The Need, Development, and Validation of the Innovation Test Instrument

Jacob Wheadon, Geoff A. Wright, Richard E. West, & Paul Skaggs

Abstract

This study discusses the need, development, and validation of the Innovation Test Instrument (ITI). This article outlines how the researchers identified the content domain of the assessment and created test items. Then, it describes initial validation testing of the instrument. The findings suggest that the ITI is a good first step in creating an innovation assessment because it is more inclusive of both divergent and convergent thinking. In comparison, past innovation assessments have only assessed either divergence or convergence. The ITI still needs further validation and improvement to make strong claims about its ability to determine the effectiveness of an innovation course.

Keywords: Innovation, assessment, validity, creativity

This article is based on the Master's Degree Thesis Wheadon, J. D. (2012). *Development and initial validation of an innovation assessment* (Master's thesis, Brigham Young University). Retrieved from http://scholarsarchive.byu.edu/etd/3326/

The Need for Innovation

In industry and education, there is an increasing push for organizations and individuals to be more innovative (Fagerberg, 1999; Wagner, 2010). Rapid technological change has created the need for organizations and individuals to adapt quickly (Christensen & Eyring, 2011). Christensen (1997) describes how disruptive innovations fundamentally change markets and require new ways of thinking for organizations to adapt and survive. He describes 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. Various economies have made claims and refocused their industries to further promote and harness innovation. The European Union (EU) reported that "the central aim of the EU 2020 strategy is to put Europe's economies onto a high and sustainable growth path. To this end, Europe will have to strengthen its innovative potential and use its resources in the best possible way" (European Commission, 2011, p. 2). Similarly, the Federal Bureau of Business and Economics of India stated: "In the

ever-changing world, innovation is the only key which can sustain long-run growth of the country . . . innovation [provides] competitive advantage" (National Portal of India, 2014). In the United States, innovation had been reported as the de facto source of job creation since the 20th century (Drucker, 1985). Drucker (1985), Wagner (2012), Former President Barack Obama (The White House: President Barack Obama, 2011), and Friedman and Mandelbaum (2011), among others, have all advocated for the growth and development and the need for people and organizations to be more innovation—to be globally competitive and marketable.

The Need to Teach Innovation

Many of these calls for increased innovation have mentioned the need for schools to teach students to be more innovative (Friedman & Mandelbaum, 2011; Wagner, 2010; Wagner 2012). They have said that for American students to remain competitive in a global market and be able to adapt to a constantly shifting playing field, they need to become innovators. Schools need to teach students the skills and behaviors of great innovators (Wagner, 2010).

In a recent study, Dyer, Gregersen, and Christensen (2011) identified the common behaviors that many of today's leading innovators share. By studying innovators' behaviors, they found that people who want to be better innovators can learn and practice behaviors that will help them create innovations. Dyer et al. give educators a set of teachable skills that students can learn to perform. They claimed that although some people might have a natural propensity for innovation, anyone can learn to be more innovative.

With the knowledge that innovation can be taught, some schools, consulting firms, and corporations have begun teaching innovation. Well-known examples include the Hasso Plattner Institute of Design at Stanford University (d.school; 2017; Stanford Graduate School of Business, 2017), IDEO (IDEO, 2017; Kelly, 2005), and Innosight (Innosight, 2011), who have all reported the great value and impact of their teaching about innovation.

The College of Engineering and Technology at Brigham Young University (BYU) has a three-fold mission statement, and innovation is central to that mission. Consequently, a faculty committee was created with the goal of developing a course to teach innovation. The course curriculum uses an active learning pedagogy, teaches students about the need for innovation, and engages them in various activities during which they practice and develop divergent and convergent thinking skills and behaviors (Howell, Skaggs, & Fry, 2010). The course is currently known as the Innovation Bootcamp, and its curriculum is focused on teaching an innovation model that promotes idea finding, idea shaping, idea defining, idea refining, and idea communicating.

The Need to Assess Innovation Teaching

The Innovation Bootcamp in various forms has been taught in the College of Engineering and Technology since 2008. The course consistently receives very positive student feedback on end of term evaluations. In addition, informal assessments asking students to report on their level of interest and ability in using innovation pre- and post-course suggested that the course was having a positive impact. However, because the informal assessments were not initially designed with the intent of a longitudinal study of testing student innovative ability, the researchers believed that an assessment should be developed to ensure that course learning outcomes were being met. In addition, they believed that an innovation assessment such as this would prove to be of significance to others interested in assessing innovative ability.

