The Architectural Creativity Test Development: A Many Facet Rasch Model Analysis to Establish Inter-Rater Reliability

Niken Titi Pratitis
Program Doctoral Psychology, Universitas Airlangga
nikenpratitis@untag-sby.ac.id

Urip Purwono
Faculty of Psychology, Universitas Padjajaran
Urip.purwono@gmail.com

Abstract

This main purpose of this study is to validate the structure of creativity test's internal consistency in the field of architecture. Method of analysis's done by Many Facet Rasch Model (MFRM) approach using Facet Program. There were 44 High School students of Public, Private, and Vocational School, also 2nd and 8th semester Architecture College students involved as participants. Three person become raters, and assessed the participants, which consist of two academicians (Architecture lecturers) and one professional Architect. Analysis of Facet Program's indicates the comparison between *exact agreements* value and *expected agreements* value is very small. So that it produce a very high reliability coefficient (> 0.8). The result shows that *rater's* interpretation is very good that it can provide stable and consistent evaluation. It also indicates the agreement for evaluation's score given. Rater's agreement also strengthen the evidence that constructed items are relevant with measuring attributes and represent overall measurable domains.

Keyword : Inter Rater Reliability, Creativity, Architecture

Received 21 August 2018/Accepted 27 November 2018

Introduction

The need to measure creativity is a critical issue in the Architecture Higher Education, particularly to map the creative skills of prospective architects. This statement is supported by Williams et al. (2010) who explained that the need to build this measurement tool is becoming increasingly important. Moreover, Ostwald & Williams (2008a; 2008b) identified three main problems related to working on creativity and design education, namely lack of understanding on pedagogical dimensions of creativity, lack of appropriate strategies to understand and assess creativity level, as well as lack of proper models to support the assessment of creativity in design.

Amireh (2013) admitted that measuring creativity in architecture is a challenge because the concept of architecture derives from both pure science and art. The main obstacle is the designing process, often characterized by difficulties in finding similar references and absence of a measured assessment technique to accommodate creativity, as well as the consequence of the process itself (Christiaans & Vanselaar, 2005; Demirkan & Afacan, 2011). Dorst and Cross (2001) reported that identifying creative ideas in a design process is difficult, even though it underlies every design project. The definition of creativity in architecture and design education also remains to be debatable (Ostwald et al., 2011).

Creativity has been defined in various ways. Some researchers considered creativity to be a part of the cognitive process, involving knowledge in generating ideas (Suharnan, 2011; Sternberg, 1999; Weisberg, 1993). Others defined creativity as the ability to create novel and useful products (Stenberg, 1999; Evans, 1994; Baron in Amanah, 2007; Munandar, 1999) or modify something (Semiawan, 2010). A more radical definition is that it is a potential that involves elements of value or appropriate thoughts for a given situation (Mohr in Weisberg, 1993), producing several creativity test kits which do not necessarily include architectural aspects.

People generally measure creativity based on four basic aspects, namely fluency in expressing ideas, flexibility, originality and elaboration (Guilford, 1974; Munandar, 2011; Kaplan & Saccuzzo, 2012). However, novel ideas that underlie the creative design in Architecture could also occur by rearranging existing knowledge based on new object association process (Mednick, 1962) or combining and finding new relationships between known facts (Kaplan & Saccuzzo, 2012).

Creativity in Architecture

Several studies indicated that creativity in Architecture and engineering design have a direct relationship with imagination (Amarta, 2013; Buzan, 2004; Eguiluz, Cavia, & Lavendero, 2003; Laurens & Tanuwidjaya, 2003; Drabkin, 1996; Ibrahim. 2012). This indicates that the existing definition of creativity, as well as the aspects measured in the creativity test kit, could not measure the design aspect in Architecture. A study on Architect students using an assessment of design products (artifacts) was considered to be unsuccessful in answering the problem

(Demirkan & Afacan, 2012). Similar studies using artifacts have also been done (Besemer & Treffinger, 1981; Besemer, 1998); Christiaans, 2002; Horn & Salvendy, 2006, 2009; O'Quin & Besemer, 1999, 2006; Demirkan & Hasirci, 2003, 2007, 2009).

The underlying issue is that general creativity test does not accommodate the technical aspects (e.g., principles, architectural design elements) and art. Thus, the indicators are considered "incomplete" in measuring the creativity level in the field of architecture. Both the figure test and work appraisal test tend to be subjective, creating unclear boundaries and imprecise aspects in describing principles and elements of Architectural design. As a result, the assessment varies from one expert to the next. Some characteristics of creativity in the field of architecture is different from general creativity, such as aspects of originality (different, unconventional, rare, extraordinary, interesting, eccentric, new, novel, unusual, unique, original), integration (coherent), equilibrium (adequate, reasonable), form of produced design (size, proportion, number, and geometric relationships) and the involvement of assembly design elements (harmony, rhythm, repetition, balance).

From those points of view, we could construct a tool to measure architectural creativity level. The constructed creativity test should be able to measure important aspects of architectural creativity, involving elements of design and design principles, in addition to general creativity aspect. Related to this, an important stage of psychological measurement tools's construction is to gather numbers of evidence that demonstrate the accuracy, reliability and credibility of constructed measuring instrument. The evidence becomes important because it give description for the capability to measure psychological attributes necessity to measure and the capability to provide the precise score with small error measurement. Evidence collection will lead to validity and reliability of measurement tool's information (Azwar, 2012). Basic assumption is that measurement instrument consider reliable if it still produce the same information while used for several times. In other word the instrument will not show the significant variation of information (Sumintono dan Widhiarso, 2014). Score stability of the instrument is need to be supported by the evidence that all of the aspects, indicators and items of measuring instrument have formed the accurate construct of the measured attributes (Azwar, 2012), which termed as validity in psychometric field.

