Examining the Value of Jury Critique for Architectural Design Studio Courses

Frederick C. Santos[1]
fcsantos3@up.edu.ph

Abstract

A common evaluation tool used in an architectural design studio course is the critique. The critique is the process by which students present their final design work as answer to a design problem, have their work examined, and receive feedback on their work from a jury while being observed by faculty and fellow students. The jury in the study were instructors and design professionals with no involvement in conducting the studio courses prior to final critique, while the faculty were instructors who officially handled said courses during the semesters. Students' works were then given scores by both faculty in-charge and jury, with the scores weighted within the students' final grades based on the evaluation method and criteria designed by the faculty in-charge.

While a common occurrence in architecture design studio courses, the value of jury critique has not been adequately examined. It is unclear if the value of jury critique lies in student grading, evaluation approaches, or both. Under these parameters, the study aims to investigate the degree of similarities and disparities between jury and faculty as well as among jurors in evaluating final output through jury critique. This was performed using statistical analysis, where the variable observed was scores given by juries to students that presented their Design course plates, in addition to surveys of jurors who have taken part in courses covered by the study. Particularly, average jury scores were compared to faculty scores and individual juror scores were compared with each other in an attempt to find a pattern of agreement or disagreement among evaluators, with survey responses used to complement and add significance to statistical analysis. A key finding was that a value of jury critique is its ability to accurately evaluate students' output without the potential biases that faculty may have from being reflected in the students' grades. This is significant as a basis to guide how to appropriately use juries as evaluators of students' plates and how grading responsibilities may be divided among jury and faculty. Recommendations were then made based on findings to maximize its effectivity within the architecture design studio course.

Keywords: Jury Critique, Jury Grading, Architectural Design Studio, Architectural Design Evaluation, Architectural Design Assessment, Evaluation Approach

I. Introduction

A. Background

One common method to evaluate the merits of students' designs in an architectural design course is through the "critique" – where the student's final work, or plate, is presented to a panel and is subjected to inquiry and evaluation. Students may be required to present their final works to an audience consisting of the faculty in-charge of the class, fellow students, and a jury whose members may be academics from the same institution, academics from other institutions, or professionals.

While not all design classes employ the jury critique method, it has been a regular and accepted part of architectural education in some form all over the world. Past literature have shown that the jury critique was started in 1795 in the Ecole Des Beaux-Arts (School of Fine Arts) in Paris as a closed-door evaluation of students' works where students were not allowed to participate and had no opportunity for defense, until it was eventually opened up to students in the early 19th century (Carlhian, 1979 & 1980; Chafee, 1977; Egbert, 1980; Kostof, 1977; and Middleton, 1982, as cited in Salama & El-Attar, 2010). As Europe was the basis of Architectural education in North America, the system was eventually adopted there in the 1800s (Kostof, 1977, as cited in Salama & El-Attar, 2010). The guidelines of evaluation were solely on "quality of presentation and drawings, ignoring many of the variables that influence architectural design" (Kostof, 1977; Salama, 1995, as cited in Salama & El-Attar, 2010).

In this study, critique is defined as the evaluation of students' works where they present their output to a jury. The output for the classes examined consisted of presentation boards and scale models. The juries were comprised of faculty with experience handling design courses from the same institution but with no involvement in conducting the studio courses in the study prior to final critique. The students were given three (3) minutes for an oral presentation and twelve (12) minutes to defend their designs to the jury whose members may ask clarificatory questions. The jury, along with the faculty, then scored the students' plates and presentations.

Today, most jury critiques have some form of visual and oral presentation. In addition to the objective of evaluating students' works, critiques are intended to be learning experiences with a potential cumulative effect (Anthony, 1987). Ideally, discussions arising from students' presentations would positively affect the successive plates of students, presenters and viewers alike.

[1] Frederick Santos earned his Bachelor of Science in Architecture from the University of the Philippines, Diliman, Quezon City, Philippines and his Master's in Business Administration from Emory University, Atlanta, Georgia, USA. He has been teaching since 2013 and is currently an Assistant Professor in the Architecture program of UP Diliman.

To what degree do critiques reach this objective, however, may be up for debate.

B. Main Problem

Past writings, particularly "Private Reactions to Public Criticism" by Kathryn Anthony (1987) and "Studio Design Critique: Students and Faculty Expectations and Reality" by Elizabeth Marie Graham (2003), have examined the benefits of jury critique and have found that its main goal of learning for the students is not fully realized. But with a system this widely used in global architectural education, it can be acknowledged that educators see inherent value in the jury critique. Whether this value lies in its application to student grading, developing evaluation approaches, or both, is still unclear. Therefore, under these parameters, the study aims to answer the question: What is the value of jury critiques in architectural design studio courses?

C. Sub-Problems

The following sub problems were addressed.

1. How valid are jury critiques relative to faculty evaluations?
2. How reliable are jury member evaluations relative to each other?
3. In what evaluation categories do jury members generally have agreement?

D. Rationale

Juries have been consistently used by architectural design courses as part of the evaluation system of students' works in architecture schools in the Philippines but the actual value they add to the course has not been studied or examined clearly. In the University of the Philippines College of Architecture (UPCA), the current practice is to conduct final thesis presentations to a jury of between three (3) and five (5) jury members. While its weight is relatively small compared to the grades given by the faculty, students are still required to successfully present and defend their thesis projects in order to pass the final design course. Considering these, a proper assessment of the value of jury critique and proper use will improve how they are used in architectural design courses.