Current Innovation Assessments

Tyler Lewis's (2011) thesis, *Creativity and Innovation: A Comparative Analysis of Assessment Measures for the Domains of Technology, Engineering, and Business*, analyzed various innovation and creativity assessments and measures. His findings suggested that innovation was either being measured in terms of creativity or divergent thinking (i.e., creativity tests often focused directly on divergent thinking; Houtz & Krug, 1995). Other creativity tests measure different aspects of divergent thinking, such as flexibility (Torrance, 1963), fluency (Houtz & Krug, 1995; Torrance, 1963), and originality (Houtz & Krug, 1995; Torrance, 1963), or focus on the environment for promoting innovation or focus on the end or implementation of the product (convergent thinking). For example, measures in Radosevic and Mickiewicz (2003) evaluated the success of innovation programs in terms of financial outputs, such as sales of a product or an increase in profits during or after the introduction of an innovation course or program. However, the measures that Lewis (2011) suggested would not be accurate for measuring people's innovative abilities.

The instructors of the Innovation Bootcamp implemented various measures such as the Torrance Test of Creative Thinking (TTCT) but found that these types of assessments, as Lewis (2011) had postulated, only measured the divergent thinking (creativity) part of innovation. Still needing a innovation assessment that would assess a person's innovative ability, the researchers decided to develop their own assessment to measure both divergent and convergent thinking.

Methodology

The faculty members involved with the development of the Innovation Bootcamp visited various recognized innovation institutions such as Innosight, IDEO, and Stanford's d.school, among others, and completed a very comprehensive literature review of innovation principles, methods, and processes. They ultimately identified five common themes in the innovation

research, which they used as the primary content stands for the Innovation Bootcamp. The five content strands, or "phases of innovation," are: idea finding, idea sharping, idea defining, idea refining, and idea communicating (see Figure 1).

The focus of Idea Finding is on helping students to be able to identify opportunities for innovation (some call this the problem-finding phase). The research on innovation suggests a wide variety of tools to help people identify or find innovation opportunities. The Bootcamp focused on teaching students three such tools in the areas of observing, experiencing, and inquiring.

The purpose of the second phase, Idea Shaping, is to help students organize, simplify, and clarify the results from their observations, experiences, or inquiries from the Idea Finding phase.

The third phase, Idea Defining, helps the students start to solve the problem that they identified from the previous two phases. Some researchers define this phase as brainstorming; however, it is more than simply generating a variety of options. This phase is concerned with associating and connecting ideas that may seem unrelated with the intent of forming ideas that are highly useful and novel.

The fourth phase is Idea Refining. During this phase, students are taught how to visualize, validate, and iterate the potential solutions that they generated in the previous phases. Other innovation researchers might connect or associate this phase with prototyping. However, the researchers at the Innovation Bootcamp believe that this phase is more than prototyping because it also promotes the need to decide the validity and value of the solution. This phases also stresses the idea of rapid prototyping in any format, from basic card stock and sketches to wire mockups and photo manipulations. The Idea Refining phase uses the motto of "anything that can quickly communicate your idea" to prompt students.

The final phase, Idea Communicating, teaches students how to communicate their solutions and ideas to others. This phase is taught by providing examples and rationale showing that presentations are insufficient to communicate an idea; there is a need to show, demonstrate, and describe within a context or situation. Meaning that a solution must be presented within the context of how the solution will fulfill the demand or problem.

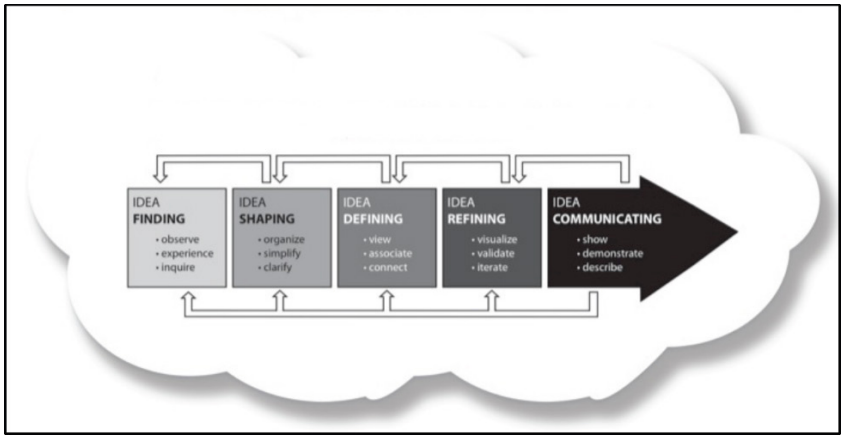

Figure 1. BYU Innovation Bootcamp model.