Validity and Interrater reliability

Validity, as written in *Standard for Educational and Psychological Testing* – AERA, APA dan NCME (1999) defined as the degree of evidence numbers and theoretical framework that support interpretation of test's score needed to adjust the rule of instrument practice. As basic things for development and evaluation, validation process including the accumulation of evidence that give scientific base for score interpretation should be taken. Five evidence categories when checking of interpretation validity correspond to the purpose of construction that need to be gathered as written in *Standard for Educational and Psychological Testing* – AERA, APA dan NCME (1999), including content, process related to subject's respond, correlation with other variables, and testing consequences. The more evidence gathered by a researcher, the higher validity and the consequences is the higher and better reliability coefficient.

One of the evidences to be the purpose of a research is evidence related to internal structure of the test. It is proven by interrater agreement testing. The testing of interrater agreement is the way to estimate instrument's reliability. Sumintono and Widhiarso (2014) suggested another way to estimate reliability by using the similarity testing based inter time to measure test's score stability, or by using pararel testing instrument to assess test's equivalent, or by using internal consistency test to assess elements within the body of measuring tools.

In inter-rater agreement test, reliability is estimated based on coefficient resulted from measuring the same subject by using the same measurement tools, but assessed by two or more assessor. The assumption is, if the score resulted from several assessors to the same subject tend to be consistent or equal, then the reliability is considered high (Sumintono dan Widhiarso, 2014). Reliability estimation technique by using rater is not popular among researchers due to several considerations, such as the difficulty to find the right assessor and time obstacles; also specified statistical analysis technique that have to be applied on data processing was hardly mastered.

The main reason to use interrater agreement as a method to prove the evidence of test internal structure is the agreement of some independence rater when assess the test score. It will prove that another person beside researcher can assess test's score objectively. However, there is a risk that the rater interpret the score test subjectively. This means, that the rater

gave the same score to same subject due to subjectivity involvement. Therefore the subjectivity chamber should be limited. This is the role of interrater agreement, to test the awareness against the rubric in order to limit subjectivity and lead the rater to agreement for the score. It will be dangerous if different raters give far different score to the same subject. If it happened, the rubric should be improved due to different interpretation and high subjectivity of the raters.

Inter rater agreements aimed to estimate reliability test in this research process with Facet statistical program based on Rasch Model. Genereally, Facet is different from Kappa Coefficient developed by Cohen (1960) and consider more notable. First, Facet can use more than two raters while Kappa is more limited. Second, Facet can use more than two category score test while Kappa generally only can be applied to two categories (code 0 and code 1) Third, Facet program processed data based on Rasch Model. This model required Logit scale (log odds unit), the scale with same interval and having linear quality, from odss ratio, not from raw score. As a result, the process of persons' estimating capability or item difficulties level will show more exact estimated score, and also the score can be compared to each other because they have the same elements. Besides that, through the use of logit scale, the resulted score will be related on occurred respons pattern, not on determined initiating score, Therefore, Rasch Model is considered independent measurement (Sumintono, 2014)

There are several advantages using Many Facet Rasch Model (MFRM) analysis in order to reveal inter-rater agreement which difficult to be done using classical test theory. The Facet software can provide the percentage of inter-rater agreement and other information. Firstly, it provides three facets output simultanously, which are reliability coefficient of rater, retee and items. Secondly, it provides information about mean, standard deviation, strata and separation value of discriminant which aid in determining the classification of raters' and subject's ability, as well as item's difficulty level. Thirdly, more detailed information about the assessment quality of each rater can be reported through rater and ratee unexpected response. Lastly, the outcome of the rating scale diagnostic information which describes the raters' apprehension toward the variation of rating scores in the assessment rubric inform us on whether assessment guideline needs to be simplified.

Based on the aforementioned advantages, this research focuses on gathering evidence of validity and reliability using Inter-Rater Reliability. The inter-rater agreement used in this study relies on two reasons stated by Widhiarso (2017) and Ebel & Frisbie (1991), namely (i) it increases the certainty that the items are relevant to the measured attributes and represent the entire domain of measurement, and ii) it is more objective. We also included non-researchers, namely experts, to score the test result using a particular assessment rubric that the researcher has made. This is done to prove that the assessment rubric can be easily understood by both researchers and experts.

In general, our constructed creativity measurement tools include both the aspects from general and architectural creativity (e.g., elements and design principles). To the best of our knowledge, this research has never been done before. Several studies on creativity in the field of Architecture often assessed their variable using a general cultural creativity tests such Torrance Test of Creative Thinking (Kvashny, 1982; Potur & Barkul, 2006; Portul & Barkul, 2009; and Cho, 2012) or task assessment studio or artifact (Demirkan & Afacan, 2012; Demircan & Hasirci, 2009; Hasirci & Demirkan, 2003; Hasirci & Demirkan, 2007). While the test of creativity in the field of Architecture was once constructed by Appulembang & Suyasa (2014), it tends to measure similar aspects to general creativity and does not include aspects that are specifically found in Architectural creativity. We expect that a specific architectural creativity test will answer the needs of the Indonesian Higher Education in Architecture, particularly to accurately predict the creative potential of future architects. Through these predictions, Indonesian Higher Education in Architecture can develop a more harmonized curriculum in sharpening and stimulating the creative potential of the students.

Method

As subjects of research, it is involved 44 students from Second Grades High School, Third Grades Vocational School, freshmen and final-year Architectural College. All subjects have passionate in design mainly in architectural fields. They come from 4 public schools, 2 private schools located in Surabaya. Vocational school specialized in construction drawing technique. The choice of specialization based on assumption that the students will continue their education to architectural school. College students in this research were including freshmen

on the 2nd semester and final-year from several privates and publics universities located in Surabaya. all respondents in this study participated in the study after signing the informed consent, so there was no compulsion for their involvement.

