BRINGING STRUCTURE TO JUDGING SUCCESS IN ARCHITECTURAL DESIGN: THE ‘TIOSE’ QUALITATIVE MEASURE

Philip D. Plowright and Matthew L. Cole

Abstract

Explicit measures of design quality are almost unknown in architectural design, particularly since judging quality in architecture, as in most design disciplines in general, resists quantification. Resistance in the judgement of design quality can be traced to requirements for evaluating complex intra- and inter-discipline factor integration, as well as cultural knowledge which is tacit in nature. This paper aims to introduce the TIOSE Qualitative Measure, an outcome evaluation tool developed for a study on team interaction, cognitive style, team processes, and quality in design. The TIOSE contains five factors for judging architectural design success: thoroughness, informativeness, organisation, synthesis, and evocativeness. The methodology for the study involved evaluating the intra- and inter-rater reliability of six architectural critics who utilized the TIOSE to judge team project quality and success among 84 graduate students working in 12 teams. Analysis found high intraclass correlations within and between critics across three rounds of ratings, suggesting the TIOSE showed excellent intra- and inter-rater reliability, respectively. Study results provide evidence that evaluating the innovation and creativity of architecture and design projects can reliably extend past ratings of efficiency (time) and productivity (volume) as objective performance measures as the testing mechanism for success. In conclusion, this study demonstrates that qualitative measures of project success can be quantified in a reliable manner and can attain internalized agreement between architectural critics. The TIOSE measure provides a rating scale with explicit definitions for the qualitative measure of architectural design quality.

Keywords

Design judgement, inter-rater reliability, qualitative factors, project success.

Introduction

A measure for design quality success was needed as part of a larger study on the interaction between inter-personal relationships, team structures based on cognitive style, and design outcomes. While the study stressed communication structures, feedback and conceptual development, the critical factor for team interaction success was final proposal quality. Architectural quality, however, is a factor that is considered contentious and ill-defined, falling into the “I know it when I see it” category (Ronn, 2011). In addition, the assessment of architectural design in normative architectural culture is often completed as an unstructured, negotiated and ad hoc event. Outside of architecture, studies needing to
determine success use objective performance measures such as task completion duration and productive outcome volumes, factors of little relevance for success in architectural design.

Accordingly, the TIOSE Measure for the Judgement of Architectural Design Success was developed in order to address the gap between objective performance measures and qualitative, unstructured assessment. The measure is focused on rating the qualitative significance and relevance of an architectural design proposal. The TIOSE was evaluated for inter-rater and intra-rater reliability using the intraclass correlation coefficient (ICC). The measure needed a high degree of validity and reliability in order for it to be useful to the overall study.

Assessing Quality in Architectural Design

Explicit measures of design quality rating are almost unknown in architectural design, particularly since knowledge held in architecture, as in most design disciplines in general, is tacit in nature and resists quantification. Architecture presents a system of knowledge which can be defined as “poorly articulated” (Cummings & Teng, 2003), making it difficult to learn, teach, and study. Research that examines innovation and creativity (design) is predominantly focused on judgement as on objective performance measure (Vanderheyden et al., 2010) which tends to rate efficiency and productivity as the testing mechanism for success (Kickul & Neuman, 2000; Priola et al., 2004; Volkema & Gorman, 1998). Research that departs from this norm utilizes poorly defined qualitative judgement processes that assume agreement in terminology (Dorst, 2001; Kruger & Cross, 2006). Nevertheless, rating scales for qualitative measures should be capable of development since internalized agreement between architectural critics has been shown to exist (Rönn, 2011; Svensson, 2010).

The focus for a rating scale addressing documented success in architectural design needs to address the quality of the final proposal. As Rönn notes “Architectural quality is the combination of elements that form a whole. This is fundamental for the assessment of projects, especially in early stages. Quality in architecture and urban design is seen as a holistic idea among professionals” (Rönn, 2011). Architecture is a systems-based discipline, meaning that it is impossible to simply add up parts to create an understanding of the whole. Instead the whole is judged as a new synthesized entity that is independent to either the time it took to create, the individual strength of isolated component or the physical volume of information about the product.