In examining the value of juries to Architectural design studio evaluations, the study may inform on how to best use such juries to align its impact to the evaluation system by correcting its weight within the final grade. It may also inform what areas of the jury critique may be adjusted and how rubrics may be changed to maximize jury value.

E. Goal of the Study

The goal of the study is to examine what value juries hold for architectural design studio critiques to align its value to how they are employed in the final evaluation process. In line with this, the study worked toward the following objectives:

1. To determine the validity of jury evaluations compared to faculty evaluations by comparing final scores given by each group and determining if they are statistically the same.
2. To determine the reliability of jury members' evaluation scores relative to each other and through similar statistical analysis determine if there is statistical agreement of final scores among the group.
3. To determine which evaluation categories that jury members have statistical agreement on by comparing the scores given by jury members for each of the evaluation categories in the forms used by the jury.

F. Scope and Limitations

The study revolved around two (2) sections of second year design classes in UPCA over three (3) semesters, from second semester of academic year 2016-2017 (AY 1617) until second semester of academic year 2017-2018 (AY 1718) and handled by the same two faculty members. The design classes were conducted as combination lecture and studio courses in each of the three (3) semesters with each project phase supported by a lecture and each class meeting having corresponding output requirements. The courses involved were the associated design courses for the first and second semesters respectively of the second year for the BS Architecture degree.

- First Semester: Arch 21 - Architectural Design III: Design & Inter-personal Spaces and
- Second Semester Arch 22 - Architectural Design IV: Design & Social Spaces

A total of 109 students enrolled in the six (6) classes, that is two (2) second-year design classes per semester over the three (3) semesters covered by the study, with each class having a minimum of fifteen (15) students and a maximum of twenty (20) students per class. Of the total number of participants, 61 percent were female and 39 percent were male as shown in Tables 1 and 2.

Table 1. Number of students per course per class

Course	Class 1	Class 2	Total
Arch22_1617	17	20	37
Arch21_1718	20	20	40
Arch22_1718	17	15	32
Total	54	55	109

Table 2. Number of students classified by sex

Course	Sex	Class 1	Class 2	Total
Arch22_1617	Male	5	7	12
	Female	12	13	25
Arch21_1718	Male	7	9	16
	Female	13	11	24
Arch22_1718	Male	7	7	14
	Female	10	8	18
Total		54	55	109

Of the 109 students, only those who were able to submit complete required outputs at the appointed time of submission were allowed to present to juries, which limited the number of students that received jury critiques. During presentations, not all students were critiqued by a complete jury for various reasons, such as jury members that could only attend part of the presentations due to professional commitments or jury members who would need to step out and miss part or a number of presentations. Jury members who missed presentations in part or in whole did not score that presentation. To ensure equal comparisons of scores where no average scores were skewed due to incomplete jury members, students that were evaluated by incomplete juries were removed from the list to be statistically analyzed.

Figures 1 to 2. Students presenting design plates to jurors
Source: Photos by Olivia Sicam

The juries over the three (3) semesters were composed of professional architects currently holding teaching positions in the UPCA, with teaching experience ranging from four (4) years to 23 years. All the jury members teach various courses of the BS Architecture curriculum but most importantly, currently teach or have taught design courses within the past two years. All have Bachelor's Degrees in Architecture from the University of the Philippines and Master's Degrees in Architecture or Planning from various educational institutions locally and abroad, with one jury member earning his doctorate at the time of the study. The criteria for choosing these faculty members to be jurors were their availability at the scheduled critique and their design courses handled.

Figure 3. Juror examining student's sketch model
Source: Photo by Olivia Sicam

G. Assumptions

The study was grounded on the assumption that, in addition to explicit questions and comments from jury members, the value of juries can be seen through the scores they give to students. Jury scores have become part of the final computations of grades, which also include faculty scores of final outputs and presentations and faculty scores of studio process and participation.

Another assumption is that while the jury members' views and priorities may vary, they can show a cohesive view which can be seen from the final scores they give. This may be revealed through the proximity of independently-given scores.

Finally, it was assumed that there are some criteria of evaluation that jurors will agree upon, which can be seen in the scores they give to specific categories of evaluation.

H. Theoretical Framework

The study focused on the value of juries as evaluators in architectural design studio projects. Value was gauged by examining the following: (1) validity of jury total scores relative to faculty scores per plate, (2) reliability of juror individual scores relative to other juror scores per plate, and (3) reliability of individual juror scores relative to other juror scores per scoring category/criterion and grouped categories/criteria. Validity was examined using t-tests and reliability was examined using f-tests, both at a 95 percent confidence level. The results were then identified as above or below the 95 percent threshold and subsequently analyzed to extract recommendations on how to maximize jury value as evaluators in critiques.

The t-test measures whether the averages of two (2) groups are statistically different from each other (Web Center for Social Research Methods, n.d.). In the study, the two groups compared were the average total scores given by the jury and the total scores given by the faculty for each critique. The t-test formula is a ratio that analyzes the difference in means relative to the variability of the given scores within groups. To be statistically different means that there is relatively little overlap between the scores given by one group as compared to the other.