Data gathered from the subjects were formed as responses for the specified picture that being scored by 3 raters. The raters are 2 architecture academicians in privates and publics universities in Surabaya. The third rater is a professional in architectural field. The three raters did not know one another. They did not make any form of communication. They scored test by using rubric reference that has been tabulated based on consultation with another Architecture academicians outside the three raters. After that, the data have been analyzed by Facet program based on Rasch Model to test the assessment agreement of three raters mentioned above. The agreement obtained was one of the validity evidence's forms and would determine reliability coefficient of measurement tools constructed.

The Architectural Creativity defined as cognitive capability of the architect in creating innovative, esthetic and original design that is assessable and accountable, though systematic design process including imagination, association and transformation of idea, by the way of managing design elements consisting of dots, lines and geometrical shapes, by means of balance, repetition, proportional unity and vocal point principles and design values including textures, colors, etc. This operational definition was built from the concept of creativity theory based on cognitive approach. This approach grounded by divergent thinking theory of Guilford (1967) and Mednick theory of association process (1962), and supported by some studies about the importance of innovation, imagination and originality in creativity (Buzan, 2004; Eguiluz, Cavia, Lavendero, 2003; Laurens, 2003; Drabkin, 1996; Antoniedes, 1990; Joseph, 2009;, Vernon, 1970; and Lumsdaine, Shelnutt, & Lumsdaine, 1999). And also the study of Demirkan & Hasirci (2009), Hasirci and Demirkan (2003, 2007) about alignment, originality, balance and assembling elements in creative design creation *Pre eliminary study* was conducted as the attempt to strengthened operational definition in this research, using interview and Focus Group Discussion with *expert judgement* (Profesional Architect and Architectural Academician). The preliminary study has strengthened the notion that Architectural creativity has a specific and detailed technical element that we need to consider. In other words, architectural creativity must be measured through a different instrument. Furthermore, item

analysis of the architectural creativity test on a field study has also been carried out twice. Both instances generated positive results, indicating that the architectural creativity items have fulfilled the requirements to be considered as good items for a measurement tool.

Architectural creativity consists of Innovative, Aesthetic, and Original aspects, Accountable aspects, and Systematic aspects in processing design elements using principles and design principles. These aspects can be measured through nine indicators, namely originality, aesthetic harmony, aesthetic diversity, aesthetic imagination, aesthetic integration, fluency in generating ideas, transformation, balance, and rationality. In this research, these indicators are then manifested in test items that are figural shapes. it demanded figurative responses of the testee. Test items consist of five commands.

The first command asked the respondents to draw a new object by combining at least two geometric shapes from the six geometric shapes provided as a stimulus (Fig. 1). The drawing must be based on two different themes (the first theme is about the house and its environment, the second theme is about education).

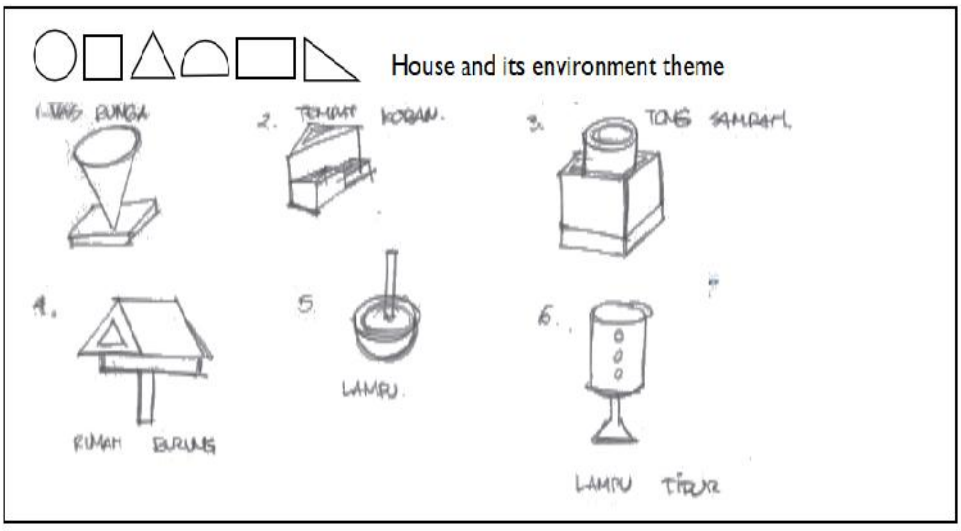


Figure 1a. Sample Answers (Item 1)

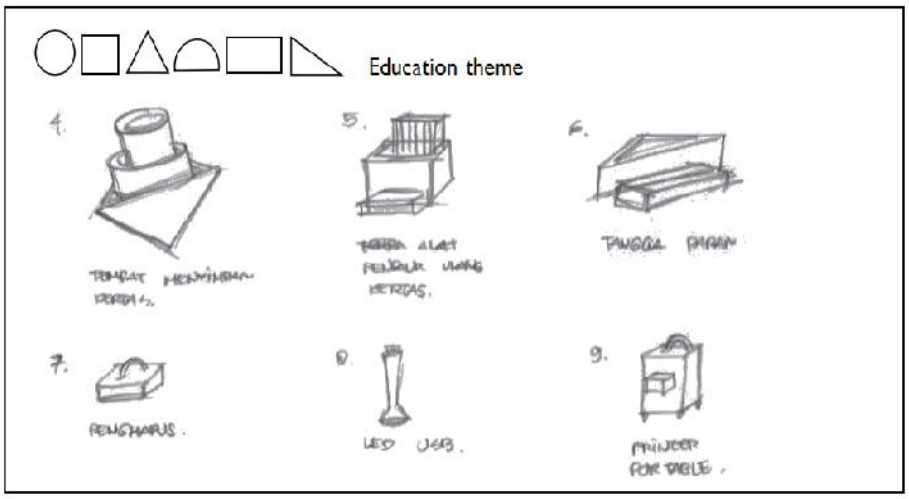


Figure 1b. Sample Answers (Item 1)

The second instruction asked the respondent to draw on six empty rectangles about new and unique patterns that are not on others mind, as displayed on Fig. 2.