Study Methodology

The TIOSE Qualitative Measure for the Judgement of Architectural Design Success is an outcome evaluation tool for the measurement of quality in an architectural design. It is not focused on physically constructed projects (built work) but on schematic design proposal. As a new tool that was developed to satisfy the need for an outcome evaluation of architectural design quality, the TIOSE was tested for soundness in psychometric properties according to reliability testing. Consequently, we evaluated the measure’s reliability among six raters (i.e., critics/judges) using the intraclass correlation (ICC) index of the reliability of a measure.
Description of the TIOSE Measure
The TIOSE Measure contains five (5) factors for architectural design quality judgement. The factors were developed by reviewing the priorities of multi-year outcomes of architectural design projects, discussions with master practitioners and design educators, and the identification of issues involved in core success of architectural projects over a multi-year period.

The five factors of the TIOSE are divided into two groups. The first group contains the three baseline factors, Thoroughness, Informativeness, and Organisation. These factors pertain to research, data gathering, communication, and project structure.

Thoroughness refers to the depth of information and addresses both lateral and longitudinal vectors. In both directions, it is fundamentally about research, data collection and knowledge. On the lateral level, the factor includes initial research breadth, exploration for opportunities and alternatives and cross-disciplinary investigation. The longitudinal vector includes evidence of rigorous application of research. On individual items and applications, there needs to be depth and specificity in content. Ultimately, scoring is dependent on evidence of clear and intimate knowledge of the research subjects.

Informativeness addresses the ability of the project to communicate its content and intentions. The factor rates the proposal’s position for clarity and its content as understandable in both graphic and written communication elements. While informativeness has strong responsibilities in helping the other factors be successful, especially synthesis and evocativeness, it is considered a separate baseline factor. Voice, style, tone, and clarity need to be addressed and integrated with project intention.

Organisation refers to the structure of the proposal, with higher ratings provided to projects with a strong organisational logic of both conceptual and physical elements. This factor refers to the project itself and not the presentation/communication, which is rated by informativeness. Instead, organisation refers to the arrangement of critical architectural elements, experiential intentions, activity containers, and formal structuring of the layers of the proposal resolution.

The second group contains the two design quality factors, Synthesis and Evocativeness. Group two factors are considered advanced factors involving complex intra- and inter-discipline integration, as well as cultural knowledge. This second group of factors requires competency in all three baseline factors in order for the final proposal to be successful.

Synthesis addresses issues of relevance and quality. To be rated highly in this factor, the project must produce a proposal resolution which has generated unique and new entity from the combination of two or more entities. Entities considered for source material are not just simply formal objects, but also can be sub-systems, patterns, experiences, activities and experiences. Part of the success of synthesis is the combination of disconnected and disassociated entities into new and significant wholes. Synthesis, in this way, is not simply the adjacency of entities, but rather, it is the combination of like and dislikes constituent elements into a new, cohesive, unified entity.
Bringing Structure to Judging Success in Architectural Design: The TIOSE Qualitative Measure

PHILIP D. PLOWRIGHT AND MATTHEW L. COLE

This factor also includes the integration of, rather than avoidance of, existing context.

Evocativeness refers to the other major factor in judging quality in architectural design—significance. A high rating in evocativeness implies that the project engaged either an evocative or a provocative position. Evocative work has the power to extend intentions and concerns into the world and address cultural values. The project should either call up memories, associations and conditions that are of concern, or produce opportunities among this particular configuration of layers. If provocative, the overall intentions and resolution of the proposal should be stimulating, insightful, and challenging to normative or baseline practices. A high rating in this factor implies that the proposal has clearly initiated a deeper conversation that draws attention to important but overlooked opportunities, issues, or potentials. The overall effect is to invoke or rouse discussion regarding the evocative and provocative intentions of the project. Either of these aspects of evocativeness should have clearly exploited unseen opportunities, rather than simply addressing an apparent ‘problem’ that the proposal needs to extend past a one-dimensional gesture. Consequently, the factor considers a significant human dimension.