F-tests are similar to t-tests in that they also assess if the means of groups are statistically different from each other. However, f-tests are able to compare the means among multiple groups while t-tests are only able to compare the means of two groups (F-test, n.d.). The groups compared in the study were the scores given by individual jury members. Total scores were compared to test their reliability per plate while category scores were compared to test reliability per scoring category and for the grouped categories of *Process*, *Design*, and *Output*.

Confidence levels in statistics is related to standard deviation, which is a measure that describes the variation or dispersion in a set of data (Bland, 1996). Put simply, a low standard deviation describes a set of data with data points close to the mean, while a high standard deviation describes a set of data with data points spread out over a wider range of values. In terms of confidence levels, a 95 percent confidence level is equal to two standard deviations in either direction away from the mean, which assumes that the range defined by two standard deviations contains 95 percent of all data points within the set (68 95 99.7 Rule in Statistics, n.d.). Thus a 95 percent confidence level assumes that the number arrived at applies to 95 percent of the population.

The words assessment and evaluations are often used interchangeably even if they differ in their meanings. Assessments attempt to establish how close students are to an intended goal while evaluations are used to estimate students' abilities (Gielen, Dochy, Onghena et al., 2011; Green & Johnson, 2010, as cited in Strang, 2015). The design of the rubric in the study attempts to gauge both. Thus, the term evaluation was used to refer to both accomplishment of goals and student ability.

Validity is the degree to which the scoring tool "provides an accurate, representative, and relevant measure of student performance for its intended purpose" (Green & Johnson, 2010, as cited in Strang, 2015). Reliability refers to the level that given scores on the tool "are consistent and stable across multiple raters, namely students, faculty or combinations of both" (Green & Johnson, 2010, as cited in Strang, 2015). In the study, jury score validity and reliability were tested using the scores they gave students based on the standard rubric given by the faculty during critiques. If the tests find that the groups being compared are statistically different, then the scores given by the jury are considered either *not valid* or *not reliable*. Tests were done for both combined and individual classes for each semester. Figure 4 shows the research design diagram.

A survey which gathered insight into jurors' priorities when critiquing students' work and their views on the value of critique sessions was given to jurors. The findings from survey responses will be used to provide context to the findings from statistical analysis, which would be the bases for conclusions and recommendations.

I. Review of Related Literature

Effectiveness of Peer Assessment in a Professionalism Course Using an Online Workshop (Strang, 2015)

Strang performed t-tests and reliability estimates on peer assessments grades submitted through Moodle Workshop on a Seminar on Professionalism, an undergraduate course taught in two campuses, to test their effectiveness in terms of validity and reliability. Moodle Workshop is an online peer assessment tool where students submit their work to be peer-assessed while they assess other students' works as well. The grades submitted through the tool were evaluated to determine if the grades provided by peers through the tool were consistent with faculty expectations. Validity was tested by comparing the means of peer assessment grades to faculty grades and reliability was tested by comparing the individual peer assessment grades to each other.

The results showed that student grades were consistent relative to faculty grades on the same assignment and therefore valid. They also showed that student grades were consistent relative to each other on the same assignment and therefore reliable.

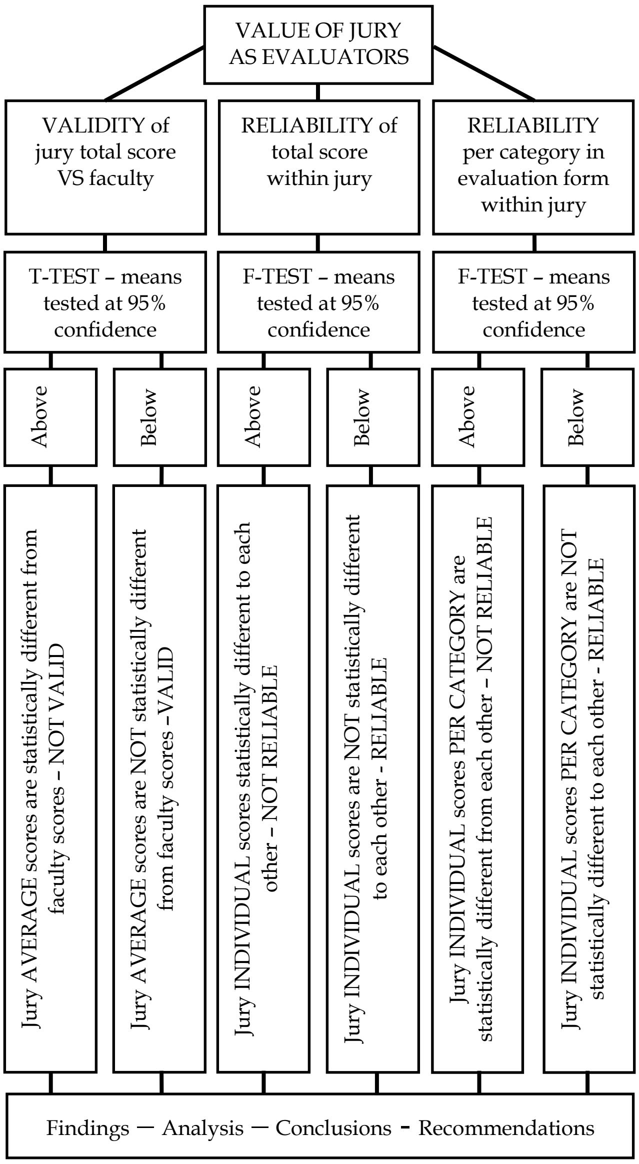
VALUE OF JURY AS EVALUATORS
VALIDITY of jury total score VS faculty
RELIABILITY of total score within jury
RELIABILITY per category in evaluation form within jury
T-TEST – means tested at 95% confidence
F-TEST – means tested at 95% confidence
F-TEST – means tested at 95% confidence
Above
Below
Above
Below
Above
Below
Jury AVERAGE scores are statistically different from faculty scores – NOT VALID
Jury AVERAGE scores are NOT statistically different from faculty scores –VALID
Jury INDIVIDUAL scores statistically different to each other – NOT RELIABLE
Jury INDIVIDUAL scores are NOT statistically different to each other - RELIABLE
Jury INDIVIDUAL scores PER CATEGORY are statistically different from each other – NOT RELIABLE
Jury INDIVIDUAL scores PER CATEGORY are NOT statistically different to each other - RELIABLE
Findings — Analysis — Conclusions - Recommendations

