 | |
| dirty cars | 5 | |
| entrance procedure | 5 | |
| see box contents | 5 | |

|  |  |  |
|---|---|---|
| organize items on shelf | 4 | 3 |
| store boots/shoes | 4 |  |
| snowboard on 1 hook | 4 |  |
| store unused things | 4 |  |
| car top carrier on cabs | 3 |  |
| containers on ground | 3 |  |
| water pipes | 3 |  |
| 2 garage doors | 3 |  |
| location of door | 3 |  |
| too many boxes | 3 |  |
| lockers don’t shut | 3 |  |
| basketball hoop | 2 |  |
| better kind of cooler | 2 |  |
| bikes scratch cars | 2 |  |
| trailer in driveway | 2 |  |
| use of vertical space | 2 |  |
| convertible top fix | 2 |  |
| floor seal | 2 |  |
| bike sizing | 1 | 4 |
| cabinet doors open | 1 |  |
| camera lens | 1 |  |
| car paint fades | 1 |  |
| door left open | 1 |  |
| items could fall | 1 |  |
| messy driveway | 1 |  |
| number of hobbies | 1 |  |
| organize tools | 1 |  |
| prioritize projects | 1 |  |
| promote organization | 1 |  |
| shape of driveway difficult | 1 |  |
| space for toys | 1 |  |
| take many bikes on a trip | 1 |  |
| VWs break down | 1 |  |
| yellowing parts on fridge | 1 |  |
| bike maintenance | 1 |  |

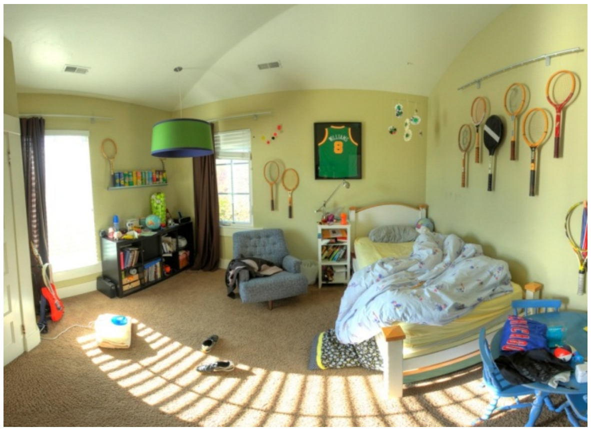

**Figure 5: Photograph for Bedroom Problem-Finding Item**

**Table 5: Response Counts for Bedroom Item**

| Response | Frequency | Score |
|---|---|---|
| bed undone | 31 |  |
| bookshelf full | 29 |  |
| clothes on chair | 22 |  |
| sun through window/blinds | 20 |  |

| | | |
|---|---|---|
| shoes on floor | 18 | 1 |
| poor lighting | 18 | |
| one leg on chair | 13 | |
| messy table | 12 | |
| no room for rackets | 11 | |
| ball storage | 9 | 2 |
| nightstand full | 8 | |
| humidifier | 8 | |
| general org/storage | 8 | |
| guitar | 7 | |
| basket for cables/games | 5 | |
| pillow on floor | 5 | |
| towel on humidifier | 5 | |
| empty floor space | 3 | 3 |
| bed storage | 3 | |
| no wall space | 2 | |
| trash can liner | 2 | |
| lack of power supply | 2 | |
| vertical space unused | 2 | |
| heated blanket | 1 | 4 |
| cd storage | 1 | |
| trash bin location | 1 | |
| paper storage | 1 | |
| workspace needed | 1 | |
| cleaning is no fun | 1 | |
| air vent location | 1 | |
| sore throat/cough | 1 | |
| room temperature | 1 | |
| need to show achievements | 1 | |
| vacuum under furniture | 1 | |
| paint fading | 1 | |
| photo lens effect | 1 | |
| cup on nightstand | 1 | |

The response counts show that the new problem finding items garnered a much larger variation in the responses. The subjects gave many more and varied responses to the items than they did in the initial test.

The mean scores and standard deviations of the problem-finding items are shown in Table 6. The table shows the overall means and standard deviations as well as the means and standard deviations of the two test groups.

**Table 6: Summary of Statistics for Problem-Finding Items**

| Overall | Garage | Bedroom |
|---|---|---|
| Mean | 13.00 | 9.69 |
| Standard Deviation | 6.14 | 5.89 |
| Group C | Garage | Bedroom |
| Mean | 12.95 | 9.20 |
| Standard Deviation | 4.98 | 4.12 |
| Group D | Garage | Bedroom |
| Mean | 13.05 | 10.21 |
| Standard Deviation | 7.15 | 7.27 |
| Item Correlation | 0.68 | |

The difference between the means of the two items suggests that they cannot be considered equivalent. There appear to be more problems to find in the garage item than in the bedroom item. In order to create more equivalent items, more pictures should be tested and analyzed.

