Empirical Results from an Experiment on Value-Based Review (VBR) Processes

Keun Lee Barry Boehm
Center for Software Engineering
University of Southern California
Los Angeles, CA 90089-0781
(keunlee, boehm)@usc.edu

Abstract

As part of our research on value-based software engineering, we conducted an experiment on the use of value-based review (VBR) processes. We developed a set of VBR checklists with issues ranked by success-criticality, and a set of VBR processes prioritized by issue criticality and stakeholder-negotiated product capability priorities. The experiment involved 28 independent verification and validation (IV&V) subjects (full-time working professionals taking a distance learning course) reviewing specifications produced by 18 real-client, full-time student e-services projects. The IV&V subjects were randomly assigned to use either the VBR approach or our previous value-neutral checklist-based reading (CBR) approach. The difference between groups was not statistically significant for number of issues reported, but was statistically significant for number of issues per review hour, total issue impact, and cost effectiveness in terms of total issue impact per review hour. For the latter, the VBRs were roughly twice as cost-effective as the CBRs.

1. Introduction

Finding more cost-effective techniques for achieving software quality is a major issue for developers. The review is one of the main processes to find defects from the initial development stage and increase the quality of the deliveries [1,2,3]. Peer reviews have been used in requirements analysis, architecture, design and coding. Many research efforts have focused on formulating effective review processes to find defects [4,5]. Studies have addressed review team composition and procedures, review preparation and duration, and criteria for focusing reviewers on sources of defects.

Initial approaches for focusing reviewers involved adding checklists into the review process [2]. With the checklist-based reviewing (CBR), it is easy to understand the process so that CBR has become the most common review-focusing techniques in current practice.

Another approach to reviewing artifacts is perspective-based reading (PBR). It focuses on different reviewers’ perspectives, like designer perspective and tester perspective [8]. Different review perspectives help to find more defects without overlaps. Another reading technique is defect-based reading, which is focusing on different defect classes [6]. Other review techniques proposed are functionality-based reading (FBR) [7], and usage-based reading (UBR) [9,10]. A number of studies have compared the effectiveness of these techniques [4,5,7,11,12]. They agree in finding that focused review methods do better than unfocused reviews; that methods’ cost effectiveness vary by the nature of the artifacts being reviewed; and generally that the preferred method to use is situation-dependent. However, the cost-effectiveness metrics used for the methods and their evaluation have been (except [10]) value-neutral, in that each defect is considered to be equally important. It means much effort is spent on trivial issues like obvious typos and grammar errors.

Here we present a new form of PBR called value-based reading, and compare the use of VBR using value-based procedures and checklists with a value-neutral CBR approach. The two approaches were performed by 28 randomly-selected professional software engineers taking a graduate software engineering project course by distance learning at USC. Their project assignment was to independently verify and validate (IV&V) the artifacts from one of 18 real-client e-services project applications that were being developed by the on-campus MS-student teams in the course[14].
Value-based review techniques and cost effectiveness metrics are explained in the next section. The experiment on Value-based review and traditional review with checklists is described in section 3. Results of the experiment results are presented in section 4. Threats to validity of the experiment are discussed in section 5. The discussions of the results and the conclusions are in section 6.

2. Value-based Verification & Validation (VBV&V)

2.1. Concepts in VBV&V

The basic idea of VBV&V, as with other value-based software engineering (VBSE) activities, is to treat each V&V activity (analysis, review, test) as a candidate investment in improving the software development process[16]. Earlier VBSE activities involve prioritizing the capabilities to be developed. VBV&V activities are then sequenced by priority and criticality. Priority is determined from negotiations and meetings with clients. In the experiment, the values of priority are High, Medium, Low (or 3, 2, 1). The criticality is how critical a review issue is to the project’s success. Generally, the values of criticalities are furnished by experts, but reviewers can also determine the values in special circumstances. The values of criticality in the experiment are also High, Medium, Low (or 3, 2, 1). The numerical values of priority and criticality are used to guide V&V activities and evaluate their effectiveness. Whereas the priority represents stakeholders’ “win” conditions, the criticality is impacted more by domain-experts or developers. For example, if one of capabilities is not important to users, its priority can be determined as “low” through clients’ meeting. However, if the capability is related to other requirements and its failure impacts seriously to other capabilities, its criticality can be determined as “high” by domain-experts or developers. The priority and criticality determine a value of each issue.