All five factors are presumed to interact with each other to yield the final outcome evaluation score. Nevertheless, factors from group one should be considered individually, whereas factors from group two should be considered from a systems-based perspective where the whole is not only generated by the parts, but also extends beyond the parts.

Research Study for which the TIOSE was developed

The TIOSE was developed as the architectural design quality outcome measure for a team-based graduate research study, “Third Coasting: Ecology Networks in the Great Lakes”. The study occurred during the summer semester, and involved research, enquiry, exploration, and a design proposal centred on ecology networks, human habitation and formal/system interventions. The study had five intervals during which team work was discussed and rated using the TIOSE. Each project interval had duration of two weeks, with the exception of interval 3 which spanned three days. While interval 1 consisted mainly of research and interval 3 was considered exploratory; intervals 2, 4, and 5 were of equal duration and complexity with matched outcomes. For this study, only time intervals 2, 4, and 5 were used for data analysis due to their content, process and outcome equivalence. These intervals are referred to respectively as Round One, Round Two and Round Three.

Participants

The study sample for the research study was comprised of 84 participants, working in 12 teams composed of 7 members each, lead through advanced design exercises by six master instructors. Participants were graduate students enrolled in architecture, urban design, or interior design Master programs at a university in the Midwest region of the United States. All participants had a minimum of 4 years formal design education. The minimum age of participants was 21, with the maximum age being 48; mean age was 28. The majority of the population was ethnically White or Hispanic.
(86%), with minority populations of South Asian (4%), East Asian (2%), and Arabic (8%). Gender was fairly balanced across all ethnicities, with the overall sample containing a higher population of males (63%) as compared to females (37%).

Teams
All participants completed the Basadur Creative Problem Solving Profile (CPSP) Inventory (Basadur & Head, 2001), and the Keirsey Temperment Sorter variation of the Myers-Briggs Type Indicator (Keirsey, 1998), prior to the beginning of the research study. Results of the Basadur CPSP Inventory were used to determine each participant’s putative cognitive style in the creative process. The study required the 12 teams to be assembled by diversity in cognitive style in order to create strong heterogeneity and implying low cohesion in the teams. Heterogeneity in cognitive styles appears to have a positive effect on team performance in creative problem solving tasks (Basadur & Head, 2001). The study engaged cohesiveness as a critical factor of increasing team success when membership was based in heterogeneity.

The composition of the study sample according to the four quadrants of the Basadur CPSP Inventory was implementer (45%), followed by conceptualizer (21%), generator (19%), and optimizer (14%). The composition of each team was made up of at least one member from each of the four Basadur quadrants, with the remaining 3 members of the team being randomly assigned for maximum variation. Gender was purposely not a feature of the research, and there was no attempt to select for equal balance of male and female participants on the teams. However, other participant characteristics (such as home location, distance from the university, and work schedule) were used to select team membership so long as they did not impact the cognitive style balance. Teams had consistent membership throughout the research study.

This project received ethical approval to conduct research with human participants from the Institutional Review Board of the University.

Procedure for Judging Team Work
Team work from the research study was evaluated by raters who utilized the TIOSE measure to judge team work quality and success. A graphic rating mechanism, rather than a numeric scale, was chosen for the measure’s scoring worksheet. This graphic approach allowed the rater to feel more natural and instinctual in their identification of success and failure as it did not ask them to translate a qualitative factor into a precise number. The gradient scale allowed a visual identification of a centreline of “meeting basic requirement” as a visual node. Three notations are identified on the grading line: “exceptional” at the left terminus, “decent” at the centre point, and “poor” at the right terminus. The rater judged the quality of the factor based on distance away from the centreline. Rating for projects identified as surpassing expectations or not meeting basic requirements were notations further to the left (or right) of each factors gradient.

