Optimizing Requirements-Based Test Reviews

Mike Rowe, Ph.D.
Computer Science and Software Engineering Department
University of Wisconsin – Platteville
and
Esterline Control Systems - AVISTA
Platteville, Wisconsin 53818
rowemi@uwplatt.edu

Thomas Bragg, Craig Hale, Renee Tigges,
Richard Kamm, and Mara Bechwar
Esterline Control Systems - AVISTA
Platteville, Wisconsin 53818
Tom.Bragg@esterline.com, Craig.Hale@ esterline.com,
Renne.Tigges@ esterline.com, Richard.Kamm@ esterline.com,
Mara.Bechwar@esterline.com

Abstract

Esterline Control Systems – AVISTA (AVISTA) is one of the few US companies that is both CMMI Level 5 and ISO 9001 certified. Part of CMMI Level 5 certification and maintenance is Organizational Innovation and Deployment (OID). OID requires that enterprises make use of process metrics to refine and optimize their processes. One of AVISTA's frequently performed activities is testing mission and life critical avionics and medical systems – basically, if one of these systems fails people may die. Thus, it is very important to optimize the number of defects found before system deployment.

This paper discusses a study done to understand and optimize the efficiency of reviews of Software Requirements-Based Tests (RBT). RBT reviews are reviews of tests to ensure the tests accurately and completely test low-level requirements of software. AVISTA has developed an automated review tool that captures detailed review data and allows the subsequent analysis of this data.

This study specifically, examined metrics from over 600 RBT reviews done between February 2007 and October 2009, The study looked at such things as number of defects detected per review, number of people that participated in each review, review size based on how many requirements were exercised by the RBTs in each review, time spent in each review, and derivatives of these metrics. Derived metrics include such things as defects per reviewed requirement, defects per review participant, review time per requirement, etc. The goal of this study was to maximize the number of defects detected while not wasting resources.

One of the findings from this study was that as review sizes increased, as measured by requirements covered per review, defects per requirement dropped. AVISTA's mean review size was 26.55 requirements per review (standard deviation = 46.19 and median = 11). Based on the results, AVISTA has set pilot guidelines to limit the size of reviews where possible at 5 requirements per review.

Other interesting review results are also presented along with future research and a discussion of compounding variables.

1 Introduction

In July 2007, Esterline Control Systems - AVISTA (AVISTA) became the 22nd U.S. and one of only a few Midwest software companies to attain Software Engineering Institute's (SEI) Capability and Maturity Model Integration (CMMI) Level 5. AVISTA has also been ISO 9001 certified since 2003. Over the past 22 years, AVISTA has provided over 2.5 million hours of safety- and mission-critical software engineering services to the world's leading aerospace and medical companies.

Capability Maturity Model (CMM) and more recently Capability Maturity Model Integration(CMMI) has been promoted by government agencies in their effort to understand what makes software system providers successful. CMMI is broken into five maturity levels. The lowest, Maturity Level 1 – Initial, is characterized by no formal process, and the highest, Maturity Level 5 – Optimizing, is characterized by organizations that are continually optimizing their processes. Today CMMI Level 3 or higher is required to perform work on many major government contracts and is of increasing importance in the competition for non-government contracts.

Maturity Level 5 consists of two process areas Causal Analysis and Resolution (CAR) and Organizational Innovation and Deployment (OID). The purpose of CAR is to determine the root cause of problems and take action to prevent them from occurring in the future. The purpose of OID is to deploy innovations that can improve an organization's processes and technologies. Both CAR and OID require measurement to support that improvement is actually occurring.

Once an organization attains CMMI Level 5 certification, it must continue to show evidence that it meets or exceeds the requirements in each of the 22 CMMI process areas.