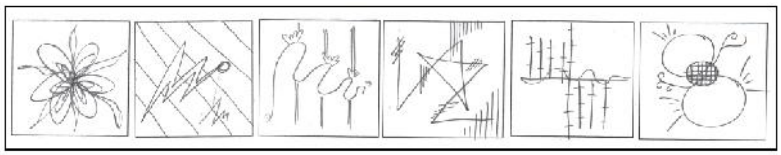


Figure 2. Sample Answers (Item 2)

The third instruction asked respondents to describe various possibilities if a paper consisting of three different forms (quadrilaterals, circles and triangles) is cut into pieces (Fig. 3).

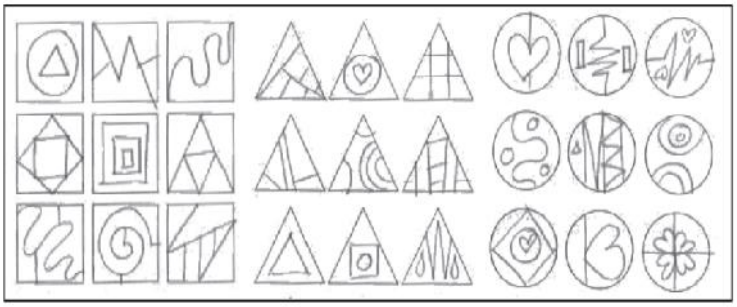


Figure 3. Sample Answers (Item 3)

The fourth test instruction, asked respondents to draw an illustration of a room containing certain objects in their respective positions using their imagination (Fig. 4).

Figure 4. Sample Answer (Item 4)

The final instruction of the test, asked the respondents to draw a picture of a unique and different table decoration design using 30 pieces of wood (10x10 cm of area) with a thickness of 1 cm (Fig. 5)

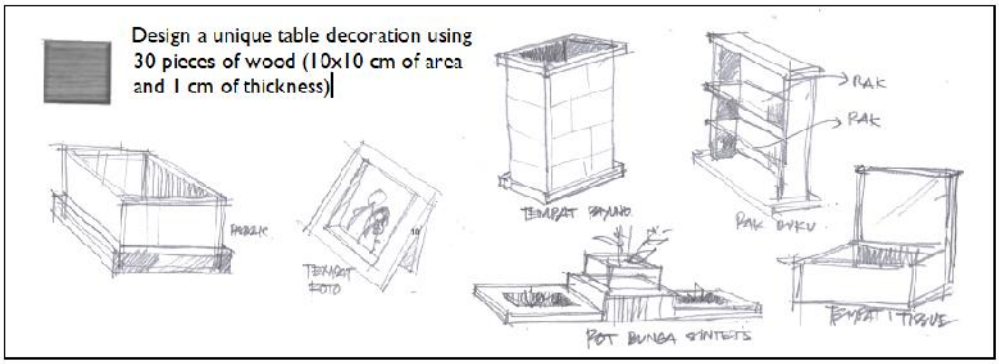


Figure 5. Sample Answer (Item 5)

Hypotheses

- Creativity Test in Architectural Field has high interraters reliability coefficient
- Creativity Test in Architectural Field has stability in assessment score

-The Items of Creativity Test in Architectural Field were proved to have a high meticulousness to measure creativity indicators in architectural field.

Result

The main result to be reported is achieved from output of *Rater Measurement Report* in *Facet* program based on *Rasch Model*. This related to *rater agreements* obtained from comparison between *exact agreements score* (30.8%) and *expected agreements score* (31.0%). Based on inter rater analysis, it displayed the information that Creativity Test in Architectural Field developed in this research relatively capable to produce the score agreement of the independent raters.

Besides the evidence of inter rater agreement, The Facet Program obtained 0.73 item reliability coefficient and 0.87 rater reliability and 0.99 rate reliability. Based on Fisher (2007) criteria, the high value of the magnitude of the coefficient of reliability's reliability, declared as categories good enough reliability to excellent. The results of the reliability coefficient of test inter rater agreement sourced from reliability, rater and rate item the encapsulated and presented in table 1.

Table 1.
Summary of Reliability Coefficient in Inter Rater Agreements Testing

Measurement	Alpha Cronbach	Limitation	Explanation
Rater	0,99	> 0,94	Excellent
Ratee	0,73	0,67 s/d 0,8	Average
Item	0,87	0,81 s/d 0,90	Good

Source : Output of Rater Measurement Report table, Ratee Measurement Report table and Item Measurement Report Program Facet Table, Rasch Model Analysis

Data analysis using the Facet also presents information that is primarily based on the output of the Measurement Report table, which produce a mean logit of three rater is -2.73; SD = 1.30; strata = 18.80; separation = 13.85 and SE Models are moving from 0.04 until 0.15 ; While the mean logit Ratee (research subjects) = 0.14; SD = 1.08; strata = 2.52; separation = 1.64 and SE Models are moving from 0.14 until 1.83; and mean logit tests item = 0.00; SD = 0.52; strata = 3.71; separation = 2.54 SE Model with moving from 0.14 until 0.30. The summary output table the Measurement Report contained in table 2.