The five phases were used to organize the learning outcomes for the course, which guided the creation of the assessment. The learning outcomes were organized into four parts: opportunity recognition (Phases 1 and 2 of the innovation curriculum), ideation (Phase 3), idea refining (Phase 4), and communication (Phase 5). The four learning outcomes were used to create a two-way chart that was used to organize what needed to be measured in the assessment. The two-way chart, called a table of specifications (Miller, Linn, & Gronlund, 2009), is a common tool used in the development of tests, assessments, and curriculum development (Table 1) in which content strands are listed on one axis and cognitive processes are listed on the other axis. Bloom's Revised Taxonomy was the foundation for the cognitive processes in the Innovation Test Instrument (ITI; Anderson & Krathwohl, 2001). Bloom's Revised Taxonomy was used because it is a well-known and respected list of cognitive processes, and this list aligned with the course's learning outcomes. The course's learning outcomes focus on application by inviting students to apply what they are learning, so two test items were created to meet this demand. Because the course teaches students how to analyze opportunities for innovation in the various problem-spotting activities, two test questions were created to align with this cognitive process. The cognitive process of evaluation was also a key element of the course's learning outcomes; therefore, two test questions were related to this process. In these two questions, students were required to justify their decisions for the newly designed innovation. Finally, in the cognitive process of creation, the desired outcome was to assess an individual's ability to prototype an idea. A prototype is defined as a strong visual manifestation. Consequently, in the two test questions related to creation, students were required to draw and annotate the new product, system, or service that they came up with.

Table 1
Table of Specifications

	Remember	Understand	Apply	Analyze	Evaluate	Create
Opportunity recognition				2		
Ideation						2
Idea refining					2	
Communication			2			

The table of specifications (see Table 1) shows the number of items created for each learning outcome. Ultimately, there were assessment items made in the apply, analyze, evaluate, and create cognitive-process areas.

The first item type corresponded with the first learning outcome and tested students' ability to find problems using a photo-identifying activity. In this activity students were asked to identify as many areas or behaviors that were problematic. Students were graded on how many problems they were able to identify within a specified amount of time. Higher scores were awarded to those who identified more novel problems (novelty was measured using student response frequency).

In the second item type, students were given a problem statement (i.e., bike seats get wet) and were asked to write out as many solutions as they could within a specified amount of time. Higher points were again awarded for more novel but feasible answers. The TTCT uses a similar grading scheme (Torrance, 1963).

The third item type assessed the students' ability to evaluate ideas by presenting a series of possible solutions to a given problem and asking them to rank order the solutions from best to worst. Their rankings should have been based on the definition of innovation used by the Innovation Bootcamp: original and useful ideas that can be implemented successfully. The student responses were compared with the responses of four technology and engineering professors who have significant experience in innovation research and industry. To ensure interrater reliability, the responses of the professors were compared and analyzed prior to comparing them with the student responses.

The final item type assessed the students' abilities to effectively communicate their ideas to others. This item required students to write out a pitch for the innovative solution that they ranked the highest on the previous ranking question. The pitch was limited to 700 characters, which meant that it had to be concise. The grading of the pitch was based on conciseness and effective communication of the value of the solution.

The final item was graded by two raters using the provided rubric. Raters were trained on how to use the rubric and then graded five questions. They graded preselected responses that were considered by the researchers to be good, mid-grade, and poor in order to ensure that the raters could be reliable at different levels of performance. The raters discussed any areas in which they disagreed. After grading the first five responses and their subsequent discussion, the raters graded five more responses and then discussed the scores. This process continued until raters achieved agreement, which was defined as a correlation greater than 0.75 because an interrater reliability above 0.75 is considered "excellent" (Cicchetti, 1994, p. 286). After the raters graded all responses, interrater reliability was estimated for all scores.

Testing Procedures

An initial pilot version of the test was first administered during the fall semester (2012) of the Innovation Bootcamp course. It was administered to three sections of the course, which had 20 students in each section (n = 60). The pilot version was done to help with initial test form equivalence and instrument validity. Following the initial pilot implementation, the results were analyzed, and the test was revised. The revised version of the test was then administered during the winter semester of the course to five sections of the Innovation Bootcamp (n = 100). Students were told that the test was a contest and that the top scores would receive a cash prize. The extrinsic motivation of a cash prize was added based on the results from the pilot test, which suggested that we needed to ensure students were motivated to do their best on their test to ensure maximal performance.

Revisions to the ITI after the initial test. After the initial test, the results were analyzed and revisions to the ITI were made in order to improve the test. The biggest problem with the initial test was that the subjects did not achieve maximal performance. Few of the subjects finished the test, and others quickly went through the items without giving much thought to them. This likely happened for a couple of reasons. The first reason is test fatigue. Subjects' performance dropped off significantly the longer they spent on the test. This was remedied by making the test shorter. The original length of the test was longer so that there would be a larger item bank for future testing. This proved infeasible for this study because the subjects could not maintain concentration over the large number of items.

The second reason for inadequate performance was that the stakes were not sufficiently high to prompt maximal performance. In order to resolve this issue, the second round of testing was done as a competition. Cash prizes were offered to subjects with the highest test scores.