Effectiveness is based on a value of each issue. The assumption of Value-based review is that each issue has different value, and if higher value issues are reviewed and fixed first, then the effectiveness of review will be increased. Thus, effectiveness can be measured based on the value of each issue. Section 2.3 will explain the process of calculating the effectiveness.

2.2 Value-based Review

We have developed an experimental set of value-based checklists for reviewing specifications to test the hypothesis that review activities will be more cost-effective if review effort is focused on the higher-priority system capabilities and the higher-criticality sources of implementation risk [13]. Our spiral approach to systems and software engineering emphasizes risk-driven activities, but our previous review guidelines made no distinctions in focusing limited reviewing resources on high vs. low priority and criticality review artifacts.

The revised approach to overcome the weaknesses in the previous approach includes the priority of system capabilities and criticality of sources of risk. Low-priority, low-criticality items are optional to review based on time availability.

The basic process for the Value-based review can be divided into four steps.

1. The system capabilities are classified based on their priorities
   First the reviewers need to determine which system capability will be reviewed first. A negotiation among stakeholders defines the priority of each system capability.

2. The high-priority system capabilities will be reviewed first.

3. At each priority level, high-criticality sources of risk will be reviewed first, basically going down the columns in Figure 2-1 below.

4. After reviewing the issues with higher criticality, the reviewers can address the next lower criticality issues, as time is available. As shown in Figure 2-1, it is optional to go all the way down the columns for the lower-priority artifacts.

Figure 2-1 Review effectiveness metric; Issue metrics and optionality guidelines

\[
\text{Effectiveness Metric} = \sum_{\text{issues}} (\text{Artifact Priority}) \times (\text{Issue Criticality})
\]

<table>
<thead>
<tr>
<th>Artifact Priority</th>
<th>H</th>
<th>M</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criticality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>9</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>M</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>L</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Generally considered optional to review
The overall review effectiveness metric is the sum of priority and criticality value. For example, if one issue has medium priority and high criticality, the impact of the issue is six from two (medium) times three (high). The overall review effectiveness metric is the sum of all the issue impacts.

### 2.3 Weight of review issues

The main goal of the Value-based review is to increase the cost effectiveness of the review. Review effectiveness metrics involve weighted sums of distinct issues reported, using the impact metrics in Figure 2-1. Each issue reported has a priority value and criticality value. The impact of each issue is the product of its priority and criticality value. For example, if one issue has medium priority and high criticality, the impact of the issue is six from two (medium) times three (high). The overall review effectiveness metric is the sum of all the issue impacts.

### 2.4 Value-Based Checklist

There are two checklists for VBR. The first one is a general criticality-oriented checklist shown in Figure 2-2. The general checklist covers general issues for reviewing any set of artifacts for completeness, consistency, feasibility, ambiguity, conformance, and risk. Each category has three levels of issue criticality: high, medium, and low. The reviewers involved in the experiment are asked to keep the general Value-based checklist in mind in a review.

The set of value-based checklists also includes artifact-oriented checklists. Since the experiment had been performed using Model-based (system) Architecting and software Engineering (MBASE), the MBASE-oriented checklist is provided to the reviewers for reviewing the various sections of the following MBASE artifacts: Operational Concept.
Description (OCD), Requirements Description (SSRD), and Architecture Description (SSAD) specifications.

3. Experiment Description

3.1 CSCI 577A course and IV&V

The CSCI 577A course is the first half of the USC Software Engineering project course, which focuses on software plans, processes, requirements and architectures. This course covers the application of software engineering process models and management approaches for developing plans, requirements and architecture of large software systems. Students work in teams and apply the Win-Win spiral model [14] and Model Based System Architecting and Software Engineering (MBASE) guidelines for software engineering of real-client projects[15].