Since AVISTA deals primarily with safety- and mission-critical software, a great deal of effort is invested in quality assurance activities. One quality assurance activity is requirements-based testing (RBT), which helps ensure that production code satisfies all requirements. Before RBT can be performed on production code, the RBT test cases must be verified to determine that they are correct and sufficient to accurately test the requirements. If defects are not detected during this review process, it is likely that correct production code will fail or defects in production code will not be detected. Neither of these outcomes is desirable.

Barry Boehm [1] reports that peer reviews early in the software lifecycle catch between 31 and 93% of software defects. There is a wide spread in the effectiveness of peer reviews. Boehm [1] indicates that some of the factors that affect the percentage of defects caught include the number and type of peer reviews performed. Due to the importance of high quality in AVISTA's software domain, AVISTA is very interested in optimizing its review process to maximize the likelihood that RBT reviews detect defects. This paper presents results of a recent study of the review of 611 RBT that test a total of 17,339 requirements that were performed between February 2007 and October 2009.

2 Data Collection

The focus of this study is the optimization of AVISTA’s RBT review process. Some customers prefer different processes and these reviews were not included in this study.

Part of AVISTA’s process includes the use an internally developed review tool, Data Item Review Tool (DIRT). This tool allows engineers to record and manage action items found while reviewing data items. DIRT has been used in over 10,000 reviews across all project phases, including: System Requirements, Interface Requirements, Software Requirements, Software Design, Software Code, System Test, Software Requirements Based Test, and Structural Coverage Test. A data item is a unit that is relevant to the type of review being performed, for instance Software Code data items are measured in Lines of Code (LOCs) and RBT reviews are measured in the number of requirements that the test case exercises.

Not all action items are defects, for instance action items can be classified as Defects, Enhancements, Issues, Duplicates, or Trivial. A trivial action item type is described as an “Error or omission which does not have an effect on product function or prevent parallel or follow-on activities from proceeding (examples include typographical error, wrong spelling, wrong word, etc.)”, whereas, a defect is described as an “Error or omission which will result in a product that does not fulfill customer, process, or project requirements if not corrected”. This study only investigates defect action items.

To understand the wealth of data available for analysis, the next sections describe the data collected and maintained by DIRT. A sample report from a bogus RBT review is included in Appendix A.

2.1 Review Data

Field	Description
Project, Job	The Project name and the Job name within a Project.
Review Date	When the review was held.
Review Method	The type of review Desk (no formal meeting, engineers independently review data items at their desk), Peer (desk review followed by an actual meeting of all participants).
Location	Conference room reserved for the Peer review.
Total Prep Time	The sum of all participant review time in minutes spent prior to a review meeting. See Review Participant Data’s Prep Time field.
SQA Signoff, Date, Time	SQA person who closed the review and the date and time of the closure.
48-Hour Notification Waiver	There is a policy which requires a minimum 48 hour announcement of a review. This can be Waived by SQA under special circumstances.

Field	Description
Created By, Created Date Last Modified By, Update Date	Who created the review and when was it created. Who last modified the review and when was it last modified.

Table 1: DIRT review data fields

2.2 Review Participant Data

Field	Description
Name	Engineer's name
Role	The engineers role in the review: Producer (of the data item), Recorder (taking notes in DIRT), Moderator (runs the review), Reviewer (people that have been asked to review the data item), SQA (a member of the SQA organization who is responsible for ensuring that the review process is followed and making sure all action items are eventually closed). Note all roles may submit action items.
Prep Time	Time spent reviewing the data item prior to the review meeting in minutes.
Attended	Yes or No – did they attend the review meeting.
Signoff Timestamp	Date/Timestamp when all action items were signed off.

Table 2: Dirt review participant data fields

2.3 Data Items to Review

Field	Description
Item Name	File name that contains the item to be reviewed.
Location	Configuration Management Path to the data item.
Review Type	The lifecycle stage of the data item.
Total Size	The size refers to a metric related to data item being reviewed. For RBT this would be requirements exercised by the data item. Note this includes both AVISTA and customer developed size.
AVISTA Size	Sometimes an item being reviewed may be fully or partially developed outside of AVISTA. This field contains the count of the AVISTA developed part of the data item. In the case of RBT, this would the number of requirements exercised by AVISTA developed test cases.