Figure 4. Research design diagram

Student Perceptions of the Architectural Design Jury (Salama & El-Attar, 2010)

Salama and El-Attar (2010) states that the jury system in architecture education has been well-documented and studied in the West but minimally so in the Arab context. To fill the void, they conducted two studies in an attempt to ascertain practices in the jury system and student perceptions by studying selected cases from educational institutions in Egypt and Saudi Arabia. Understanding the jury system in this setting as well as its challenges resulted in recommendations to improve the system's use.

Literature review in the study found that the main value of the jury system lies in helping students learn to solve architectural problems as well as giving them a framework to follow to improve their current or future projects. However, the system has been criticized that due to innate pressures within it, students claim to not have learned much from jury comments or state that they cannot remember anything about the projects not their own.

The studies by Salama and El-Attar found that students preferred the involvement of juries to increase objectivity of evaluation but the learning goals were seldom met. Students stated that verbal presentation skills, rather than solving architectural problems, were the most important aspect learned from juries and board layout became a primary focus to attract the attention of the juries.

Studio Design Critique Students and Faculty Expectations and Reality (Graham, 2003)

Graham (2003) investigated the use of criticism in the context of Landscape Architecture studio. She discovered from her research that criticism is more about the critic rather than the one criticized. There will always be some degree of bias in a critique because criticism is a behavior where people express their own views of an object being criticized in the "interest of a more adequate perception."

The study found that faculty and student expectations are achievable but not always fulfilled. Students expect juries to be focused and give equal time to all students, but this is not always the case due to the differences in the students presenting, project quality, and jury construction. They also expect criticism to be constructive, delivered tactfully, and to benefit all students whether as presenter or viewer. However, these too are occasionally missed.

Instructors feel that the best juries are those that fuel student interaction and discussion but the reality is that many times students are not engaged unless they are the ones presenting due to reasons such as lack of sleep. Instructors also felt that juries should generate a grade from the critique because students want critiques to directly relate to their grades however, most students seem to not feel the same way.

Even with these drawbacks, critiques still can achieve the goals for learning. After experiencing an effective jury, both faculty and students believe it reasonable to expect experiences of similar quality from successive juries.

Private Reactions to Public Criticism: Students, Faculty, and Practicing Architects State Their Views on Design Juries in Architectural Education (Anthony, 1987)

Anthony (1987) investigated the jury system in architecture education by looking at four (4) aspects relating to it: (1) how educationally valuable the jury system is and to whom, (2) if interim and final juries were equally effective teaching techniques, (3) how students cope with public criticism, and (4) how behavior patterns of architecture students differ from other fields. A study was conducted over the course of one (1) academic year with phase one using the case of an Architecture school at a western university in the United States and the phase two being conducted at an ACSA Teachers' Seminar.

Results of the study directly relating to the initial questions found that (1) since academic juries and professional juries are fundamentally different, the reasoning of jury critiques benefitting students due to their simulating professional environment is questionable; (2) interim juries tended to be a more effective learning tool than final juries as jury comments can immediately be applied to the current project; (3) students tend to react defensively and nervously to jury comments that were typically not delivered tactfully or not perceived to be constructive; and (4) studio culture inherent to architecture education can foster students with a sense of belonging but may be detrimental to their physical and mental health due to lack of sleep and improper nutritional habits.

II. Methodology

A. Observation and Documentation

Critiques were conducted twice per semester at the end of each plate. Both faculty in-charge were present during all critique sessions and observed presentations and interactions among students and jury members. Photos of critiques were taken during each session and photos of boards and models taken after critique completion.

B. Desk Research

Research was performed to find relevant writings about both jury and peer evaluations, specifically applied to Architecture education if available. The goal was to find a framework of testing to examine value of jury evaluations of students works and a background of context to better understand any findings. The framework found was adjusted to be applied to the study.

C. Juror Survey

A survey form was sent to all former jury members involved in the three (3) semesters covered by the study. Aside from general data, the survey gathered insight into their priorities when critiquing students' work and their views on the value of critique sessions. Six (6) out of the seven (7) jury members answered the survey.