Previously, the review guidelines in CSCI 577 have been value-neutral. Every artifact and issue was equally important and was reviewed from the first to the end with same weight.

The remote students (off-campus students) in the course are responsible for "Independent Verification and Validation" (IV & V). V & V is a collection of analysis and testing activities across the full life cycle and complements the efforts of other quality-engineering functions. It determines that the software performs its intended functions correctly, performs no unintended functions and measures the quality and reliability of the software.

![Figure 3-1. Independent Verification and Validation in CSCI 577A](image)

Experiment, the off-campus IV&Vers were divided into two groups, A and B. The randomly-selected Group A performed verification and validation with the Value-based checklists (VBR in Figure 3-1), discussed in Section 2. Group B used the traditional CS 577 value-neutral checklists (CBR). Both groups had otherwise-equal education on MBASE and IV&V processes through the lectures and the assignments. If the IV&Vers find any issue, they register potential Problems as Concerns on the Concern logs. The IV&Vers’ Concern logs are sent back to the developers and collected for data analysis. The author determines which Concerns need fixing and registers them as Problems. The author was then asked to submit a Problem list that includes identified Concerns as Problems. In our context, the term “Problem” carries the general meaning of “defect”. We use the generic term “Issue” to cover both “Concerns” and “Problems”.

The Concerns logs and Problems lists used were basic forms to collect data for analysis. All collected data was vetted, and if the review processes or forms were applied incorrectly, the IV&Vers were asked to resubmit the data.

3.2 Experiment Design

3.2.1 Materials

The Value-based checklists used in the experiment covered three documents in MBASE which are directly related to the product development.

- Operational Concept Description (OCD)
  The OCD describes how the proposed system operates. It gives a general understanding of the proposed system and its operational concepts to stakeholders. It defines the objectives and scope of the proposed system, and describes the system’s key operational stakeholders and scenarios, along with prototype summaries.

- System and Software Requirement Description (SSRD)
  The SSRD describes all requirements of the project. It includes capability requirements, Level of service requirements and global constraints (Interface requirements, Project requirements)

- System and Software Architecture Description (SSAD)
  The SSAD describes the results of analyzing the requirements and concept of operation for the system, designing architecture, and designing an implementation. The SSAD serves as a development blueprint for the Construction phase.
3.2.2. Variables

• Independent variables
Two different review techniques, CBR and VBR, are used in the experiment. The VBR is assigned to group A and the traditional value-neutral CBR is assigned to group B.

• Dependent variables
Number of issues: The number of concerns and problems found by IV&Vers are compared in both groups.
Impact of issues: using the effectiveness metric, the effectiveness of each issue is calculated (figure 2-2). The total Impacts and sum of each issue effectiveness are compared in both groups.
Effort: generally, the cost of performing a software review includes the individual preparation of each reviewer before the review and the effort of reviewers during the review activity. The total effort IV&Vers spent is the sum of the preparation time and review time. Preparation for review includes understanding background of the artifacts, understanding the checklist to use in the review, communication with the development team related to the review and other efforts to prepare the review.
Cost-effectiveness: Numbers or impact of issues found per hour of effort.

3.3 Hypotheses

We defined the hypotheses for the experiment as follows.

Overall Hypothesis: The mean numbers, rates, impacts and cost-effectiveness of concerns and problems will not differ between the Value based review group and the traditional review group.

Individual Hypotheses:

1. The number of concerns IV&Vers found does not differ between groups.
2. The number of problems IV&Vers found does not differ between groups.
3. The Impact of concerns in both groups does not differ between groups.
4. The Impact of problems in both groups does not differ between groups.
5. The number of concerns IV&Vers found per hour does not differ between groups.
6. The number of problems IV&Vers found per hour does not differ between groups.
7. The cost effectiveness (impact per hour) of concerns does not differ between groups.
8. The cost effectiveness (impact per hour) of problems does not differ between groups.