Field	Description
Prior Version, Reviewed Version, Post Review Version	The version of the data item refers to a configuration management checked in version. Prior Version refers to any previously reviewed version of the data item, the Reviewed Version is the version currently being reviewed, and the Post Reviewed Version contains all resolved action items.
Configuration	A "Baseline" review reviews all parts of a data item, whereas a "Changes-Only" review considers only those parts of a data item changed since the last time the data item was reviewed.
Conclusion	At the end of the review meeting, if only a few minor action items need to be addressed the meeting attendees may agree that only "Revise [with] No Review" is sufficient. If many action items are found then the attendees may agree to have another meeting "Revise and Review", if no action items need to be resolved the attendees will most likely conclude to "Accept" the review as complete.

Table 3: DIRT Data Item Fields

2.4 Action Items data

Field	Description
Data Item(s)	The data item in which the action item was found.
Originator	Person who submitted the action item.
Responsibility	The person responsible for researching and resolving the action item.
Line/Page Reference	Location of the item. For instance if this is code written to automatically test a requirement this would refer to the line in the source code.
CR Reference	If this action item pertains to software from an external source, the CR (Change Request) will reference the CR number sent to the external source.
Checklist Ref	Many projects use review checklist to guide the reviewing process. This field refers to the checklist item(s) that is violated for defects.
Action Item Type	Defects, Enhancements, Issues, Duplicates, or Trivial.
Severity	Defects are categorized as High, Medium and Low Severity. Other Action Item Types do not use this field.
Injection Source	Whether the defect was injected by AVISTA "Internal" or an "External" company.
Injection Point	The lifecycle stage in which the defect was introduced into the system. For instance with a RBT, a defect could be the RBT test case, the requirement that the test case is based on, or any other lifecycle stage up to and including RBT.

Field	Description
Detection Point	The lifecycle stage in which the defect was discovered. This would include any lifecycle stage from the Injection Point on to customer deployment.
Description	A detailed description of the action item.
Resolution	A detailed description of how the action item was fixed or handled.
Resolved By	Name of the engineer who resolved the Action Item.
Time to Resolve	Resolution time in minutes
Resolution Date	Date/time the resolution was submitted.
Verification	A detailed description of how the Resolution was verified.
Verified By	The engineer who verified the resolution.
Time to Verify	Verification time in minutes.
Verification Date	Date/time the verification was submitted.

Table 4: DIRT Action Item Data Fields.

3 Results

3.1 Raw Data Descriptive Statistics

The below tables characterizes the data set used in this study.

Data Item	Value
Number of reviews	611
Total number of Requirements covered by reviewed tests	17339
First Review analyzed	February 2007
Last Review analyzed	October 2009
Total number of different customers	10
Total number of projects, some customers had multiple projects	21
Total number of jobs, some projects had multiple jobs	34
Mean number of requirements per review	26.55
Median number of requirements per review	11
Standard deviation of requirements per review	46.19
Mean number of participants per review	6.11
Median number of participants per review	5
Standard deviation of participant per review	3.07
Mean preparation (review) minutes per requirement	30.47
Median preparation (review) minutes per requirement	16.67
Standard deviation of preparation (review) time per requirement	46.18

Table 5: Characterizations of the data analyzed.

3.2 Data Relationships: Results

AVISTA considers defect rates proprietary so relationships involving number of defects will be obscured.