Fixing these two problems with the test strengthened evidence of construct validity. Problems with fatigue and lack of incentive hurt the construct validity of the test. Problems in the test procedure affected scores enough that they did

not accurately describe a person's ability to perform the tasks. By fixing these problems, a stronger claim of construct-related evidence can be made.

Test form equivalence. Because a major part of this study was to create equivalent forms that can be used for pre- and post-testing, two forms of the test were created and given to the students at the same time. To find the forms equivalent, corresponding items should have similar means and standard deviations for the same group of test subjects. Also, student rankings by total score should be the same for both forms of the test.

Results

Overall Results for the Initial Test

The initial (or pilot) test was given to the three sections of the Innovation Bootcamp in the fall semester. The participants were split into two groups. Half of the students from each class were put into Group A, and half were put into Group B. Table 2 lists the participant scores and the means and standard deviations for the groups.

Table 2
Summary of Overall Scores for the Initial Test

	Group A			Group B		
	Overall	Form 1[a]	Form 2	Overall	Form 1	Form 2[a]
Mean	75.83	44.92	30.92	98.17	46.33	51.83
SD	36.95	15.67	21.88	43.58	21.60	23.60
Correlation	.93			.86		

[a] Indicates which form was taken first by each group (Group A started with Form 1, and Group B started with Form 2).

These data show that scores declined as test time increased, meaning that, regardless of the test form, averaged scores were lower on the second test form. For example, Group A's mean scores decreased from 44.92 to 30.92, which was similar to Group B's decrease from 51.83 to 46.33. Although the decline was lower in Group B, because both groups experienced a decline, this was attributed to (a) test fatigue and (b) lack of incentive.

Observation showed that the subjects became fatigued because of the length of the test and the number of items. For example, many of the subjects did not attempt to complete later items on the second form. Because of this finding, the test was modified into a significantly shorter version. Originally, each form of the test was going to have two items of each type; however, only one item of

each type was included on each form of the revised version to reduce test fatigue.

Another limitation of the results is that many of the students failed to achieve maximal performance on the test items because they were not interested enough in completing the test (not enough incentive). Some subjects skipped essay questions or answered them with only a few words, which was problematic because the test was designed to score participants based on subjects' maximal performance of cognitive tasks. In the initial trial of the test, stakes were not high enough to prompt maximal performance. Consequently, incentives were offered for high performance on the revised version of the test.

Analysis of Individual Items

Analysis of the scores and responses for individual items were used to gather evidence of validity and to find ways to improve the items for future tests. Even though the initial test's issues of length and test fatigue limited what could be learned from these results, there were still important things shown. Some of the items did not perform as expected and were revised for the second round of testing. The problem-finding items did not generate a large enough variety of responses and were modified. Also, the communication items needed better instructions and were modified to help the subjects understand better what was expected of them.

Analysis of Problem-Finding Items

In the problem-finding items, subjects tried to identify problems from photographs provided in the test. A rater counted all of the responses to find out which responses were more common than others. Figures 1–4 show the pictures used in each item.

Figure 1. Photograph from the man on couch problem-finding item.

Figure 2. Photograph from the leaky drain problem-finding item.

Figure 3. Photograph from the printer problem-finding item.

Figure 4. Photograph from the street cracks problem-finding item.

The mean scores and standard deviations are shown in Table 3, which includes the overall means and standard deviations as well as the means and standard deviations for the two test groups.

Table 3
Summary of Statistics for Problem-Finding Items

	Overall		Group A		Group B	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Man on couch	7.75	3.94	9.17	4.47	6.33	2.66
Leaky drain	7.88	5.24	8.17	6.15	7.58	4.11
Printer	7.33	5.91	6.58	5.68	11.08	6.78
Street cracks	6.71	5.59	5.75	5.83	7.33	5.47

These statistics show that there was a significant order effect. The subjects tended to perform better on items that they completed earlier in the test. This makes establishing equivalence between the items difficult because it is unknown whether the change in scores was a result of those items being more difficult or a result of the order in which the subjects completed the items. Notwithstanding the order effect, some claims can be made about the difficulty of the items. Both groups scored higher on the printer item than the street cracks item. Because these items were placed in the same section of the test, this difference can likely be attributed to difficulty of the items. The other scores were inconclusive. Even though the man on couch and leaky drain items were in the same section of the test, Group A performed better on the man on couch item, and Group B performed better on the leaky drain item. The man on couch and street cracks items showed less divergence in their responses. This led to the decision to test different photographs in the second round of testing. In this initial test, problem-finding photographs were taken of specific problems similar to the ones that students identify in the Innovation Bootcamp; however, in the revised version, the problem-finding items had pictures that were taken of scenes from a home without focusing on specific problems. It was hoped that these photographs would give subjects the opportunity to identify a wider range of problems and that having to identify problems from a broader scene would be closer to the experience of problem finding that students face in the Innovation Bootcamp and that innovators face in real-world practice.