Figure 4-1. Number of Concerns and Problems by IV&Ver
4. Experiment Results

The total number of IV&Vers involved in the experiment was 29, but one did not follow the MBASE guidelines. Thus the number of samples collected was 28: 15 in Group A and 13 in Group B. The total number of concerns IV&Vers found is 4,641 and the number of problems is 2,765. The average number of concerns IV&Vers found is 165.75 (Group A 189.13 and Group B 138.77) and the average number of problems is 98.75 (Group A 111.80, Group B 83.69). The average preparation time for review is 23.36 hours (Group A 19.57 hrs, Group B 27.73 hrs) and average review time is 43.29 hours (Group A 40.97 hrs, Group B 45.97 hrs).

Basically the results show that group A reviewers tended to find more concerns and problems than group B. (Figure 4-1). Figure 4-2 shows group A subjects tended to have a higher impact on concerns and problems, but again not uniformly across all the subjects in both groups. Below we will test for statistical significance of the results.

4.1 Comparative Effectiveness

Table 4-3 shows the comparative average numbers of concerns and problems found by IV&Vers in Groups A and B. The Group A mean is 34% higher for number of concerns, and 51% higher for number of problems. It looks like the IV&Vers in group A had found a higher number of concerns and problems than group B. But the t-test p-values are 0.202 for number of concerns and 0.056 for number of problems. The hypothesis that there is no difference between the two groups for number of concerns and problems cannot be rejected at the 0.05 level.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Group</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>% Group A higher</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerns</td>
<td>A(VBR)</td>
<td>178.15</td>
<td>98.89</td>
<td>34%</td>
<td>0.202</td>
</tr>
<tr>
<td></td>
<td>B(VNR)</td>
<td>133</td>
<td>60.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problems</td>
<td>A(VBR)</td>
<td>106.46</td>
<td>52.79</td>
<td>51%</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>B(VNR)</td>
<td>70.73</td>
<td>27.82</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4-4. Number of concerns and problems found by IV&Vers in the both groups

Figure 4-4 explains more clearly the t-test results. It shows numbers of concerns and problems found by IV&Vers with boxplots. In the figure, the maximum, minimum and mean of the two groups are shown and the box in the middle shows data range of 75% of total data in the two groups. Because of some anomalies with the data reporting, we have dropped the highest and lowest data points in each group. It looks like the numbers of concerns and problems found by IV&Vers in Groups A and B have a good deal of overlap, although the Group A yields are uniformly higher. Fifteen data points (9 in group A, 6 in group B) report the numbers of concerns the IV&Vers found are in the range 100 to 250, which is 62.5% of the total data (15 data of 24 total data). For problems, 16 data points (8 in group A, 8 in group B) show the number of problems the IV&Vers found to be in the range 50 to 150. This is 66.7% (16 data of 24 total data). More than 60% of the data points are overlapped in the middle range for concerns and problems, making it difficult to reject the equal-means hypotheses.

Table 4-5. Comparison of Group A and Group B: Mean Impact based on Concerns and Problems

<table>
<thead>
<tr>
<th>Impact</th>
<th>Group</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>% Group A higher</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerns</td>
<td>A(VBR)</td>
<td>992.23</td>
<td>552.58</td>
<td>65%</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>B(VNR)</td>
<td>601.91</td>
<td>302.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problems</td>
<td>A(VBR)</td>
<td>604.92</td>
<td>334.43</td>
<td>89%</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>B(VNR)</td>
<td>319.55</td>
<td>133.27</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4-5 shows the comparative effectiveness in terms of concern and problem impacts found by IV&Vers in Groups A and B. The difference between Group A and Group B means is considerably higher for impact of concerns and problems than it was for numbers of concerns and problems. For concerns, Group A’s mean is 65% higher on impact vs. 34% on number. For problems, Group A is 89% higher on impact vs. 51% on number. Also the results of the t-test of Impact of Concerns and Problems show that two groups are different.