The first analysis that was performed was to produce a correlation matrix of review data showing the correlations among:

1. Total Preparation Time per Review (TPT/R)
2. Defects per review (D/R)
3. Requirements per Review (Rq/R)
4. Preparation Time per Requirement (PT/Rq)
5. Defects per Requirement (D/Rq)
6. Number of Participants (NP)
7. Preparation Time per Participant (PT/P)

	D/R	**Rq/R**	**PT/Rq**	**D/Rq**	**NP**	**PT/P**
TPT/R	0.319 0.000 0.102	0.240 0.000 0.042	0.293 0.000 0.086	-0.100 0.015 0.010	0.025 0.539 0.000	0.252 0.000 0.048
D/R		0.320 0.000 0.102	-0.096 0.020 0.009	0.148 0.000 0.022	0.114 0.006 0.008	-0.104 0.011 0.011
Rq/R			-0.252 0.000 0.035	-0.186 0.000 0.035	0.092 0.026 0.062	-0.249 0.000 0.062
PT/Rq				0.285 0.000 0.004	-0.064 0.123 0.004	0.915 0.000 0.837
D/Rq					-0.070 0.088 0.005	0.275 0.000 0.076
NP						-0.258 0.000 0.067

Table 6: Correlation Matrix of review variables. Within each cell, the top number is the correlation coefficient, the middle number is the probability estimate for the correlation, and the bottom number is the R-squared value. The yellow highlighted cells are analyzed in more detail below in this paper. Due to paper length restrictions, only four of the cells are discussed.

Figure 1 shows the rather expected relationship between the Total Preparation Time spent on a review and the number of Defects found in that review. The best-fit trend line is a power function, y = 0.2925 * POW(x, 0.4699). The R-squared values for this and the Linear, Exponential, and Natural Log trend lines are listed across the top of the chart body. The power trend line with an exponent significantly less than 1.0 indicates that as more time is spent there are fewer additional defects discovered, thus as expected there are diminishing returns in performing extended reviews.