Having more equivalent items could also improve the item correlation. The author hypothesized that the same phenomenon causing the difference in means could be negatively affecting the item correlation. Further testing with different prompts will help researchers understand whether the difference in item difficulty affects item correlation. If it is found that difficulty does affect correlation, it may mean that there are multiple factors being measured in these items.

*Results for Solution Items*

The mean scores and standard deviations of the solution items are shown in Table 7. The table shows the overall means and standard deviations and the means and standard deviations of the two test groups.

**Table 7: Summary of Statistics for Solution Items**

| Overall | Headphones | Garbage Liner |
|---|---|---|
| Mean | 8.95 | 11.15 |
| Standard Deviation | 4.85 | 6.24 |
| Group C | Headphones | Garbage Liner |
| Mean | 8.95 | 9.60 |
| Standard Deviation | 5.04 | 5.67 |
| Group D | Headphones | Garbage Liner |
| Mean | 8.95 | 12.79 |
| Standard Deviation | 4.64 | 6.39 |
| Item Correlation | 0.46 | |

The data in this table show that the headphone and garbage liner items are not likely equivalent because of the large difference in the means. There was a large difference in performance between the two groups on the garbage liner item. This may be due to the sample size of the groups. Future testing with more items and larger samples should be done to create and identify equivalent items. As with the problem-finding items, the item correlation may be improved with more equivalent items.

*Results for Communicate Items*

The communicate items use the same prompts, but the students base their pitches on the solutions they generated in the previous item. The mean scores show that this creates a difference in the difficulty of the communicate items. Some of the differences may come from the differences in the problem statements from the solution items. More testing would need to be done with different prompts in the solution items. It may be found that solution items with more equivalence could lead to communicate items with more equivalence also.

**Table 8: Summary of Statistics for Communicate Items**

| Overall | Headphone pitch | Garbage Liner pitch |
|---|---|---|
| Mean | 8.62 | 8.28 |
| Standard Deviation | 1.41 | 1.28 |
| Group C | Headphone pitch | Garbage Liner pitch |
| Mean | 9.10 | 8.20 |
| Standard Deviation | 1.37 | 1.50 |
| Group D | Headphone pitch | Garbage Liner pitch |
| Mean | 8.11 | 8.37 |
| Standard Deviation | 1.25 | 0.98 |
| Item Correlation | 0.43 | |

Inter-rater reliability for the second test was also high. The correlation between the raters' scores on the two items were 0.76 and 0.74 respectively. This is enough to confidently claim good inter-rater reliability (Cicchetti, 1994).

*Results for Ranking Items*

The ranking items gave subjects a problem statement and four potential solutions. Subjects ranked solutions using the Innovation Bootcamp's definition of innovation: original and useful ideas implemented successfully. Before the test was administered, the ranking items were given to four Industrial Design faculty. Their rankings were used to create a key to grade the students' scores by summing the point values from their rankings and then ranking the totals. Table 9 Table 11 show the problem statements and possible solutions for the ranking items and Table 10 and Table 12 show the professors' rankings.

**Table 9: Problem Statement and Experts' Rank Order for Bike Seat Item**

| **Bike seats are often exposed to the weather and become wet or absorb water, which causes discomfort to the rider.** | |
|---|---|
| 1 | A plastic cover with elastic around the edge (like a hairnet) that protects the seat from becoming wet. |
| 2 | The seat has ridges that channel water away from the rider and off the surface of the seats. |
| 3 | Small, removable seat that the rider can take with them while not riding the bike. |
| 4 | A wide fender that folds down to protect the rider from water that splashes from the tire while riding. While not riding, the fender folds up and shields/cover the seat from the weather. |

**Table 10: Expert Responses for Bike Seat Item**

| | Plastic cover | Fender | Ridges | Removable |
|---|---|---|---|---|
| Professor 1 | 1 | 3 | 4 | 2 |
| Professor 2 | 1 | 4 | 3 | 2 |
| Professor 3 | 3 | 4 | 1 | 2 |
| Professor 4 | 2 | 3 | 1 | 4 |
| Total | 7 | 14 | 9 | 10 |
| Rankings | 1 | 4 | 2 | 3 |

**Table 11: Problem Statement and Experts' Rank Order for Toilet Item**

| **People don't like to sit on public toilets. How do we make them more sanitary?** | |
|---|---|
| 1 | A toilet that automatically sprays disinfectant after every flush. |
| 2 | Seats with multi-layered tissue, one layer is removed after each use. |
| 3 | Toilet with no seat and people hold on to handrails and squat down. |
| 4 | Removable toilet seats with a seat washer in the bathroom. |