Analysis of Solution Items

The solution items gave subjects problem statements and asked them to generate as many solutions as they could. The scoring of these items followed a similar procedure to the problem-finding items. Students received points for the solutions that they generate, and more points were awarded for novel (less common) responses.

The responses show that some of the items gave the subjects greater opportunities for different answers than others. The bakery item (i.e., a local

supermarket has to discount their leftover baked goods after they are a day old) performed particularly poorly in this regard. It did not generate a very large number of different responses from the subjects. The garbage liner (i.e., garbage can liners often slip down inside of the cans when they are full of garbage) item performed best, followed by the headphone item (i.e., headphone wires get tangled in people's pockets), and then the corner-cutting item (i.e., people often cut across the lawn in places around campus, which leaves ugly dead patches in the grass). Other than the bakery item, these items garnered more responses than the problem-finding items. Table 4 shows the overall means and standard deviations as well as the means and standard deviations for the two test groups.

Table 4
Summary of Statistics for Solution Items

	Overall		Group A		Group B	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Garbage liner	7.33	5.91	5.50	2.25	9.17	7.61
Headphone	6.71	5.59	5.83	3.08	7.58	7.17
Bakery	5.71	4.25	4.50	3.75	6.92	4.37
Corner cutting	9.88	8.91	5.33	4.17	14.42	10.00

As with the problem-finding items, it is difficult to determine item equivalence based on the data shown here because of the order effect, which is attributed to test fatigue. These data show that for both groups, the bakery item was the most difficult. The other scores do not conclusively describe the equivalence of the other items.

The data from the solution items show that they performed better than the problem-finding items. In most of the items, the subjects gave a larger number of different responses than in the problem-finding items. Thus, the garbage liner and headphone items were chosen for more testing (to be used in the second round) because their means were closer than the others and because they had a large number of different responses.

Analysis of Ranking Items

The ranking items gave subjects a problem statement and four potential solutions. Participants ranked solutions using the Innovation Bootcamp's definition of innovation: original and useful ideas implemented successfully. Prior to administering the test, the ranking items were given to four engineering and technology professors. 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 5 shows the overall and group means and standard deviations for the ranking items.

Table 5
Summary of Statistics for Ranking Items

	Overall		Group A		Group B	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Bike seats	4.92	3.08	5.58	2.98	4.25	3.03
Toilets	6.71	2.78	6.42	2.87	7.00	2.65
Lawnmowers	3.92	2.83	3.67	3.27	4.17	2.27
Outlets	2.88	2.11	3.00	2.24	2.75	1.96

The data show that the outlet item is more difficult than the other items because both groups did significantly worse on it than on the other three items. The lawnmower item also appears to have scored much lower, but in Group B, the lawnmower item scored close to the bike seat. Group A and the overall scores for the lawnmower item were lower. Because of this, the bike seat and toilet items were chosen to be retested in the revised test.

Analysis of Communication Items

The communication items followed the ranking items in the assessment. The communication items asked the subjects to create a pitch for the innovation that they ranked highest on the second ranking item. They were asked to create a convincing pitch that would persuade others to adopt the innovation that they chose. Table 6 shows the overall and group statistics for the communication items from each form of the instrument.

Table 6
Summary of Statistics for Communication Items

	Overall		Group A		Group B	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Form 1 item	4.33	3.57	4.25	3.42	4.42	3.71
Form 2 item	3.63	3.84	2.08	3.28	5.17	3.74

These data show that subjects in both groups performed poorly on both of the items. Although a total of 12 points were possible on the items, the means of the

responses were less than half of that. A few problems with the items were observed when looking at individual responses.

The first problem was that many of the subjects gave very limited responses to these items. It appeared that the subjects did not care enough about the test to go through the effort of constructing a good response to this item. Also, many subjections did not finish the item. The researchers attempted to remedy this problem in the second round of testing by making the second round a competition with prizes for those with the highest scores on the test.

The second problem was that most subjects wrote the pitch as if the raters already understood the problem statement and the solutions. It was difficult for them to write about the problem and how the innovation fixed it when they were given both the problem and the solution. For this reason, in the revised version of the test, communicate questions were tied to the solution questions rather than the ranking questions. After the students generated their solutions from the given problem statement, the communication item was placed next so that students could explain the benefits of the innovation that they came up with rather than the innovation that they were given.

The third problem was that subjects did not always understand what they were supposed to write in the pitch. Some subjects described their rationale for choosing one of the responses over the others. Others failed to mention what the problem was or how their choice would solve that problem. To remedy this issue, clearer instructions were created for this item.

One aspect of these items that worked well was their rating. Using the grading rubrics, the raters scored the items with high reliability levels: 0.94 for the item from Form 1 and 0.97 from Form 2. Cicchetti (1994) said that reliability scores above 0.80 are considered "nearly perfect." This high reliability could be due to the training procedure explained in the methods section above but is also likely a result of so many of the responses being poor (raters easily agreed on responses that were severely lacking).