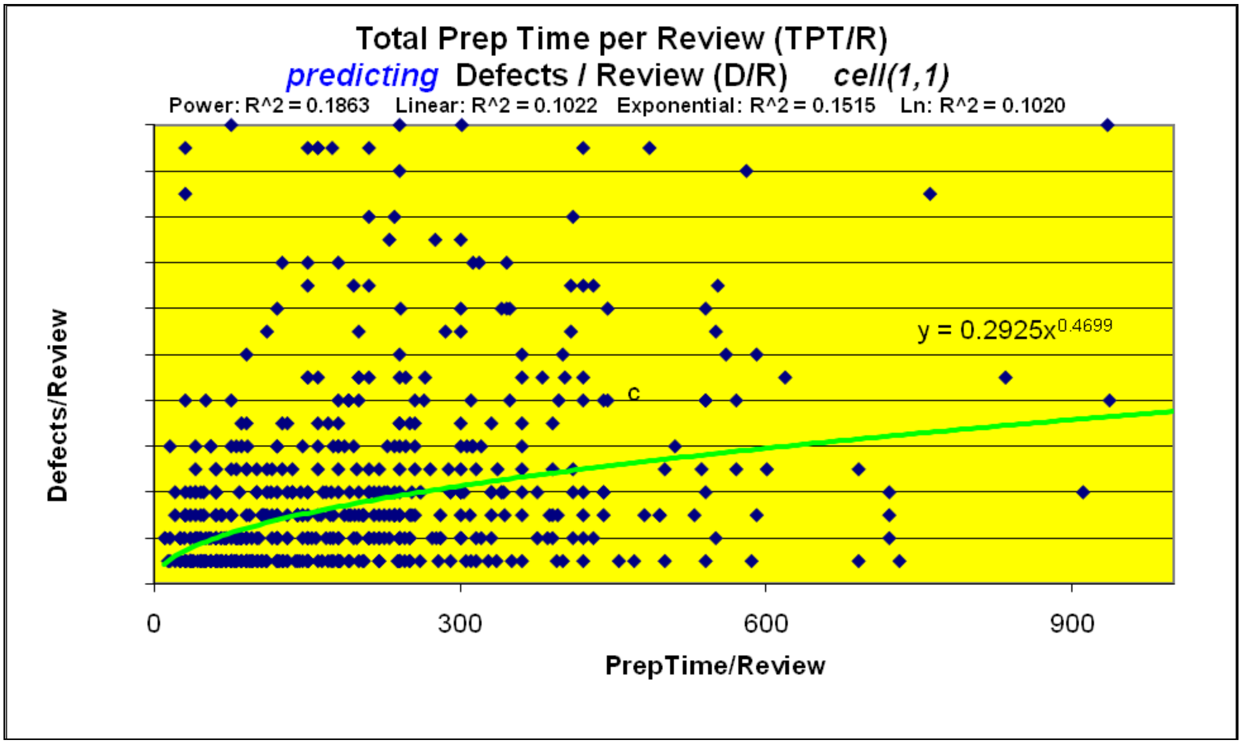

Figure 1: Chart showing the relationship between Total Prep Time (TPT) and Defects per Review (D/R). This is cell (1, 1) of the above correlation matrix (see Table 6). Note the y-axis scale is not labeled to protect the proprietary nature of Defects per Review.

Figure 2 shows the rather expected relationship between Requirements per Review and the Total Preparation Time spent on a review. As the size of the reviews grew, the total review time also increased. The best-fit trend line is a power function, y = 82.587 * POW(x, 0.2771). The R-squared values for this and the Linear, Exponential, and Natural Log trend lines are listed across the top of the chart body. The power trend line with an exponent significantly less than 1.0 indicates that as reviews increase in size, less and less time is spent reviewing per requirement. A conclusion might be that reviewers get bored with long reviews and that one should consider guidelines related to maximum review size. When these results are considered with the findings in Figure 1, as more time is spent fewer additional defects are found, it would seem to support a guideline restricting the size of reviews.