The p-values are 0.049 for Impact of Concerns and 0.012 for Impact of Problems. This result is enough to reject the hypothesis that there is no difference in the mean impact of concerns and problems between the two groups.

Figure 4-6. Impacts of concerns and problems in the both groups

Figure 4-6 shows the individual impacts of the concerns and problems in both groups of boxplots. It can be seen that the differences between the two groups are much greater than they were for a number of concerns and problems in figure 4-5. For the impact of concerns, the number of data in group A over mean impact (992.23) is 5 (of 13 data, 38.5%), while the number of data in group B over the group A’s mean impact is 1. Most data in group B are placed in the range 200 to 900 (8 of 11 data, 72.7%) which is less than group A’s mean impact. For the impact of problems, the number of data in group A over the average impact of group B (319.55) is 10 of 13 in group A, 76.9%.

The differences in impact were also clear from examining the review results. Group B reported many more low-criticality issues (e.g. non-misleading typos and grammatical errors in figure 2-1) than Group A.
4.2 Comparative Cost-Effectiveness

Table 4-7 shows the comparative average numbers of concerns per hour and problems per hour found by IV&Vers in Group A and B. “Hour” is measured as a personnel effort hour spent for review. The differences are clearly higher than the differences in mean numbers of concerns and problems found. For concerns, Group A’s mean is 51% higher for concerns/hour as compared to 34% higher for number of concerns. For problems, Group A’s mean is 61% higher for problems/hour as compared to 51% for number of problems. The differences between group A and group B are proved statistically with low p-values (0.026 for concerns/hour, 0.023 for problems/hour).

<table>
<thead>
<tr>
<th>Issue per hour</th>
<th>Group</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>% Group A higher</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerns</td>
<td>A(VBR)</td>
<td>2.9</td>
<td>1.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B(VNR)</td>
<td>1.88</td>
<td>0.77</td>
<td>55%</td>
<td>0.026</td>
</tr>
<tr>
<td>Problems</td>
<td>A(VBR)</td>
<td>1.78</td>
<td>0.85</td>
<td>61%</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>B(VNR)</td>
<td>1.11</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4-8. Number of concerns and problems found by IV&Vers per hour in both groups

In figure 4-8, boxplots of the number of concerns and problems found by IV&Vers per hour are shown. The figure also shows that the IV&Vers in group A had found a larger number of concerns and problems per hour. 11 of 13, 84.6%, in group A are higher than the average of group B,1.88, for the number of concerns the IV&Vers found per hour. For problems per hour, the number of data points in group A over the data points with maximum value (1.78) in group B is 7 (of 13 data, 53.8%).

Table 4-9 compares the mean impact per hour for concerns and problems between Groups A and B. Again, the differences are obviously higher than the differences for mean impact. For concern impact per hour, Group A’s mean is 105% higher than Group B’s, vs. 65% for concern impact mean. For problem impact per hour, Group A’s mean is 108% higher vs. 89% for problem impact mean difference. It shows Group A had roughly twice the cost effectiveness of group B.

The p-values of cost effectiveness of concerns and problems are 0.004 (Concerns) and 0.007 (Problems) in table 4-3. These results strongly show that the two groups have different cost effectiveness means statistically.

<table>
<thead>
<tr>
<th>Issue per hour</th>
<th>Group</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>% Group A higher</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerns</td>
<td>A(VBR)</td>
<td>16.75</td>
<td>8.46</td>
<td>105%</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>B(VNR)</td>
<td>8.16</td>
<td>3.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problems</td>
<td>A(VBR)</td>
<td>10.14</td>
<td>5.67</td>
<td>108%</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>B(VNR)</td>
<td>4.87</td>
<td>1.93</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4-10. Cost Effectiveness of concerns and problems per hour in both groups

Figure 4-10 shows cost effectiveness of the two groups with boxplots. The differences of cost effectiveness of issues in the two groups are shown clearly in the figure. For both concerns and problems
cost effectiveness, there is no data in group B over the mean (16.75(concerns), 10.14(problems)) of group A.