Overall Results for the Revised Test

The revised test was administered to 100 students in five sections of the Innovation Bootcamp. They were incentivized with cash prizes for the top 15 scores. To reduce test fatigue, the revised test also had half the number of questions that the initial pilot version did. The results show that having a shorter test with an incentive increased performance (see Table 7) and consistency—making the comparisons between items more helpful.

Table 7
Summary of Scores for the Revised Test

	Group C			Group D		
	Overall	Form 1[a]	Form 2	Overall	Form 1	Form 2[a]
Mean	69.45	35.15	34.30	73.74	35.26	38.47
SD	17.95	9.74	9.81	21.28	9.76	13.28
Correlation	.69			.70		

[a] Indicates which form was taken first by each group (Group C started with Form 1, and Group D started with Form 2).

Results for Problem-Finding Items

The problem-finding items on the revised version of the test used the same format as the initial version but with different pictures with a broader focus that the original pictures. The pictures used in the revised version of the test are shown in Figures 5 and 6.

Figure 5. Photograph from the garage problem-finding item.

Figure 6. Photograph for the bedroom problem-finding item.

The response counts revealed 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 for the initial test. The mean scores and standard deviations of the problem-finding items are shown in Table 8. The table shows the overall means and standard deviations as well as the means and standard deviations for the two test groups.

Table 8
Summary of Statistics for Problem-Finding Items

	Overall		Group C		Group D		
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Item correlation
Garage	13.00	6.14	12.95	4.98	13.05	7.15	0.68
Bedroom	9.69	5.89	9.20	4.12	10.21	7.27	

These data show that the revised version of the test had a smaller order effect than the initial version. With the reduced order effect, the equivalence of the items could be studied. The difference between the means of the two items suggests that they cannot be considered equivalent. There appeared to be more

problems to find in the garage item than in the bedroom item. In order to create two items that are more equivalent, more pictures should be tested and analyzed.

Results for Solution Items

The solution items on the revised test remained unchanged from the original test items. They appeared to be working well in the first test, but it was unclear how equivalent they were because of the order effect, so they were tested again in the revised test. The mean scores and standard deviations for the solution items are shown in Table 9.

Table 9
Summary of Statistics for Solution Items

	Overall		Group C		Group D		
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Item correlation
Headphones	8.95	4.85	8.95	5.04	8.95	4.64	0.46
Garbage liner	11.15	6.24	9.60	5.67	12.79	6.39	

The data in this table show that the order effect was also reduced for the solution items. The second round of testing gave a clearer view of the equivalence of the items. Because of the large difference in the means, the headphone and garbage liner items are likely not equivalent. These data also show that there was a large difference in performance between the two groups on the garbage liner item, which may be due to the sample size of the groups. Future testing with more items and larger sample sizes 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. It could also be that there are other confounding factors at work in these measurements. For example, if a person's past experience had led them to deal with one of these problems before, they may already have solutions in mind for these problems. Future researchers may need to look for problems to use as prompts that are either universally familiar or universally unfamiliar to the population that is being tested.

Results for Communication Items

For the revised test, the communication items were changed to go with the solution items rather than the ranking items. The instructions were also changed to be clearer and describe what the raters were looking for in the items. Table 10 contains the resulting data.

Table 10
Summary of Statistics for Communication Items

	Overall		Group C		Group D		Item correlation
	Mean	*SD*	Mean	*SD*	Mean	*SD*	
Headphone pitch	8.62	1.41	9.10	1.37	8.11	1.25	0.43
Garbage liner pitch	8.28	1.28	8.20	1.50	8.37	0.98	

These data show that even though the communication items use the same wording, they are not necessarily equivalent. The difference between the scores was more pronounced in Group C than in Group D. It is not clear why this happened, but it could be that a larger data set is needed to stabilize the results. There may be some statistical anomaly in one of the groups that would disappear with a larger test sample. 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 communication items with more equivalence also. Because the communication items rely so heavily on the solution items, the lack of correlation for the solution items is likely contributing to the lack of correlation for the communication items. In future studies, researchers should see how the item correlations for the communication items change as the item correlations for the solution items improve.

Interrater reliability for the revised 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" interrater reliability (Cicchetti, 1994).

Results for Ranking Items

The ranking items were chosen from the items in the first round of testing. The bike seat and toilet items were chosen for the revised test because they were the higher scoring items from the previous test. Table 11 shows the summary statistics.

Table 11
Summary of Statistics for Ranking Items

	Overall		Group C		Group D		Item correlation
	Mean	*SD*	Mean	*SD*	Mean	*SD*	
Bike seat	4.64	2.90	4.15	2.85	5.16	2.85	0.09
Toilet	7.21	2.40	7.30	2.22	7.11	2.57	

The data in the table show that the order effect and fatigue problems were reduced 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.