Table 2
Summary of Rater, Ratee and Item Measuremet Report

Measurement	Mean Logit	SD	Strata	Separation	Model SE
Rater	-2,73	1,30	18,80	13,85	0,04 s/d 0,15
Ratee	0,14	1,08	2,52	1,64	0,14 s/d 1,83
Item	0,00	0,52	3,71	2,54	0,14 s/d 0,30

Source : Output Tabel Measurement Report

Other information that results from the test program on Facet inter rater agreements was rater and ratee unexpected response, which describes the consistency or quality of the rater and ratee research. Based on rater unexpected response obtained results that 2 of the third rater research, that is rater B and C most often give assessment under ideal value that should accrue to the subjects of research (28^{th} times), although rater C never deliver value higher (6^{th} times) of the test results of the research subjects. While rater A detected never give a lower assessment 13^{th} times and 3^{rd} times higher than the ideal value should be obtained subjects of research. On the other hand, having regard to the ratee unexpected response, obtained information from research subjects, subjects number 10 was the subject most often rated rater "doesn't fit" (higher or lower than the ideal values which should be achieved). Results more loading on table 3 and table 4.

Table 3
Summary of Unexpected Responses

Rater	Scoring Under The Ideal Score	Scoring Over The Ideal Score
A	13 x	3 x
B	28 x	-
C	28 x	6 x

Source : Output Tabel Unexpected Responses, Program

Table 4
Summary of Subject with Creativity Ability Difficult to Be Scored

Quantity of Subject's Appearance	Subject Number	Subject Amount
8x	10	1
7x	-	0
6x	-	0
5x	11 dan 36	2
4x	12, 13, 35, 37 dan 41	4
3x	26, 38, dan 44	3
2x	21, 31, 32, 39, 40, 42, dan 43	7
1x	1, 2, 3, 4, 5, 6, 7, 8, 9, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 27, 28, 29, 30, 33, 34, dan 44	27

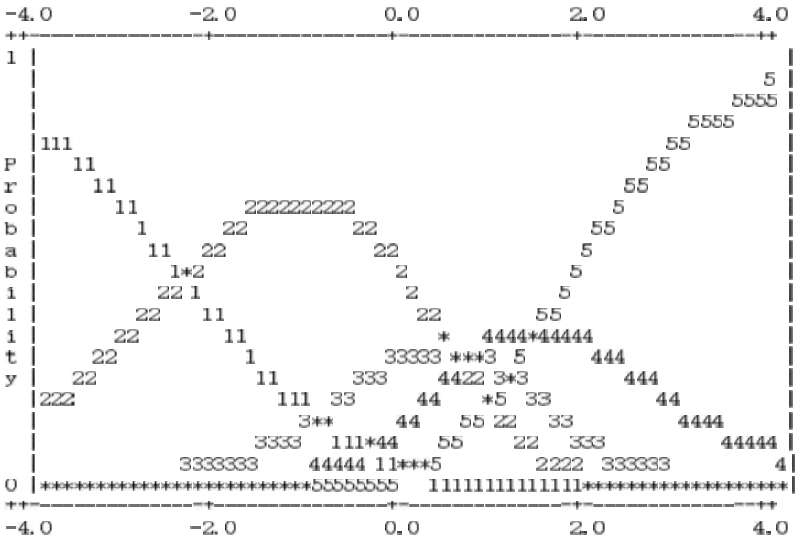
Source : Tabel Unexpected Responses, Proram Facet

Other results can be reported from test inter rater agreement using the facet is the rating scale diagnostic indicated from increased average measure and index and rich threshold. In this study, the average value obtained proof that the measure is moving from -0.26 until 2.59 and index and rich threshold on rating 3 and 4 under 1.40 logit. Specifically the results described through graph probability curves (graph 1) and summary table index and rich threshold (table 5).

Table 5
Summary of Index Andrich Threshold

Rating	Index Andrich Threshold	Inter Rating Difference
1	-	-
2	-2,51	2,51
3	0,40	2,91
4	0,70	0,30
5	1,41	0,71

Source: Output Facet Tabel Rating Scale Diagnostic



Graphic 1. Probability Curves
Source : Output Probability Curve, Program Facet

Discussion

The score agreement among all three raters, as presented in the result section, shows good consistency and stability in assessing and understanding the assessment rubric used for scoring the architectural creativity of 44 research subjects. In other words, the assessment scores were relatively stable among raters because each of them had the same understanding of the rubric. This proves the objectivity of the raters in evaluating the test results. It also highlights an important finding that raters who come from different backgrounds and independent work field could still assess these test results provided the same guideline (rubric) be used. Indirectly, the high-reliability coefficient between raters also proves the stability of the test score.

The important things that need to be reported, with attention to the score mean logit achieved (good rater, ratee nor the item), then it can be informed there are three conditions that can be discussed. First, by observing the score mean logit rater, then a third rater in the study even though it is proven to have a good agreement (agreement) in assessing but all three include lenient (tend to give the score a high test results in assessing the results of tests

on each of the subjects of the research). Second, the ratee (subject) from the score mean logit ratee, shows that the average ability of creativity in the field of Architecture is good. Third, based on the mean logit item, then in the field of creativity tests item Architecture has average difficulty level.

The tendency of the rater to lenient, presumably influenced by the rater subjectivity spaces still, so these three raters tend to give positive assessment against the subjects assessed. The actual decision making errors (such as giving a value higher than the value of the actual achieved ideal subject), including reasonable happens in an assessment, especially when concerning the assessment of the other person. Most likely, a tendency he gives positive value by the assessment against the subjects of the study, is because the feeling of same interest and educational background that is the field of architecture. In addition tendency hallo effect may also be experienced, i.e. While the evaluator gave a high score on one aspect or indicator, there is a tendency they also gave high scores on indicators of assume that if on one indicator the subject is able to respond properly then it's likely the subject will also respond well on other indicators.