Rating of team work was completed during a group judging session in which each team’s project was presented as a hard-copy poster (7 panels of 24”x36”). All presentations occurred in the same presentation/critique space to maintain stable and consistent
review conditions for the raters (Flemenbaum & Zimmermann, 1973). Each team initially had 10 minutes to orally present the important qualities of the work. As needed, presentations were extended to a hard cap of 30 minutes in order to ensure comprehension of important points and clarification of any deficiencies between intentions and results. Raters were instructed to use the hard-copy poster for grading, and use the oral presentation for orientation only. Judging occurred in a semi-blind process in that none of the raters had access to the worksheet of other raters. Worksheets were submitted to the study’s Principal Investigator after each judging period, and raters were not allowed access to worksheets, nor were they given the opportunity to change ratings once submitted.

**Raters**

Six raters (critics/judges) performed the judging of team work quality and team success at Round One, Round Two, and Round Three. Raters consisted of 4 males and 2 females who were considered to have reached, or were above, a level of mastery in design. All raters had a terminal degree in their sub-field of design (architecture, urban), and all acted in accordance with the definition of an “expert performer” involved in deliberate practice with continued and sustained training and effort (Ericsson, 2006). The six raters had an average of 18 years of continuous practice experience in the study of design (range 10-25 years).

Raters were introduced to the TIOSE measure and the scoring worksheet during a one-hour training session which occurred before the first rating period. Each of the factors was presented for definition and examples of successes and failures for each factor were provided. Discussion occurred after each factor presentation, as well as a general discussion of the overall measure at the conclusion of the session. The raters had minimum questions and the factors were agreed upon as being understood and instinctual by the raters.

**Data Analysis**

All raters evaluated the five TIOSE factors for each of the 12 team projects on Rounds One, Two, and Three. After raters completed scoring, ratings were entered into a spread sheet. A grading key translated the graphic marks on the measure’s worksheet into a 100 point scale. The conversion occurred after all rating was completed and independent to the rater or any inter-rater discussion. Importantly, both groups of factors and the corresponding rating process are for project results only, i.e., the visual representation which is the mode that architecture communicates the intentions of the proposal. Other factors, such as individual, team interaction and development/learning ratings, are not addressed by the TIOSE measure. Consequently, all data was analysed (no missing data). Submitted rating values for the five categories were combined into an overall Project Success category. Data was analysed using Minitab 16.1 statistical analysis software to determine the intraclass correlation (ICC) index of the reliability of the TIOSE.

**Results**

ICCs and mean summary ratings between raters for each round are presented in Tables 1-3; ICCs and mean ratings between- and within-raters across all three rounds are presented in Table 4. For all results, ICC values 0.75 or higher reflect excellent agreement within- and between-raters of the same team project.
As shown in Table 1, excellent agreement in the individual team ratings between raters on Round One was found for 4 out of 5 categories and the composite Project Success construct (ICC values range from 0.78-0.85). The between-rater agreement on Round One for Evocativeness was 0.70, which is slightly below the criterion cut-off for excellent agreement. In contrast to the excellent agreement between raters on Round One, agreement between raters was moderate on Round Two and Round Three (see Table 2 and Table 3, respectively), with Round Two containing the lowest ICCs across all three rounds. Specifically, ICCs on the TIOSE rating categories during rounds Two and Three ranged from 0.47-0.62, and 0.64-0.79, respectively. ICCs for the rounds Two and Three composite category Project Success were 0.58 and 0.73, respectively.

ANOVA found significant differences in the Round One mean ratings between raters 2 and 5 on the rating categories Thoroughness and Informativeness. Significant differences in the Round Two mean ratings were found between raters 5 and 6 on the rating categories Informativeness and Organisation, and between raters 1-2 and 5-6 on Evocativeness. In Round Three, significant differences in mean ratings were found between raters 1-3 and 2-5 on the rating categories Organisation and Synthesis, and between raters 1-2 and 5-6 on the rating category Evocativeness. Taken together, these ANOVA results support ranking of ICCs on the Project Success composite rating category across the three rounds: Round One (ICC = 0.79) > Round Three (ICC = 0.73) > Round Two (ICC = 0.58).