D. Critique Evaluation Form

Jury evaluation forms were given to all jury members for each critique session to use as scoring sheets and to guide the jury in their scoring. For the first and second semesters included in the study, the form contained five (5) scoring categories, shown in Table 3, each with a weight of 20 percent. In the third semester of the study, the critique evaluation form was changed to a 10-category format, shown in Table 4, with each category weighted equally at 10 percent. The change was decided by the faculty members in-charge in order to have a more accurate and detailed guide in scoring.

Table 3. Evaluation Categories for Arch 22, second semester AY 1617, and Arch 21 first semester AY 1718

Category	Weight
Design Concept & Translation	20%
Responsiveness to Site & Context	20%
Elements of Design (Line, Color, Massing, Programming, Aesthetics)	20%
Creativity, Originality, & Innovation	20%
Workmanship/Craftmanship & Oral Presentation	20%

Table 4. Evaluation Categories for Arch 22, second semester AY 1718

Category	Grouped Category	Weight
Approach	Process	10%
Development		10%
Concept		10%
Site and Context	Design	10%
Form		10%
Space		10%
Creativity & Innovation		10%
Boards	Presentation	10%
Model		10%
Oral Presentation		10%

E. Quantitative Analysis

Validity and reliability of jury scores were determined using scores given on critique evaluation sheets. Validity was tested by comparing the proximity of jury scores to faculty scores through t-tests while reliability was tested by comparing the proximity of individual juror scores to each other through f-tests. All scores were tested on a 95 percent confidence level and the corresponding p-values were determined. Score comparisons with p-values greater than 0.05 were considered not statistically different and therefore *valid* or *reliable*, while those with p-values less than 0.05 were considered statistically different and therefore *not valid* or *not reliable.*

III. Findings and Analysis

A. Findings

1. Validity

T-tests were conducted to examine the validity of the scores given by juries relative to faculty. When two (2) classes per semester were combined and scores per plate were tested, it was found that scores were not statistically different and therefore *valid* for half of the plates tested – the first three (3) plates were not valid but the latter three (3) plates were valid. When the classes were tested individually, Class 1 jury scores were valid for two (2) out of the six (6) plates and Class 2 jury scores were valid for five (5) out of six (6) plates as shown in Table 5.

Table 5. Validity per plate

Course	Plate	Combined	Class 1	Class 2
Arch22_ 1617	Plate 1	Not Valid	Not Valid	Not Valid
	Plate 2	Not Valid	Not Valid	Valid
Arch21_ 1718	Plate 1	Not Valid	Not Valid	Valid
	Plate 2	Valid	Valid	Valid
Arch22_ 1718	Plate 1	Valid	Not Valid	Valid
	Plate 2	Valid	Valid	Valid

2. Reliability per Plate

F-tests were conducted to examine the reliability of total scores given by individual jurors relative to other jurors' total scores. It was found that the jury scores of three (3) out of the available five (5) plates tested were not statistically different and therefore *reliable* for the combined class. Jury scores for four (4) out of five (5) plates were reliable in Class 1. Class 2 jury scores were reliable for all five (5) plates tested. A summary is shown in Table 6.

Table 6. Reliability per plate

Course	Plate	Combined	Class 1	Class 2
Arch22 _1617	Plate 1	Reliable	Reliable	Reliable
	Plate 2	-*	Reliable	-*
Arch21 _1718	Plate 1	Not Reliable	Not Reliable	Reliable
	Plate 2	Reliable	Reliable	Reliable
Arch22 _1718	Plate 1	Not Reliable	Not Reliable	Reliable
	Plate 2	Reliable	Reliable	Reliable

*Test skipped - individual jury scores unavailable

3. Reliability per Evaluation Category

F-tests were conducted for jury scores of both plates of Arch 22 AY 17-18 to test their reliability for individual scoring categories, shown in Table 7, where it was found that only two (2), Form and Creativity & Innovation, of the ten (10) scoring categories were reliable for Plate 1. Three (3) categories, namely Approach, Development, and Concept, could not be tested for Plate 2 due to the jurors choosing not to score them individually but as a group. Of the seven (7) remaining categories that could be tested individually for Plate 2, six (6) were reliable.

Table 7. Reliability per evaluation category

Categories	Grouped Categories	Plate 1	Plate 2
Approach	Process	Not Reliable	-*
Development		Not Reliable	-*
Concept		Not Reliable	-*
Site and Context	Design	Not Reliable	Reliable
Form		Reliable	Not Reliable
Space		Not Reliable	Reliable
Creativity & Innovation		Reliable	Reliable
Boards	Presentation	Not Reliable	Reliable
Model		Not Reliable	Reliable
Oral Presentation		Not Reliable	Reliable

When related categories were grouped and new grouped categories of Process, Design, and Presentation were formed, f-tests were conducted again to test reliability for the new groups. It was found that the grouped categories of Output and Presentation were not statistically different and therefore *reliable*, with Process as the only grouped category that was not reliable. Table 8 shows a summary.