Interestingly, the analysis with facet, is also able to show the sequence of raters from the most lenient to a less lenient. Rater B more lenient than rater A and C and rater C is the least lenient compared to two other raters. Even with paying attention to unexpected response, it can be noted that evaluator B and C are likely to provide under a 28^{th} times ideal value which should be accepted by the research subjects, while the raters with only 13^{th} times value lower than the value of the ideal subjects. This is quite interesting, because raters A and B has academics backgrounds (as lecturers) and rater C is the practitioner of architecture (as professional Architects). Presumably the background as lectures which are encouraging raters A and B tend to more easily assess the positive test results because as educators they can appreciate the process of the whole subject is entirely "students" although from a variety levels of education. While rater C which is a practitioner, more likely to see the final result in accordance with their performance, so far where creative design works in the field of architecture that directly perceived is the result. Subjectivity that appear on the raters that presumably that making a distinction between the value still exact agreements and expected agreements, although in general the difference is too small. That is,

although such differences exist and shows that there is an element of subjectivity in the assessment of the performance of the research subject, however small the percentage difference is pointed out that such subjectivity spaces can be restricted by either. It is also supported by the value of the strata and separated the rater indicates that the score given by the rater have a high reliability (value and separated strata > 5 includes having excellent reliability; in Sumintono and Widhiarso, 2015; Fisher, 2007), in addition it generates value Model logit SE Rater that overall under 0.5 so that illustrates the level of carefulness of rater's third research.

The high precision of the third rater in giving judgment against the subjects of research, certainly gives an overview that regardless of the elements of subjectivity, the third independent raters are very responsible, careful and conscientious in giving judgment. This means that the researcher is quite right in choosing the third raters, although all three of them have a different background jobs and do the assessment separately. Raters in this research provides assessment rubrics based on the judgments given researchers, reflected also from a long discussion conducted researchers with each of them. Before the assessment process progresses, reviewers asked in detail about the meaning of the description of each rating in the rubric. In fact they also provide input and advice are clear descriptions of each rating to measure each indicator based on their understanding as people who know architecture properly. The rater also discuss completely about the description of each of the indicators and the rating so that researchers can refine and produce the rubric assessment be disallowed with great detail.

The high precision of the third rater in giving judgment against the subjects of research, certainly gives an overview that regardless of the elements of subjectivity, the third independent raters are very responsible, careful and conscientious in giving judgment. This means that the researcher is quite right in choosing the third raters, although all of them have a different background jobs and do the assessment separately. Seriousness of this research in the rater provides assessment rubrics based on the judgments given researchers, reflected also from a long discussion conducted researchers with each of them. Before the assessment process progresses, reviewers asked in detail about the meaning of the description of each rating in the rubric. In fact, they also provide input and advice are clear descriptions of each

rating to measure each indicator based on their understanding as people who know architecture properly. The rater also discuss completely about the description of each of the indicators and the rating so that researchers can refine and produce the rubric assessment be disallowed with great detail.

Information about the quality of the subject's research is supported with the resultant value of the strata and separation/ratee which generate information that is generally based on the score the test results, the subject can be grouped in 3 categories the level of ability creativity, i.e. subjects with high creativity category (10 subject), creativity is the average (27 subject) and low creativity (7 subject). More information about the categories of ability of creativity research subjects can be seen in table 5.Indirectly (although the categorization of also having regard to the price of a Model SE ratee research shows most of the subjects have the creativity that average), the differences of the research subjects into 3 categories based on these test results score important evidence that tests the creativity in the field of Architecture have a different score. The power of such a good score difference illustrates that test score of creativity in the field of Architecture capable of differentiating both groups the subjects of different ability, which is the indication of a measuring instrument that has a good level of reliability (Mardapi, 2012; Azwar, 2012).

Furthermore, by checking out the mean logit items, it concludes that the average items of creativity have a medium level of difficulty. Specifically based on value strata and separating items, generated, it can be concluded that the average item tests the creativity in Architecture field has a difficulty level. Specifically, based on the value of the strata and separate in item test, items are also reported to be clumped into 3 levels of difficulty, i.e. the item test that rated as high difficulty level are 6 items; and the group of item with an average level of difficulty are 23 items; and the groups of item with low levels of difficulty are 4 items (table 6). Although the deference of the three categories of difficulty item level, but pay attention to the price of a Model SE items smaller than 0.5, indicates that all items of the creativity test in the field of architecture constructed in this research have a good precision in measuring the charge indicators will respectively.

In general, the findings strongly support the results of the item analysis that we conducted as part of the first and second field study. The two field studies found that the range of reliability coefficients is between 0.85 - 0.985, illustrating that the architectural creativity test items are well-distributed based on the item difficulty and have an excellent discriminant ability. Thus, this study proves that the constructed creative test has great item quality and assessment. Additionally, it is also considered a reliable tool to measure, not only the fundamental aspects of creativity but also the design aspects in evaluating architectural creativity.

Conclusion

Result of data analysis by using interrater agreement test shows high interrater reliability coefficient in Creativity Test in Architectural Field. The test has proved to have a stabilized score. The items have proved to have high meticulousness in measuring architectural creativity indicators. For that reason, all of the hypotheses were accepted.

Our findings indicate that the constructed test is a reliable tool for measuring architectural creativity. However, several limitations remain to exist. The first limitation is related to the figurativetest response. It requires raters to have a full understanding of the assessment rubric. Ensuring that each rater to have the same perception on scoring each item, despite being architects, remains to be a challenge. Therefore, the assessment rubric should be made in detail so it can be easily used as a reference. Secondly, there are more variables outside of inter-rater reliability that could influence the decision for people to believe that this is a reliable test for measuring architectural creativity. As a result, other tests are needed to strengthen people's trust that the architectural creativity test items are valid and can produce realistic and stable scores even when different assessors conduct the assessments. Lastly, the scores on the test rubric by determined range tend to yield difficulties for the raters. The raters considered the rating of 3-4 and 4-5 to be unclear (collapse rating).