<table>
<thead>
<tr>
<th>Rating Categories</th>
<th>ICC (95% CI)</th>
<th>Total</th>
<th>Rater1</th>
<th>Rater2</th>
<th>Rater3</th>
<th>Rater4</th>
<th>Rater5</th>
<th>Rater6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thoroughness</td>
<td>.85</td>
<td>83.96</td>
<td>86.00</td>
<td>89.75</td>
<td>81.92</td>
<td>81.50</td>
<td>78.75</td>
<td>85.83</td>
</tr>
<tr>
<td></td>
<td>(.65-.95)</td>
<td>(1.06)</td>
<td>(1.54)</td>
<td>(2.74)</td>
<td>(3.40)</td>
<td>(2.76)</td>
<td>(1.83)</td>
<td>(2.03)</td>
</tr>
<tr>
<td>Informativeness</td>
<td>.80</td>
<td>85.06</td>
<td>86.42</td>
<td>91.33</td>
<td>84.83</td>
<td>82.08</td>
<td>78.42</td>
<td>87.25</td>
</tr>
<tr>
<td></td>
<td>(.55-.93)</td>
<td>(1.08)</td>
<td>(1.90)</td>
<td>(2.75)</td>
<td>(3.18)</td>
<td>(2.42)</td>
<td>(2.08)</td>
<td>(2.13)</td>
</tr>
<tr>
<td>Organisation</td>
<td>.78</td>
<td>85.38</td>
<td>89.58</td>
<td>87.00</td>
<td>85.46</td>
<td>84.00</td>
<td>80.67</td>
<td>85.58</td>
</tr>
<tr>
<td></td>
<td>(.52-.92)</td>
<td>(1.95)</td>
<td>(1.84)</td>
<td>(2.63)</td>
<td>(3.18)</td>
<td>(2.42)</td>
<td>(2.08)</td>
<td>(2.13)</td>
</tr>
<tr>
<td>Synthesis</td>
<td>.83</td>
<td>82.33</td>
<td>85.33</td>
<td>83.42</td>
<td>83.96</td>
<td>79.54</td>
<td>79.00</td>
<td>82.75</td>
</tr>
<tr>
<td></td>
<td>(.62-.94)</td>
<td>(1.06)</td>
<td>(2.03)</td>
<td>(2.66)</td>
<td>(3.43)</td>
<td>(2.87)</td>
<td>(1.95)</td>
<td>(2.29)</td>
</tr>
<tr>
<td>Evocativeness</td>
<td>.70</td>
<td>83.61</td>
<td>89.42</td>
<td>82.42</td>
<td>84.92</td>
<td>82.50</td>
<td>78.67</td>
<td>83.75</td>
</tr>
<tr>
<td></td>
<td>(.36-.90)</td>
<td>(1.03)</td>
<td>(1.17)</td>
<td>(3.11)</td>
<td>(3.48)</td>
<td>(1.99)</td>
<td>(1.64)</td>
<td>(2.42)</td>
</tr>
<tr>
<td>Project Success</td>
<td>.79</td>
<td>84.07</td>
<td>87.35</td>
<td>86.78</td>
<td>84.22</td>
<td>81.93</td>
<td>79.10</td>
<td>85.03</td>
</tr>
<tr>
<td></td>
<td>(.68-.86)</td>
<td>(1.46)</td>
<td>(1.28)</td>
<td>(1.39)</td>
<td>(1.10)</td>
<td>(0.81)</td>
<td>(0.99)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Inter-rater Reliability and Mean Ratings on Round One.