**Table 12: Expert Responses for Toilet Item**

| | Spray | Removable | Tissue | No Seat |
|---|---|---|---|---|
| Professor 1 | 2 | 4 | 1 | 3 |
| Professor 2 | 2 | 3 | 4 | 1 |
| Professor 3 | 1 | 4 | 2 | 3 |
| Professor 4 | 2 | 3 | 1 | 4 |
| Total | 7 | 14 | 8 | 11 |
| Rankings | 1 | 4 | 2 | 3 |

The summary statistics of the second test ranking problems are shown in Table 13. The data in the table show that the order effect and fatigue problems have been resolved, but that the difference in the item difficulties became more pronounced. Both groups performed better on the toilet item than on the bike seat item. More items should be created and tested to find items that are more equivalent.

**Table 13:Summary of Statistics for Ranking Items**

| Overall | Bike Seat | Toilet |
|---|---|---|
| Mean | 4.64 | 7.21 |
| Standard Deviation | 2.90 | 2.40 |
| Group C | Bike Seat | Toilet |
| Mean | 4.15 | 7.30 |
| Standard Deviation | 2.85 | 2.22 |
| Group D | Bike Seat | Toilet |
| Mean | 5.16 | 7.11 |
| Standard Deviation | 2.85 | 2.57 |
| Item Correlation | 0.09 | |

The item correlation for these items is very low. This shows that there is a serious problem with these items. This problem likely stems from the lack of agreement between expert rankings. With more consensus in the expert rankings, the item correlations will improve because there will be a stronger standard against which students can be compared. As consensus on the correct ranking improves, the items will more consistently discriminate between students who can rank the innovations well and those who cannot.

## Discussion of Results

In conjunction with the development of this test, an initial validation was performed. It is not a full and conclusive validation of the instrument, but serves as a foundation for further, in-depth validation studies. In this initial validation study, researchers looked for any major problems with the test and ensured that the test is aligned with the content domain. They also checked for reliability among the raters of the test and for equivalence between the two forms of the test.

Data from the two rounds of testing performed in this study show that the test has great potential for validity in measuring subjects' ability to innovate. Evidence gathered from this study allowed researchers to improve the test and make a case for initial validity. The ITI appears to measure the subjects' ability to perform tasks within the process of innovation. Reliability of the scores on the rater-scored items was high. These findings show that a more in-depth validation study of this instrument would be valuable.

### *Limitations of Findings*

After analyzing the data, some limitations were noted. These limitations should be addressed in future study and validation of the Innovation Test Instrument. One limitation was the sample size for the tests. Some of the response data from the items show significant differences between the groups that cannot be attributed to order effect. These differences may be the result of samples being too small. With large enough samples, the anomalies noted in the data will likely be resolved.

One other limitation was noted in the ranking items. In order to grade the ranking items, they were given to five industrial design professors. These professors ranked the innovations and their rankings were combined to create an overall ranking against which subject responses would be scored. The problem with this is that the professors were not all in agreement on their rankings. This likely caused the low correlation between students' scores on the ranking items. The validity of the ranking items could be greatly strengthened by developing responses that all the experts could agree upon rather than just combining their scores.

*Recommendations for Future Study*

Based on the findings of this research, there is potential for future studies that can further develop the ITI innovation assessment. Some of these recommendations apply to individual items from the instrument. Others apply to future validation studies that would be performed on the test as a whole.

The items on the second version of the test had varying levels of equivalence. These items should continue to be modified over time to improve pre-post-testing of the Innovation Bootcamp. The problem-finding items work better when the photographs are of rooms or scenes rather than of individual problems because they gave subjects a wider variety of possible responses.

The limitation of the ranking items that was discussed in the previous section needs to be addressed before the test can be used to evaluate the Bootcamp. Demonstrating better consensus among the expert rankings would add to the evidence of validity of these items. This could be done in one of three ways. One would be to get the experts together and have them discuss their rationale for choosing each ranking and then have them come to an agreement about how the innovations should be ranked. The second option would be to continue adjusting and testing the items until the faculty all agree on a ranking. The third option would be to get a much larger sample of experts and then total all the scores to create the rankings (as was done with the small sample in this study).

Future validation studies should be done to strengthen the claims of validity for this instrument. In this study, construct validity was only studied at a surface level. Confirmatory factor analysis would help establish that the theoretical construct that this instrument attempts to measure are valid. It determines whether or not the factors the test is intended to measure really work the way researchers hypothesize that they do. In this study, four major factors are hypothesized to measure a person's skill at innovation. A confirmatory factor analysis could tell researchers if there are other factors that these items are measuring and if their hypothesized model is right. This type of analysis was not done in this study because it requires a larger data sample than was available. Future studies with larger data sets would allow a confirmatory factor analysis to be done.