The item correlation for these items was very low, indicating that there is a serious problem with these items. The problem likely comes from the lack of agreement between expert rankings. With more consensus in the expert rankings, it is likely that the item correlations will improve because there will be a stronger standard against which students can be compared.

Conclusion

The Innovation Test Instrument (ITI) was create to address the need for an innovation test that assesses an individual's ability to perform all of the different parts of the process of innovation (Lewis, 2011). The purpose of this article was to outline the design, development, implementation, and validation of the ITI, which was designed to test an individual's innovative capacity in the skills identified from the literature: idea finding, idea shaping, idea defining, idea refining, and idea communicating. The findings from this study helped the researchers to improve the test and argue for initial validity based on the high reliability from interrater scores. Nonetheless, a more in-depth validation study of ITI would be valued. Below, the issues of validity and reliability are discussed briefly.

Validity

Although more testing should be done to further establish validity of the scores from this instrument, this study showed that there is a good case for some types of validity-related evidence: content-related evidence, consequence-related evidence, construct-related evidence, face validity evidence, and criterion-related evidence of validity.

Content-related evidence is the degree to which an instrument covers the content within a specific domain (Babbie, 1990). The evidence criterion is fulfilled by the description of the processes of innovation as outlined in this paper, and used to design the instrument (as described above). In addition, the

method of development and implementation of the ITI also helped to establish a link between the instrument and the content that is to be tested. The review of literature showed that the BYU Innovation Bootcamp curriculum is aligned with other innovation processes and models, and the methods employed shows that the ITI is aligned well with the Bootcamp curriculum.

According to Miller (2009) consequence validity describes the thoughtfulness of the consequences of use and interpretation of assessment results. In this study, the stakes of the test results were very low. Results were not used to establish grades for students or determine whether they should be admitted to certain programs or positions. The only real consequence of the results of this instrument in its current form is that the results could affect how the Innovation Bootcamp is taught in the future. The results of this instrument should not be used for other considerations without further study.

In this article, the development of the test items was described, showing that the test items were developed using generally accepted test development practices. This can be a positive initial step in establishing construct-related evidence of validity. Construct validity refers to how well the measurements taken in an assessment relate to each other according to theoretical constructs (Babbie, 1990). Showing that appropriate methods were used does not establish construct validity on its own, but it does show that construct validity is more likely than if they had not been used.

Construct-related evidence was also addressed in the revisions that were made between the two rounds of testing. Changing the pictures in the problem-finding items, moving the communication items, revising the communication items' instructions, shortening the instrument, and adding incentives were all ways that the researchers reduced construct-irrelevant variance.

Face validity is a type of validity that refers to how much the respondents perceive that the test is relevant or important (Miller et al., 2009). The first round of testing showed that the instrument had some face validity for the students of the Innovation Bootcamp. Even though test fatigue caused results that made some interpretations difficult, the fact that so many students participated as much as they did demonstrates a level of face validity. This improved more in the second round of testing because students were more invested in completing the test well. Some students commented that they enjoyed taking the test or thought that it was an interesting way to practice what they had learned in the Innovation Bootcamp. The fact that students felt that the test was relevant to what they had learned is a strong piece of evidence in favor of face validity.

Criterion-related evidence refers to how well a measured variable can predict other variables. In this test, a claim of criterion validity would say that scores on this test are a good predictor of how likely a person is to be a strong innovator. This type of validity was not formally studied in this research. Notwithstanding, the researchers of this study made anecdotal observations that

support criterion validity. The researchers of this research also assisted in the instruction of the Innovation Bootcamp. The researchers noted that the top scorers on the test were also students who had many innovative ideas during the Innovation Bootcamp. This alone is not enough to establish criterion validity, but it's an initial value to be considered.

Reliability

In this study, two types of reliability were studied: test form equivalence and interrater reliability. The results discussed in detail the equivalence of the items. Because of the differences in the means scores of the items, all of the item types in this instrument need additional work before they can be used for pre–post testing of the Innovation Bootcamp. Even though this instrument did not achieve form equivalence, it is a strong first attempt that will facilitate future instrument development in the area of innovation assessment.

Although the means and standard deviations for the items show that these items are not equivalent, they can still be used as pre- and post-test items to measure the impact of the Innovation Bootcamp. This can be done by using the data from this sample to compute *z*-scores for the responses to each item. For example, in this study, the garage item had a mean of 13.00 and a standard deviation of 6.14, and the bedroom item had a mean of 9.69 and a standard deviation of 5.89. If a student did the garage item in a pretest and scored 11, the *z*-score (in relation to the sample group from this study) would be -0.33. If the student did the bedroom item as part of a posttest, and scored 10, the *z*-score would be +0.05. In this case, the positive change in the *z*-score would show that the student performed better on the posttest item than on the pretest item.