Table 8. Reliability per evaluation category

Categories	Grouped Categories	Plate 1	Plate 2
Approach	Process	Not Reliable	Not Reliable
Development			
Concept			
Site and Context	Design	Reliable	Reliable
Form			
Space			
Creativity & Innovation			
Boards	Presentation	Reliable	Reliable
Model			
Oral Presentation			

4. Juror Survey

It was found that 83 percent of respondents cited presentation boards and model, the visual part of the required output, should be students' first priorities to successfully convey information about their designs, with one respondent stating that "Design must be appreciated with minimal verbal or oral intervention trying to explain a designer's intent. [Architects'] medium is visual and should be appreciated and understood as is." However, also of note is that one responded ranked oral presentation as first priority, stating that visuals, which are the drawings and model, "should support the narrative" conveyed through oral presentation. Table 9 provides a summary of the responses for Priority 1 and 2, with Priority 3 filled in through process of elimination. Figures 5 to 6 show examples of presentation boards, figure 7 of a model.

Table 9. Jurors' priorities in evaluation of students' work

Juror	Priority 1	Priority 2	Priority 3
J01	Oral Presentation	Presentation Boards	Model
J02	Model	Presentation Boards	Oral Presentation
J03	Presentation Boards	Model	Oral Presentation
J04	Presentation Boards	Model	Oral Presentation
J05	Presentation Boards	Oral Presentation	Model
J06	Model	Oral Presentation	Presentation Boards

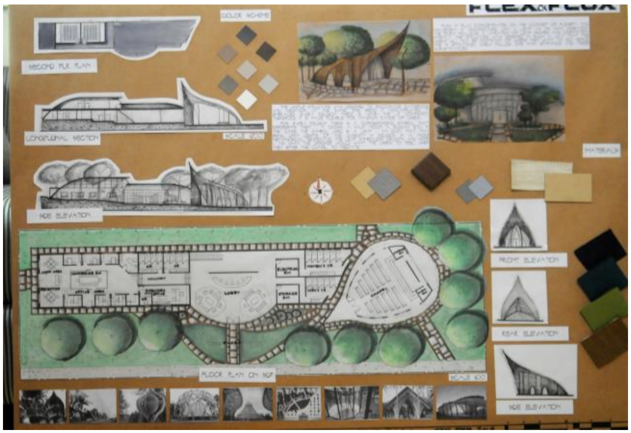

Figure 5. Examples of boards presented during critiques
Source: Photo by Olivia Sicam

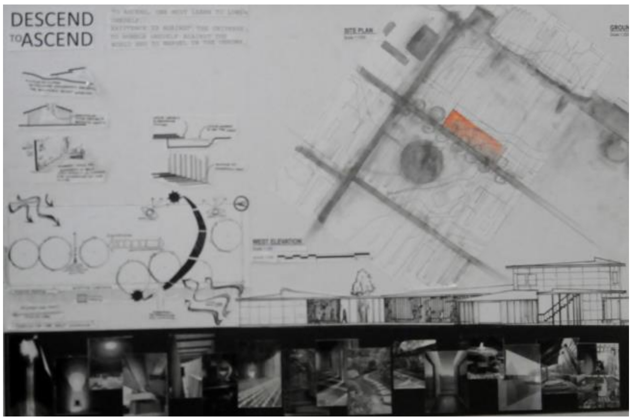

Figure 6. Examples of boards presented during critiques
Source: Photo by Olivia Sicam

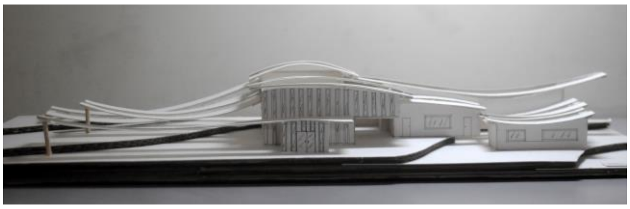

Figure 7. Example of models presented during critiques
Source: Photo by Olivia Sicam

Sixty-seven percent responded that process was more of an influence than final output when asked about which influences jury members more in scoring students' designs, shown in Table 10.

Table 10. Jurors' main influence in scoring students' work

Juror	Jurors' bigger influence in scores given to students – Process or Output
J01	Process
J02	Process
J03	Output
J04	Output
J05	Process
J06	Output

When asked if jury critique is valuable to architectural education, 100 percent responded that it was valuable. However, responses were evenly split between more valuable and equally valuable when comparing the jury critique to a faculty critique. Notable is that there was no response saying that jury critique had no value to architectural education. A summary is shown in Table 11.

Table 11. Jurors' perceptions on value of critique

Juror	Jurors' perception if jury critique has value to architecture education	Jurors' perceived value of jury critique relative to faculty critique
J01	Yes	Equal
J02	Yes	More Valuable
J03	Yes	Equal
J04	Yes	More Valuable
J05	Yes	Equal
J06	Yes	More Valuable

Fifty percent responded yes when asked if average jury score should ideally be close to scores given by faculty while only 33 percent responded the same when asked if jury total scores, the sum of the scoring categories, should ideally be close to each other. Similarly, only 17 percent of the respondents answered yes when asked if jury criteria scores, the scores per category on the evaluation form, should ideally be close to each other. Summary is shown in Table 12.