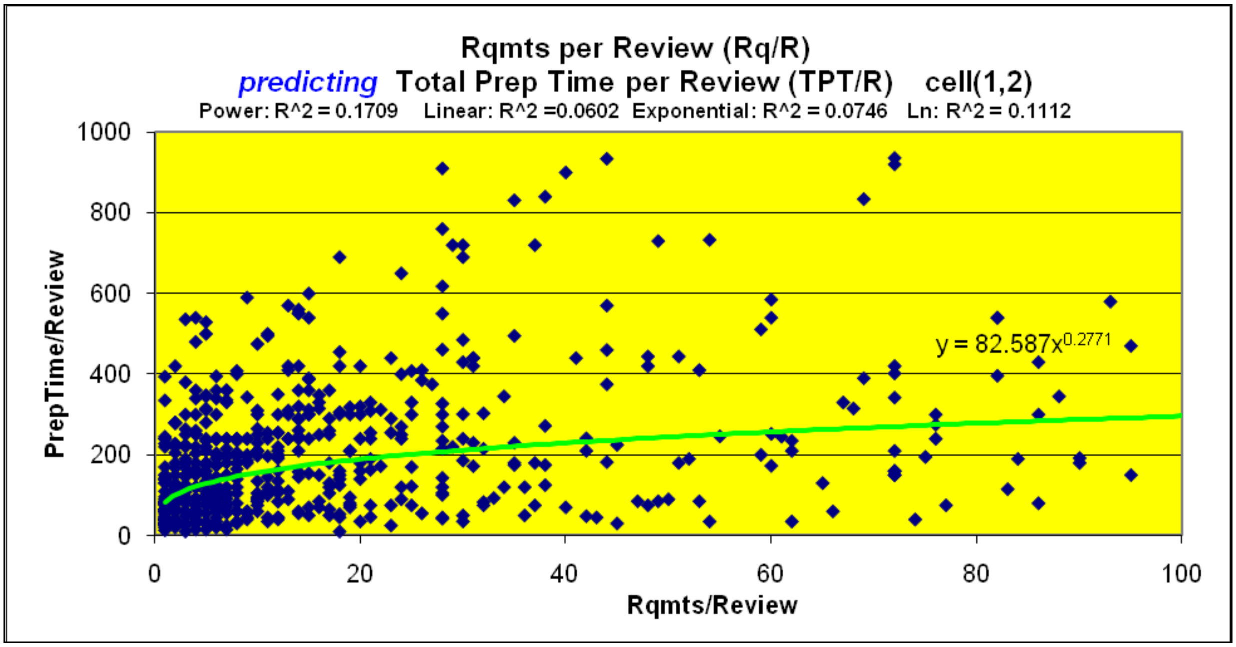

Figure 2: Chart showing the relationship between Requirements per Review (Rq/R) and Total Preparation Time per Review (TPT/R). This is cell (1, 2) of the above correlation matrix (see Table 6).

Figure 3 shows the relationship between Requirements per Review and Defects found per Review. The best-fit trend line is a power function, y = 1.6359 * POW(x, 0.2775). The R-squared values for this and the Linear, Exponential, and Natural Log trend lines are listed across the top of the chart body. The power trend line with an exponent significantly less than 1.0 indicates that as reviews increase in size, fewer defects are found per review. The conclusion is in line with the findings of Figures 1 and 2, that reviewers get bored with long reviews, which results in fewer defects being detected in a review, and that we should consider guidance related to maximum review size.

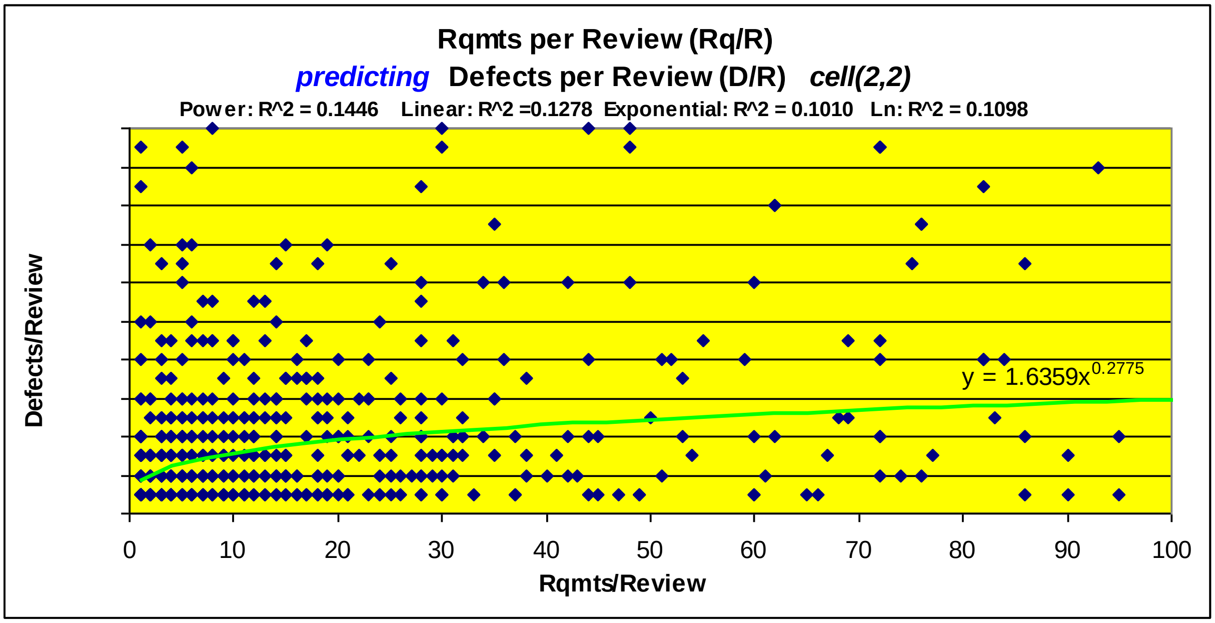

Figure 3: Chart showing the relationship between Requirements per Review (Rq/R) and Defects per Review (D/R). This is cell (2, 2) of the above correlation matrix (see Table 6). Note the y-axis scale is not labeled to protect the proprietary nature of Defects per Review.