Intraclass correlations (ICC) for inter-rater reliability derived from a 2-way, random effects ANOVA model according to the procedure of Shrout and Fleiss (1979); 95% confidence interval (CI) in parentheses. Mean ratings of each category for Total and individual Raters on round; standard error of the mean (SEM) in parentheses. *p < .05 **p < .01 significant difference between ratings according to ANOVA.
Table 4 presents both inter-rater reliability and intra-rater reliability across all three rounds. The lowest ICC was, surprisingly, found in Organisation while the highest was in Thoroughness. Out of the three base-line categories, Organisation is the more difficult to assess as it relates to issues of clarity on the part of proposal but also the rater’s ability in pattern recognition, often across multiple scales. Thoroughness, consistently high in agreement between raters, can be defined as a fairly quantitative category and had excellent agreement when compared across
all three rounds. Overall, there was consistency between the raters on all categories from less tacit to more tacit, with a range from 0.68 to 0.80 ICC.

Table 4: Inter-rater and Intra-rater Reliability and Mean Ratings across All Three Rounds.

<table>
<thead>
<tr>
<th>Rating Categories</th>
<th>ICC a (95% CI)</th>
<th>ICC b (95% CI)</th>
<th>Total Ratings</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
<th>Rater 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thoroughness</td>
<td>.80 (.68-.89)</td>
<td>.73 (.47-.90)</td>
<td>87.60</td>
<td>88.61</td>
<td>91.78</td>
<td>88.22</td>
<td>85.53</td>
<td>85.67</td>
<td>85.81</td>
</tr>
<tr>
<td>Informativeness</td>
<td>.71 (.53-.83)</td>
<td>.84 (.67-.94)</td>
<td>86.40</td>
<td>88.56</td>
<td>90.11</td>
<td>87.97</td>
<td>85.33</td>
<td>80.33</td>
<td>86.11</td>
</tr>
<tr>
<td>Organization</td>
<td>.68 (.49-.82)</td>
<td>.77 (.55-.92)</td>
<td>87.10</td>
<td>91.64</td>
<td>88.22</td>
<td>88.51</td>
<td>86.22</td>
<td>81.33</td>
<td>86.64</td>
</tr>
<tr>
<td>Synthesis</td>
<td>.73 (.58-.85)</td>
<td>.77 (.53-.92)</td>
<td>83.72</td>
<td>88.17</td>
<td>84.58</td>
<td>83.38</td>
<td>83.96</td>
<td>79.39</td>
<td>82.86</td>
</tr>
<tr>
<td>Evocativeness</td>
<td>.72 (.54-.84)</td>
<td>.71 (.45-.90)</td>
<td>86.46</td>
<td>91.72</td>
<td>89.92</td>
<td>86.33</td>
<td>85.89</td>
<td>81.44</td>
<td>83.44</td>
</tr>
<tr>
<td>Project Success</td>
<td>.74 (.67-.80)</td>
<td>.78 (.69-.85)</td>
<td>86.26</td>
<td>89.74</td>
<td>89.82</td>
<td>86.88</td>
<td>85.39</td>
<td>81.63</td>
<td>84.97</td>
</tr>
</tbody>
</table>

Table 4: Inter-rater and Intra-rater Reliability and Mean Ratings across All Three Rounds.
aICCs for inter-rater reliability across all rounds. bICCs for intra-rater reliability across all rounds. cMean ratings of each category for Total and individual Raters across all rounds; standard error of the mean (SEM) in parentheses. *p < .05 **p < .01 significant difference within ratings according to ANOVA.

Discussion

Summary and Implications
In summary, when looking at intra-rater reliability for the measure, the agreement within raters was excellent for 3 of the 5 rating categories and the composite Project Success construct. The coefficient was the highest for the categories Informativeness, Organisation, and Synthesis. The lowest ICC was found for the category Evocativeness, suggesting that this category was the most difficult for each rater to judge. The fact that Synthesis rated with excellent agreement speaks to the ability to have consensus, without prior discussion, on the core nature of design as a qualitative, tacit and systems-based composition.