Table 12. Jurors' perceptions on proximity of scores

Juror	Perception if juror average scores should be close to faculty scores	Perception if juror total scores should be close to other jurors' scores	Perception if juror scores per scoring category should be close to other jurors' scores
J01	Yes	Yes	Yes
J02	No	No	No
J03	Yes	Yes	No Opinion
J04	No	No	No
J05	No	No	No Opinion
J06	No Opinion	No	No

B. Analysis of Results

1. Validity

When classes were combined, results showed only half the plate scores were found valid over the three (3) semesters. Valid scores began appearing in the second plate of Semester 2 and the first and second plates of Semester 3. That is, the second half of the duration of the study. However, when the scores were tested per class, Class 1 only had two (2) plates out of six (6) that were valid, while Class 2 had five (5) out of six (6). Considering that the jury was the same for both classes and students from either class presented alternately thus equalizing jury potential fatigue that may affect their scoring for both classes, the validity appears to be heavily influenced by the differences in faculty scoring. One possible source of this difference is the different scoring standards of the faculty in that even with a standard rubric used as a guide, design evaluation is still highly subjective and differences in evaluations are to be expected. However, there is perceived value in the differences that jurors offer to the critique as only half the respondents believe that jury scores should ideally be close to faculty scores and less than half believe juror scores should be close to each other. This is further bolstered by comments of the jury members surveyed such as "the function of the jury is sometimes to look at a project in a different light" and "the jury should be independent in giving their scores and it would also depend on the professional experience of the jury which may not be the same as the other jury members." While one respondent mentioned that "scores are a measure of objectivity," this also assumes that all evaluators have the same or similar design values, which can be influenced by their experience as professional architects and educators and may not be the case for all instances as well as difficult to control for in the composition of the jury.

Another possible reason for the differences in validity of the scores of individual classes may be the faculty's exposure to the studio meetings. While the scores were purely based on the students' output, it would be difficult for the faculty to fully compartmentalize output from process during studio. There is the prospect that the observations made by the faculty during studio still influenced scoring for output, and this influence may have been more evident in Faculty 1's scores. Jurors also understand this possibility as shown by the comment, "the faculty in-charge can sometimes develop certain biases or tendencies that the jury are free from."

The fact that the scores during the first half of the study were not valid but valid for the second half may be the result of the faculty and jury not only learning to use the evaluation method but also learning the expectations and values of other evaluators whether faculty or juror. The interactions among jurors during critiques and their interaction with the faculty afterward could alter the expectations of the evaluators and may eventually be seen from the scores they give to students' works.

2. Reliability per Plate

Results indicated that the main influencer of reliability may be the jury's understanding of the students' works. It was observed that all of the second plates per semester were reliable, which may be due to an increased preparedness of the students in producing and presenting their output. Salama & El-Attar (2010) found that students judged "development and improvement of verbal presentation skills" as their most important learning in jury critiques and it would make sense that students would improve on their oral presentations for the successive critiques. Also, given that 83 percent of respondents say that visual output should be students' priority in conveying information about their designs and visuals play a large role in architectural education as well as profession, improved visual output would help better communicate their designs to the juries. From observation, jurors would often comment on the quality, whether good or bad, of the boards and models to convey information. These comments would also push the students to improve on visual output for the next plates. Of note is one (1) respondent cited oral presentation as first priority and while having a different opinion, understanding still was the reason for the response, taken from the comment "The narrative is key to understanding the design process".

In addition to the presentations themselves, it was observed that discussions between the students and the jury brought up more relevant matters and as they continued, agreement among the jury about the students' plates seemed to increase. Where in a 1993 round table discussion at Harvard University, participating faculty members agreed that the jury system is "an opportunity for developing theoretical discourses for ideas to thrive utilizing the work of students as a catalyst for discussion" (Dilnot et. al. 1993, as cited in Salama & El-Attar, 2010), jurors involved in critiques during the study would occasionally also delve into discussions of ideas kicked off by the plate and these could influence the agreement among jury members in their evaluations of students' works. One juror believed that extended discussions could help jurors understand the students' works more clearly and would adjust the scores accordingly from the comment, "I wonder if the effectiveness of the critique corresponds with the length of the session. I surmise that the longer the critique session, the jury will get a better sense of the proper grade."

3. Reliability per Evaluation Category

Results showed that two (2) of ten (10) individual scoring categories for Plate 1 and six (6) of seven (7) for Plate 2 were reliable. The differences in reliability of categories between Plate 1 and Plate 2 may be due to the different jurors involved in the critiques. This shows that potential swings in scores depended on jury composition. An interesting thing to note is that once similar categories were grouped, a certain pattern appeared - Process was not reliable but Design and Presentation were consistently reliable for both plates.

While scores of Process were not reliable, 67 percent of the respondents say that they prioritize process over output when scoring students' works. When looking at the comments of the two that cited output as bigger influence, it would seem that one of the respondents still prioritized process and judged its quality through the output, commenting that "[the] output is the result of the process. A poor process will definitely show in the output." The other respondent prioritized output because of its major weight in the scoring sheet, commenting "[it] usually is the bigger component in the grading sheet." Therefore, while it cannot be determined if more than 67 percent of the respondents prioritizes process, the comments hint at such. This seems to show that jurors place importance on process but agreement only occurs when judging design and presentation. It is possible that this agreement comes from the fact that design and presentation can be seen in the output while the jury relies solely on the students' presentations at the time of critique to learn about their processes. This signals the importance of presentation skill during critiques, in that jury evaluation of students' processes are heavily influenced by how well it was conveyed through the visual output or through oral presentation.