Criterion validity is another type of validity that should be studied for this test. This could be done in a number of ways. One would be to use this instrument to test students of the Innovation Bootcamp and then have raters score the performance of the same students as they participate in the course. By comparing the results, researchers could see how well the assessment predicts student performance in the Bootcamp. Studies could also be done that compare students' scores on this instrument with other validated instruments that measure parts of what this assessment does. Scores on this instrument could be compared with scores on other instruments like the ones mentioned in Lewis (2011). Another study would be a longitudinal study of students who take the assessment to see how well it predicts how innovative they are in their later careers. This could be another way of seeing how well the assessment predicts future innovation skill.

## Conclusion

This paper described the need for an innovation test to evaluate the effectiveness of innovation courses. It described the content that needed to be tested for and the procedures that the author went through to create the Innovation Test Instrument. It also showed the results of initial validation testing for the test Innovation Test Instrument.

This study is an important step in creating methods of testing students' innovation skills. Based on the testing performed in this study, the Innovation Test Instrument will help researchers understand the effectiveness of the Innovation Bootcamp at improving students' innovation skills. Future testing and development should be done to improve the item equivalency. Even with the items that are not currently equivalent, much of this instrument could be used to begin evaluating the impact of the Bootcamp. By using z-scores for the test items, researchers can compare the scores on the items to see how students have improved as a result of the Bootcamp. Once the problem of the experts' lack of consensus on ranking items is fixed, this instrument will be ready for use.

Overall, there are encouraging signs that testing students' skills at performing specific parts of the innovation process has value in measuring their overall innovation skill. This study can be used as a springboard to more research in the process-based approach to innovation measurement.

## References

Christensen, C.M. (1997). *The Innovator's Dilemma: Why Great Companies fail*, Harper Business.

Christensen, C.M., Eyring, H. J. (2011). *The Innovative University: Changing the DNA of Higher Education from the Inside Out*, John Wiley and Sons.

Cicchetti, D. V. (1994). "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." *Psychological Assessment*, 6(4), 284-290.

Drucker, P. F. (1985). *Innovation and Entrepreneurship*, Harper Collins.

Dyer, J., Gregersen, H., Christensen, C. M. (2011). *The Innovator's DNA: Mastering the Five Skills of Disruptive Innovators*. Harvard Business Review Press.

Fagerberg, J. (1999). "The Need for Innovation-Based Growth in Europe." Challenge, 42(5), 63-79.

Friedman, T. L., Mandelbaum, M. (2011). *That Used to Be Us: How America Fell Behind in the World It Invented and How We Can Come Back*, Farrar, Straus and Giroux.

Getzels, J. W. (1975). "Problem-Finding and the Inventiveness of Solutions." *The Journal of Creative Behavior*, 9(1), 12-18.

Howell, B., Wright, G., Fry, R., & Skaggs, P. (2011). "The Innovation Lab." *Proceedings of the International Conference on Engineering Design (ICED11)*.

IDEO (2011). "About IDEO." http://www.ideo.com/about/ (accessed 22 April 2011).

Innosight (2011). “Our Approach.” http://www.innosight.com/our_approach/create_or_reshape _process.html?gclid=CIPCytTUsKgCFSUZQgod0BZJHA (accessed 22 April 2011).

Kelly, T. (2005). *The Ten Faces of Innovation: IDEO's Strategies for Defeating the Devil's Advocate and Driving Creativity Throughout Your Organization*. Currency/Doubleday.

Lewis, T. (2011). “Creativity and Innovation: A Comparative Analysis of Assessment Measures for the Domains Of Technology, Engineering, and Business.” Master’s Thesis, Brigham Young University.

Mednick, S. A. (1962). “The associative basis of the creative process.” *Psychological Review*, 69, 220-232.

Obama, B. (2011). “Winning the Future.” *State of the Union Address*. Washington, D.C.

Rogers, E. M. (2003). *Diffusion of Innovations, 5th Edition*. Free Press.

Runco, M. A. (2006). *Creativity*, Academic Press.

Stanford dSchool (2010). “Bootcamp Bootleg.” http://dschool.stanford.edu/wp-content/uploads/2011/03/BootcampBootleg2010v2SLIM.pdf (accessed 14 May 2012).

Stanford dSchool (2011). “Design Thinking Boot Camp: From Insights to Innovation.” http://www.gsb.stanford.edu/exed/dtbc/ (accessed 22 April 2011).

Wagner, T. (2010). *The Global Achievement Gap: Why Even Our Best Schools Don't Teach the New Survival Skills Our Children Need--and What We Can Do About It*, Basic Books.

Wagner, T. (2012). *Creating Innovators: The Making of Young People Who Will Change the World*, Scribner.

Wright, G., West, R. (2010). “Using Design Thinking to Improve Student Innovation.” *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare,*