Consequently, there are some suggestions disclosed by the author. The rubric should be constructed and examined well, ensuring accurate description of each score and able limiting the subjectivity of the raters. Next, the validity and reliability evidence should be obtained to strengthen people's trust in using the architectural creativity test. Reliability estimation (using

test-retest reliability) or testing reliability based on equivalence (based on the similarity between two instruments, such as the number of items, the difficulty level, and the administration) can be used to obtain more evidences.In term of a collapse rating, a rating of 1-10 scale is recommended to be simplified into that of a 4 or 5 scale.

Finally, we hope that the architectural creativity test will be used appropriately now that it has been proven to be reliable and valid. The use of the architectural creativity test should be used to map the capability of future architects as well as be part of the student selection process for entering Higher Education in Architecture.

References

Amanah, D. (2007). Pentingnya pengembangan kreativitas. (The importance of creativity development). *Jurnal Madani, 8*(2), 45-62.

Amarta, R. (2013). *Agar kamu menjadi pribadi kreatif. (How to b*e a creative person). Jakarta: Sinar Kejora. .

Amireh, O. M. (2013). An introduction to creative thinking in architectural design. *International Journal of Engineering & Technology IJET-IJENS, 13*(5), 44-53.

Antoniades, A., C. (1990). *Poetics of architecture: Theory of design*. New York: Van Nostrand Reinhold. ISBN: 9780471285304-0471285307.

Appulembang,Y.A., &Suyasa, P.T.Y.S. (2014). Pengembangan alat ukur kreativitas pada mahasiswa jurusan Teknik Arsitektur. (Development of creativity tool for Architectural Students). *Jurnal Provitae*, *6*(1), 1-18.

AERA-APA and NCME. (1999). *Standard for educational and psychological testing*. Wasington, DC: American Educational Research Association.

Azwar, S. (2012). *Tehnik penyusunan skala pengukuran*. (*Constructing measurement scale technique*). Yogyakarta: Universitas Gajah Mada Press.

Besemer, S. P. (1998). Creative product analysis matrix: Testing the model structure and a comparison among products-three novel chairs. *Creativity Research Journal, 11*(4), 333-346.

Besemer, S. P., & Treffinger, D. J. (1981). Analysis of creative products: review and synthesis. *Journal of Creative Behavior, 15*(3), 158-178.

Buzan, T., & Buzan, B. (2004). *The mind map book (*Milineum ed.). Batam : Interaksara

Christiaans, H. H. C. M. (2002). Creativity as a design criterion. *Creativity Research Journal,14*(1), 41-54.

Christiaans, H.H.C.M., & Vanselaar, K. (2005). Creativity in design engineering and the role of knowledge: Modelling the expert. *International Journal of Technology and Design Education.,15*(3), 217-236.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 7-46.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Ohio, USA : Cengage Learning. ISBN-13: 978-0-495-39591-1.

Demirkan, H., & Afacan, Y. (2012). Assessing creativity in design education : Analysis of creativity factors in the first-year design studio. *Design Studies Journal*, *33*(3), 262-278.

Demirkan, H., & Hasirci, D. (2009). Hidden dimensions of creativity elements in design process. *Creativity Research Journal*, *21*(2-3), 294-301.

Dorst, K., & Cross, N. (2001). Creativity in design process: Co-evaluation of problem-solution. *Design Studies, 22*(5), 425-437.

Drabkin, S. (1996). Enhanching creativity when solving contradictionary technical problems. *Journal of Professinal Issues in Engineering Education and Practice, 4,* 78-82.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement (5th*Ed.) New Delhi : Prentice Hall of India. ISBN : 087692-700-2.

Eguiluz, L. I., Cavia, M. A., & Lavandero, J. C. (2003). Creativity Test applied in engineering. *Proceeding International Conference on Engineering Education.* July 21-25. Valencia, Spain.

Evans, R. J. (1994). *Berpikir kreatif dalam pengambilan keputusaan dan manajemen*. (*Creative thinking in decision making and management*). Jakarta: Erlangga.

Fisher, W. P. Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transaction*, *21*(1), 1095. Retrived August 30, 2017, from http://www.rasch.org/ rmt/rmt211m. htm.

Fleenor, J. W., Fleenor, J. B., & Grossnickle, W. F. (1996). Interrater reliability and agreement of performance ratings: A methodological comparison. *Journal of Business and Psychology, 10*(3), 367–380.

Guilford, J. P. (1967). *The nature of human intelligence*. New York : McGraw Hill

Guilford. J. P. (1974). *Characteristic of creativity*. Springfield, 11: Illinois State Office of the Sperintcndent of Public Instruction, Children Section.

Guilford. J. P. (1975). *Creativity: A quarter century of progress.* In I. A. Taylor, & J. W. Getzcls, (Eds), *Persprctives in Creativity*. Chicago: Aldine

Haik, Y. (2003). *Engineering design process*. Pasific Grove, CA : Thomson Learning.

Hair, J., Anderson, R. E., Tatham, R. L., & Black, W. C. (2010). *Multivariat Data Analysis* (7thEd.). USA : Prentice Hall

Hasirci, D., & Demirkan, H. (2003). Creativity in learning environments: the case of two sixth grade art-rooms. *Journal of Creative Behavior*, *37*(1), 17-42.

Hasirci, D., & Demirkan, H. (2007). Understanding the effects of cognition in creative decision-making: a creativity model for enhancing creativity in the design studio process. *Creativity Research Journal*, *19*(2e3), 259-271.

Horn, D., & Salvendy, G. (2006). Product creativity: conceptual model, measurement and characteristics. *Theoretical Issues in Ergonomics Science, 7*(4), 395-412.

Horn, D., & Salvendy, G. (2009). Measuring consumer perception of product creativity: Impact on satisfaction and purchasability. *Human Factors and Ergonomics in Manufacturing, 19*(3), 223-240

Ibrahim, B. (2012). *Exploring the relationships among creativity, engineering knowledge and design team interaction on senior engineering design projects*. Unpublished doctoral dissertation, Colorado State University.