IV. Conclusions and Recommendations

A. Validity

The difference of validity per class combined with the subjective nature of design illustrates the challenges to consistently have valid jury scores relative to faculty scores. The differences in evaluation standards of jurors and faculty as well as the numerous combinations of faculty and jury composition would all have influence on the validity of given scores, and these cannot be immediately judged before any actual critique.

Another challenge for validity is the faculty's involvement in studio meetings. The frequency of studio meetings and the number of students per class could create situations of the faculty missing potentially significant points in students' processes or designs. Additionally, this close interaction among faculty and students in studio may make it increasingly difficult for faculty to remain objective when only looking at students' output.

The challenges mentioned as well as survey responses lead to the conclusion that the value jury critiques possess would be the fresh perspective they offer. The jury critique could also be used to balance the partiality of the faculty. Therefore, it is recommended that the weight of jury scores in the final computation of grades be reviewed given its value as an impartial evaluation of students' final output. The proper weighting of jury scores could lessen the effects of faculty's potential biases that stem from their involvement in studio and interaction with the students prior to the critique, that may be seen even in their given scores for output. It would also mitigate skewing of scores for process against those who do not do well with visuals and oral presentation.

B. Reliability per Plate

The synthesis of multiple points of view during discussions would help students in their next plates but more time would be needed to make these discussions meaningful. If used to supplement faculty scoring, a more reliable jury score would be the true measure of the merits of students' output and could counter potential biases of or missed points by the faculty. However, the 3-minute presentation and 12-minute discussion per student critique in the study seem to be too short to fully convey and understand a student's work which is relied upon by the jury when scoring students' output. To increase reliability of jury scores, an increase in the length of critique sessions is recommended. A lengthier and deeper discussion among the jurors and students about the work could lead to a better understanding of the design by the jury, helped by the prompting of other jurors as well as students' answers, especially with the weight juries give to process which may not be easily seen in the output. In the current practice of thesis deliberations in the college for fifth year students, sixty (60) minutes is allotted for the student to present and defend their work – fifteen (15) minutes to present, ten (10) minutes for board inspection, and thirty-five (35) minutes for questions. A critique that more closely simulates the thesis model may prove advantageous in this regard. While this increased time for critique per student may be difficult to achieve with the current class size of fifteen (15) to twenty (20) per class, the lesser number of students critiqued per session may be beneficial in lessening the repetition of some parts of presentations and deepening the conversation about the students' designs.

C. Reliability per Evaluation Category

Based on juror responses, process is generally a greater priority for juries; however, it cannot be immediately seen from students' output and greatly depends on how well students' show their process through either visuals or oral presentation. Therefore, the reliability in design and presentation would show that letting juries score these grouped categories would give a more accurate evaluation of students' work in these particular areas. The lack of reliability when scoring process would suggest that this should be scored purely by the faculty involved, as they are the ones immersed in studio and are able to observe how the students' work during studio. This may change in higher-level design courses as students' will be more experienced in design presentation and better able to convey their processes through their visual output and narrative.

D. Recommendations for Future Studies

The study had a generally homogeneous composition of jury members in terms of sex, age, and educational background. A more diverse composition may yield different results. Future studies may take this into consideration in the assessment of jury composition.

During critiques themselves, better efforts must be made to ensure that all students present to complete juries. This would increase the number of critique scores to be tested and would provide a more robust set of findings and conclusions.

It is finally suggested that future studies review the rubric design for grading. There may be ways that the categories for grading can be altered to better guide the jurors in scoring as well as be more appropriate to jury critique evaluation.

Acknowledgment

The researcher wishes to thank the following people for sharing their time, effort, and knowledge: Ging Santos for helping direct the data analysis; Olivia Sicam for guidance in writing; Marc Espino, Asia Guerrero, and the other students of the Design courses involved in the study; Prof. Johannes Chua and the other jurors who gave their valuable insights.

References

68 95 99.7 Rule in Statistics. (n.d.). Retrieved from Statistics How To: http://www.statisticshowto.com/68-95-99-7-rule/

Anthony, K. H. (1987). Private Reactions to Public Criticism: Students, Faculty, and Practicing Architects State Their Views on Design Juries n Architectural Education. *Journal of Architectural Education*.

Bland, J. M. (1996). Measurement Error. *The BMJ*.

F-test. (n.d.). Retrieved from Statistics How To: http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/f-test/

Graham, E. M. (2003). *Studio Design Critique Students and Faculty Expectations and Reality.* Louisiana State University and Agricultural and Mechanical College.

Russell, M. K. (2012). *Summative Assessments, Classroom Assessment. Concepts and Applications (7th ed., Vol. 5).* Mcgraw-Hill.

Salama, A. M., & El-Attar, M. T. (2010). Student Perceptions of Architectural Design Jury. *International Journal of Architectural Research*, 27.

Strang, K. D. (2015). Effectiveness of peer assessment in a professionalism course using an online workshop. *Journal of Information Technology Education: Innovations in Practice, 14*.

The T-Test. (n.d.). Retrieved from Web Center for Social Research Methods: https://socialresearchmethods.net/kb/stat_t.php