5. Threats to Validity

5.1 Nonuniformity of Projects

Each project covered a different application with different client and technical characteristics. Some were highly COTS-intensive with fewer artifacts to review. Some were done very well with fewer problems to find. Some were highly dynamic, leading to version mismatches between developers and IV&Vers. Some had relatively good communication between on-campus developers and off-campus IV&Vers, while other projects had significant communications problems. These are all sources of higher variability across projects, leading to higher standard deviations of each group, but not sources of bias between groups. In situations with two IV&Vers on a project, we randomly assigned one to Group A and one to Group B, which tends to reduce some of the variability across projects.

5.2 Nonuniformity of Subjects

In both Groups A and B, there were outlier performers who were either tremendously effective or in over their heads. In some cases, this also led to unreliable data reporting. As a result, we excluded the highest and lowest performers in Groups A and B. Subjects also exhibited considerable variability across reviews, particularly if their job duties were heavy at review times. Again these differences led to higher standard deviations, but not sources of bias.

5.3 Nonuniformity of Motivation

We presented the experiment as a comparison of tried and untied review methods. We provided equal training time to both groups, and indicated that we were not sure which method would work better. Grading criteria (primarily Number of Problems) were uniform across the two groups, but the two groups were graded on separate curves, with very similar grade distributions. Again, these tended to reduce bias between groups.

5.4 Treatment Leakage

We tried to avoid leakage of the value-based guidelines to Group B IV&Vers, and do not know of any complete access to the value-based guidelines by Group B subjects. But as we were teaching value-based methods elsewhere, some concept leakage was inevitable. Some of the better Group B subjects showed a more value-based orientation, but that could have come through innate value orientation or job experience.

5.5 Nonrepresentativeness of Subjects

Although the on-campus development teams are primarily full-time MS-level students, the IV&Vers are almost all full-time professional employees taking the course via distance learning. Their review schedule conflicts were similar to review schedule conflicts on the job. Thus the results should be reasonably representative of industrial review practice.

6. Discussion and Conclusions

<table>
<thead>
<tr>
<th>Table 6-1. The P-values from the T-tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>By Number</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Average of Concerns</td>
</tr>
<tr>
<td>Average of Problems</td>
</tr>
<tr>
<td>Average of Concerns per hour</td>
</tr>
<tr>
<td>Average of Problems per hour</td>
</tr>
</tbody>
</table>

The T-test is done to estimate the probability of equality of the two groups’ means. The T-test p-values, which are explained in the previous section, are shown in Table 6-1. The results show that the equal-means hypothesis can be rejected at the p=0.05 level in six cases (average number of concerns and problems per hour, impact of concerns, impact of problems, concern cost-effectiveness, problem cost-effectiveness).

The results of the T-tests in Table 6-1 show that one cannot reject the hypothesis that there are no differences in number of concerns and problems found by IV&Vers between the Value based review group and the traditional review group, even though we found that there are some gaps between the two groups’ means.

However, we can reject the hypothesis that the numbers of concerns and problems per hour found by IV&Vers do not differ in both groups (0.026 for concerns, 0.023 for problems). The mean effort that the IV&Vers had spent for reviewing was lower for group A, which causes the difference. Also, there are
The overall 65% to 89% higher mean effectiveness metrics of the value-based review group over the value-neutral review group and the overall 105% to 108% higher mean cost-effectiveness of the value-based review group is strongly indicative evidence that value-based review techniques were roughly twice as cost-effective than value-neutral techniques.

One of the other major benefits of using a value-based checklist is educational. Qualitative IV&Ver feedback indicated that the value-based checklists provided stronger support for understanding review objectives and for prioritizing review effort.

As a result, we have committed to the value-based checklists for both intra-project and IV&V reviews for the current projects and for future projects.

7. Acknowledgement

This research was partially supported by USC’s Value-Based Science of Design grant with the National Science Foundation and by NASA/CMU’s High Dependability Computing program.

8. References