The agreement between raters was highest on the first round (Table 1), dipped for the second round (Table 2), and then came back up for the third round (Table 3), with all of the Round Two ICCs for inter-rater reliability within the moderate agreement range of 0.40-0.75. Since one of the needs for the TIOSE measure was to handle “real-world” variation in project presentation, there was a variation of requirement between the delivered project proposals between rounds one-three. Round Two contained a shift of priorities and complexity for the teams, with greater emphasis placed on Synthesis and Evocativeness, thereby requiring evidence of a deep understanding of the topic. This finding seems to be reflected in the range of rater’s assessment, and implies that Thoroughness,
Informativeness and Organisation are not as well defined or agreed upon when there is an increase in complex and system-based information. We also recognized an issue that raters, in judging design work, may be biased by projecting their own interpretation onto a project, creating a dimension of possibility that doesn’t actually exist in the work itself. The rating on the measure then relates to the rater’s view of the project’s potential rather than objectively rating the project itself. The differences in ratings seen between Round 2 and Round 3 may be related to raters recognizing this intrinsic bias for self-evaluation.

As mentioned above, agreement between the six raters on Round Three was higher than on Round Two, with 2 out of 5 categories on Round Three satisfying the inter-rater reliability criterion for excellent agreement. Again, it was the composite and more highly tacit categories of Synthesis and Evocativeness which had lower ICCs than the baseline categories. The categories of Thoroughness, Informativeness and Organisation followed the same pattern of ratings in Round Three as in Round One, with Thoroughness rating the highest followed by a decrease in each subsequent category. There was more agreement on Evocativeness in Round Three than the previous rounds, although it didn’t quite reach the criterion for excellent agreement.

Overall, the reliability of the TIOSE measure was excellent since the Project Success ICC between- and within-raters across all three rounds of project evaluations was excellent (mean = 0.76, see Table 4). The decrease in reliability in Round Two (Table 2) possibly represents change in project focus and an increase in the use of tacit knowledge, as well as more advanced design integration skills. As the study progressed, personal interaction between the raters and the teams increased in a constant manner. Additionally, there was less tacit knowledge available early in the study, and Round Two judging may have therefore been confounded by the personal interactions with the teams, thereby increasing personal bias amongst the raters. It was observed that in the final round (Round Three), the raters appeared to be more objective, and remove any personal bias. Thus, Round Two appeared to be instrumental in ‘training’ the teams as to the expectations of the raters, as well as preparing the raters to withdraw personal claims on the presented work. In terms of the teams, expectations of success in design were not made available as explicit instructions, but were implied through the discussion and question periods of the project development. As the raters acquired experience with each team, and the teams acquired more experience to the overall expectations of success, reliability increased.

Lower ICCs reflect less agreement between raters, less ability to access the knowledge held in the TIOSE categories and address the distinction between tacit versus explicit knowledge. For example, lower ICCs were obtained for those categories which are traditionally considered as systems-based and more difficult to define (i.e. Synthesis, Evocativeness), compared to the higher ICCs for the more explicit and quantitative categories (Thoroughness, Informativeness, Organisation). In terms of the raters, higher ICCs are indicative that there was more agreement amongst held tacit knowledge (i.e., the raters had more access to that knowledge), and
more clarity in the category.

Conclusion

The TIOSE measure, overall, showed excellent agreement within and between raters to depict agreement of success in architectural design projects. A key feature of the research was increased reliability when raters appeared to have access to deep, disciplinary knowledge (expert performer). Consequently, increases in variation of the experience of the rater group will contribute to an increase in the reliability of the measure. Key factors identified in the study were inter-personal interaction with the students/teams to be rated, and the experiential level of the rater.

There are several concerns and suggestions to consider for subsequent applications of the TIOSE measure. First, training will be necessarily minimal since the measure identifies currently accepted implicit practices of judgement to make them explicit, and does not focus on the construction of new factors. Training is meant for orientation and cohesiveness of judgement content for reliability factors in a tacit environment, rather than for educating the rater on unknown procedures or specific aspects of content. Training is also meant for general discussions to occur between raters so that a consensus of operational definitions is determined.