Joseph, E. J. (2009). *Effectiveness of Khatena training method on the creativity of form four students in a selected school*. Unpublished master thesis, Faculty of Education, University of Malaya: Kuala Lumpur.

Kaplan, R. M., & Saccuzzo, D. P. (2012). *Pengukuran psikologi, prinsip, penerapan dan isu.* (7th Ed.). (*Psychological testing, principle, applications and issues*). Jakarta: Penerbit Salemba Humanika.

Kvashny, A. (1982). Enhancing creativity in landscape architectural education. *Landscape Journal, 1*(2),104-112.

Laurens, J. M., & Tanuwidjaja, G. (2012) Melalui pendekatan desain inklusi menuju arsitektur yang humanis. (Through an inclusive design approach towards humanist architecture). *Proceeding in Seminar Nasional Menuju Arsitektur yang Berempati*, 04-05-2012-05-05-2012. Retrieved from http://repository.petra.ac.id

Laseau, P. (2001). *Graphic thinking for Architects and Designers* (*3rd*Ed.). Canada : John Willey & Sons, Inc. ISBN : 0-471-35292-6

Lumsdaine, E., Shelnutt, J.W. & Lumsdaine, M. (1999). Integrating creative problem solving and engineering design. *ASEE Annual Conference & Exposition, Charlotte, NC: American Society for Engineering Education*, *Session 2225*. Retrieved from https://peer.asee.org/integrating-creative-problem-solving-and- engineering-design.pdf

Mardapi, Dj. (2012). *Pengukuran penilaian dan evaluasi pendidikan*. (*Measurement of assessment and evaluation in education*). Yogyakarta : Nuha Medika.

Marlinda, E.S., Barliana, M.S., & Krisnanto, E. (2013). Hubungan pengalaman berarsitektur dengan kreativitas desain mahasiswa. (Relationship between Architectural experience with student creativity design). *Jurnal Invotec*, *9*(1), 1-16.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*, 201-232.

Mednick, M. T., & Andreas, F. M. (1967). Creative thinking and level intelligence. *Journal Creative Behavior*, *1*(4), 428-431.

Munandar, S C. U. (1999). *Kreativitas dan keberbakatan. strategi mewujudkan potensi kreatif dan bakat.* (*Creativity and giftedness: Strategies to actualizing potential of creativity and talent*). Jakarta: PT. Gramedia Pustaka Utama.

Musta'ama, A. H., Norman, E., Jabor, M. K., & Buntat, Y. (2012). Does CAD really encourage creative behaviors among its users: A case study. *Proceeding ofInternational Conference on Teaching and Learning in Higher Education (ICTLHE) in conjuction with RCEE & RHED 2012*, *56,* 602-608. Retrieved from http://www.sciencedirect.com.

Potur, A. A., & Barkul, Ö. (2006). Creative thinking in architectural design education. *Proceeding of 1st International CIB Endorsed METU Postgraduate Conference Built Environment and Information Technologies*. Ankara. Retrieved from http://www.irbnet.de/daten/iconda/06059008097.pdf

Potur, A. A., & Barkul, Ö. (2009). Gender and creative thinking in education: A theoritical and experimental overview. *ITU A|Z, 6*(2), 44-57. Retrieved from https://www.researchgate.net/profile/Ayla_Potur/publication

O'Quin, K., & Besemer, S. P. (1999). Creative products. In M. Runco, & S. R. Pritzker (Eds.).*Encyclopedia of creativity*. 413-422. Boston: Academic Press.

O'Quin, K., & Besemer, S. P. (2006). Using the creative product semantic scale as a metric for results-oriented business. *Creativity and Innovation Management, 15*(1),34-44

Semiawan, C., dkk. (2010). *Kreativitas keberbakatan: Mengapa, apa dan Bagaimana*. (*Giftedness of creativity: Why, what and how*). Jakarta : Indeks.

Sternberg J.R. (1999). *Handbook of Creativity.* USA: Cambridge University Press.

Suharnan. (2011). *Kreativitas (Teori dan Pengembangan)*. (*Creativity: Theory and development*). Surabaya: Penerbit Laras.

Sumintono, B. (2014). Model Rasch untuk penelitian sosial kuantitatif. *Conference Paper*. Retrieved from https://www.researchgate.net/ publication/268688670_Model

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model RASCH untuk penelitian ilmu-ilmu sosial.*(Revision Ed.). (*Application of RASCH model for social sciences research*). Cimahi: Trim Komunikata Publishing House

Widhiarso, W. (2010). Melibatkan rater dalam pengembangan alat ukur. (Involving rater in the development measuring instrument). *Article*. Fakultas Psikologi Universitas Gadjah Mada, Retrieved from http://widhiarso.staff.ugm.ac.id/files/

Ostwald, M. J., & Williams, A. (2008a). Understanding architectural education in Australasia. Volume *I*: *An analysis of architecture schools, programs, academics and students*. Sydney : ALTC.

Ostwald, M. J., & Williams, A. (2008b). Understanding architectural education in Australasia. *Volume 2: Results and recommendations*. Sydney : ALTC.

Ostwald, M. J., Askland, H. H., & Williams, A. (2011). Assessing creativity as an aspired learning outcome : a four-part model. *Proceeding of 45th Annual Conference of the Architectural Science Association, ANZAScA 2011*, University of Sidney.

Vernon, P.E. (1970). *Creativity*. Harmonsworth: Penguins Books.

Weisberg, R, W. (1993). *Creativity-beyond the myth of genius*. (2nd Ed.). New York: W.H. Freeman.

Williams A., Ostwald, M. J., Askland, H. H. (2010). *Creativity, design and education. theories, position and challenges*. Sydney: ALTC.

Yamin, S. (2014). *Rahasia olah data Lisrel*. (The secret of processing Lisrel data). Jakarta: Mitra Wacana Media