The interrater reliability for the communication items was also tested. In the first round of testing, interrater reliability levels were 0.94 and 0.97, and in the second round, interrater reliability levels were 0.76 and 0.74. According to Cicchetti (1994), interrater reliability between .60 and .74 is considered "good." This leads the researchersto be confident in the interrater reliability of the scores for the communication items.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.) (with Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C.). (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Abr. ed.). New York, NY: Longman.

Babbie, E. (1990). *Survey research methods* (2nd ed.). Belmont, CA: Wadsworth.

Christensen, C. M. (1997). *The innovator's dilemma: When new technologies cause great firms to fail.* Boston MA: Harvard Business School Press.

Christensen, C. M., & Eyring, H. J. (2011). *The innovative university: Changing the DNA of higher education from the inside out*. San Francisco, CA: Jossey-Bass.

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. doi:10.1037/1040-3590.6.4.284

Drucker, P. F. (1985). *Innovation and entrepreneurship: Practice and principles*. New York, NY: Harper & Row.

Dyer, J., Gregersen, H., & Christensen, C. M. (2011). *The innovator's DNA: Mastering the five skills of disruptive innovators*. Boston, MA: Harvard Business School Press.

European Commission. (2011, December 12). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Open data: An engine for innovation, growth and transparent governance*. Brussels, Belgium: Author. Retrieved from http://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2011/0882/COM_COM%282011%290882_EN.pdf

Fagerberg, J. (1999). The need for innovation-based growth in Europe. *Challenge*, *42*(5), 63–79. doi:10.1080/05775132.1999.11472122

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.* New York, NY: Farrar, Straus and Giroux.

Getzels, J. W. (1975). Problem-finding and the inventiveness of solutions. *Journal of Creative Behavior*, *9*(1), 12–18. doi:10.1002/j.2162-6057.1975.tb00552.x

Hasso Plattner Institute of Design at Standford University. (2017). Bootcamp bootleg. Retrieved from https://dschool.stanford.edu/resources/the-bootcamp-bootleg?rq=Bootcamp%20Bootleg

Houtz & Krug (1995). *Assessment of creativity: Resolving a mid-life crisis*. Educational Psychology Review, 7, 269-300.

Howell, B., Skaggs, P., & Fry, R. (2010). The innovation boot camp. In W. Boks, C. McMahon,W. Ion, & B. Parkinson (Eds), *Proceedings of the 12th International Conference on Engineering and Product Design Education* (pp. 216–221). Backwell, Bristol, United Kingdom: The Design Society. Retrieved from https://www.designsociety.org/publication/30129/the_innovation_boot_camp

IDEO. (2017). About IDEO. Retrieved from http://www.ideo.com/about/

Innosight. (2011). Our approach. Retrieved from http://www.innosight.com/our_approach/create_or_reshape_process.html?gclid=CIPCytTUsKgCFSUZQgod0BZJHA

Kelly, T. (with Littman, J.). (2005). *The ten faces of innovation: IDEO's strategies for defeating the devil's advocate and driving creativity throughout your organization*. New York, NY: 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). Retrieved from http://scholarsarchive.byu.edu/etd/2865/

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). Measurement and assessment in teaching (10th ed.). Upper Saddle River, NJ: Pearson.

National Portal of India. (2014). Innovation and business: Importance or benefits. Retrieved from http://www.archive.india.gov.in/business/innovation/benefits.php

Radosevic & Mickiewicz (2003). Innovation Capabilities in Seven Candidate Countries: An Assessment, Vol. 2.8. Brussels: Enterprise Directorate General, CEC, 2003.

Rogers, E. M. (2003). Diffusion of innovations (5th ed.). New York, NY: Free Press.

Stanford Graduate School of Business. (2017). Design thinking boot camp: From insights to innovation. Retrieved from https://www.gsb.stanford.edu/exec-ed/programs/design-thinking-bootcamp

Torrance, E. P. (1963). *Creativity*. Washington, DC: American Educational Research Association & National Education Association.

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*. New York, NY: Basic Books.

Wagner, T. (2012). *Creating innovators: The making of young people who will change the world* (Rev. ed.). New York, NY: Scribner.

The White House: President Barack Obama. (2011, January 25). Remarks by the President in State of Union Address. Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2011/01/25/remarks-president-state-union-address

About the Authors

Jacob Wheadon (jacob.wheadon@gmail.com) is a graduate student of Technology and Engineering Studies at Brigham Young University.
Geoff A. Wright (geoffwright@byu.edu) is Associate Professor of Technology and Engineering Studies at Brigham Young University.
Richard E. West (rickwest@byu.edu) is Associate Professor of Instructional Psychology and Technology at Brigham Young University.
Paul Skaggs (paul_skaggs@byu.edu) is Associate Professor of Industrial Design at Brigham Young University.