In addition, the blind nature of the rating needs to be maintained. In later applications, multiple events of bias and influence occurred between raters when the blind and discrete nature of the rating was not maintained. This interpersonal bias lead to the rating to normalized around the strongest personality of the raters.

The rating can be, and has been, used as a teaching tool to architectural design students. Discussing the factors of judgement of design allows for transparency of the judgement and grading process. Students understood their responsibilities clearer and were able to address factors both as combined and singular issues. This allowed the instructor to clearly articulate concerns as well as strengths and weaknesses of the final proposal and the design process.

Future Research

In the present study, the TIOSE measure was tested for reliability in terms of the ability of the tool to yield agreement between- and within-raters. Future research will investigate the validity of the tool by comparing it to similar tools (convergent validity). For example, the TIOSE will be examined in relation to the Jackson and Messick (1964) criteria for assessment of the creativity of a product (in terms of unusualness, appropriateness, transformation and condensation), and the creative quality of team conceptualization and optimization.

Future research will also address the issue of rater competence. Specifically, if the rater is not at mastery level, there is concern that identification of the judgement factors in design proposals may lack clarity. Thus, the impact of training and its effect on competency utilizing the TIOSE measure will be investigated.

As noted, teams in the present study were formed according to each member’s presumed cognitive style in the creative process. Cognitive styles influence not only the creative problem
solving process (Cools & Van Den Broeck, 2007), but also team cohesion and team performance (Bradley & Herbert, 1997). Since architecture maintains little to no research in the effect of team structure, cognitive style, and team interaction in the success of design outcomes (Hill, Dong, & Agogino, 2002), future research will use the TIOSE measure to investigate how diverse cognitive thinking styles engage in various phases of the design process, and how tacit knowledge is transferred to complete a successful design outcome. Team member characteristics, such as gender and personality, will also be explored for their impact on team cohesion and quality design outcome.

In summary, this study demonstrates that qualitative measures of project success can be quantified in a reliable manner and can attain internalized agreement between architectural critics. The TIOSE measure adds to the arsenal of available rating scales for the qualitative measure of architectural design quality.

References


inferences about some intraclass correlation coefficients. Psychology Methods, 1, 30-46.


Acknowledgement

The authors would like to acknowledge the support of the College of Architecture and Design at Lawrence Tech (Southfield, USA) as well as the comments and discussions extended by Mason White, Ralph Nelson, Constance Bodurow, Beverley Geltner, Margaret Wong, and Aaron Jones.

Philip Plowright

Philip Plowright is an Associate Professor of Architectural Design and History/Theory at Lawrence Technological University in Southfield, Michigan, USA; a registered architect in Michigan; a founding member of the synchRG research group; principal of the design research practice rlab.a; and a director for the Architectural Research Centers Consortium (ARCC). He holds degrees in studio art, art history and architecture from the University of Guelph and the University of British Columbia, Canada. His research is focused around issues of meaning in design and includes explorations in architectural theory, design methodology, cognitive science and judgment/decision theory. The shared interest in these knowledge categories serves notions of quality in architecture and the application of conceptual frameworks into application. His upcoming book, Architectural Design Methodologies: Applying Thinking Frameworks to Practice Strategies, will be published by Routledge in early 2014. He can be contacted at pplowrigh@ltu.edu

Matthew L Cole

Matthew Cole is an Assistant Professor with the College of Management, Lawrence Technological University, Southfield, MI, USA. Dr. Cole teaches Business Statistics, Biostatistics, Research Design-Quantitative Methods, and Research Methods for the Behavioral Sciences. He received his PhD from Wayne State University (Detroit, MI) in Cognitive and Social Psychology across the Lifespan, where he conducted research on growth modeling of risk behaviors and perceptions among adolescents from the U.S. and Vietnam. Dr. Cole holds an MA degree in Biopsychology from Wayne State University, and an MS degree in Clinical Behavioral Psychology from Eastern Michigan University. His research interests focus on the science of teams, cognitive styles in problem solving, entrepreneurship, and research methodology. He can be contacted at mcole@ltu.edu