Figure 4 shows the relationship between Requirements per Review and Defects found per Requirement (note that the y-axis of scale is not labeled to protect the proprietary nature of defect data). The best-fit trend line is a power function, y = 1.6359 * POW(x, -0.7225). The R-squared values for this and the Linear, Exponential, and Natural Log trend lines are listed across the top of the chart body. The power trend line with a negative exponent indicates that as reviews increase in size, fewer defects are found per requirement. This is consistent with the findings of Figure 3, provides additional support of the conjecture that review sizes should be limited, and also allows us to help set the review size guidance. When the trend line is examined, one can see an inflection at five requirements per review. This indicates that reviews that have more than five requirements are less effective at detecting defects.

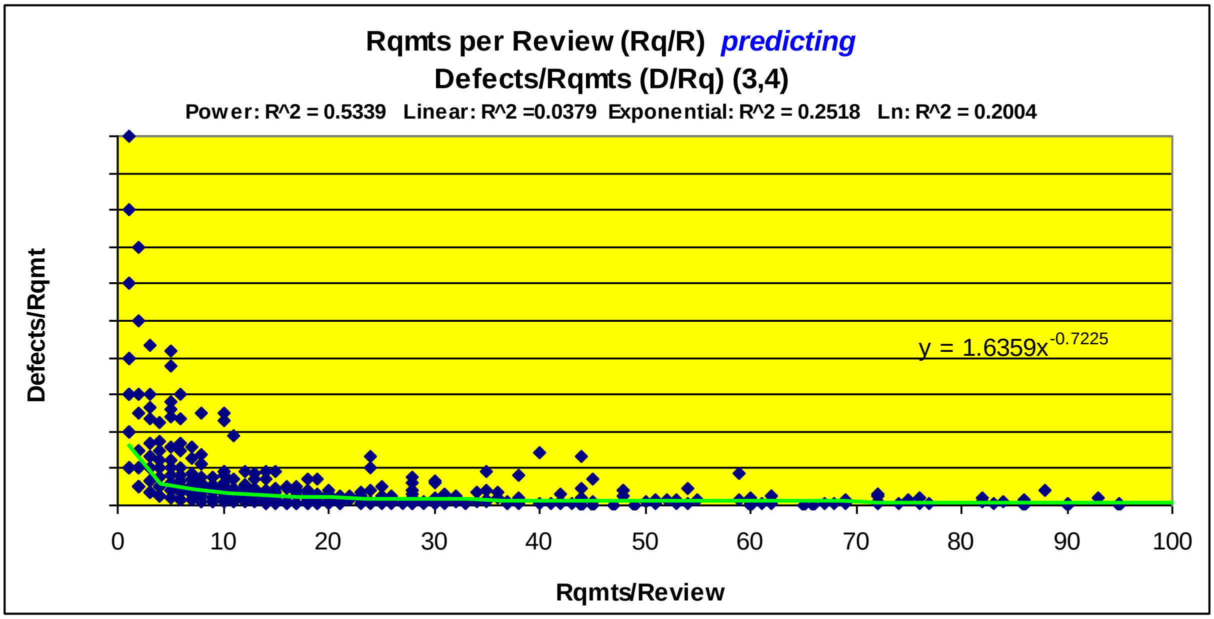

Figure 4: Chart showing the relationship between Requirements per Review (Rq/R) and Defects per Requirement (D/Rq). This is cell (3, 4) of the above correlation matrix (see Table 6). Note the y-axis scale is not labeled to protect the proprietary nature of Defects per Review.

4 Conclusion

AVISTA's historical review size, during the February 2007 to October 2009 period, was much larger than would seem to be optimal (mean = 26.55 requirements per review, median 11, standard deviation 46.19). Based on the findings in Figures 3 and 4, AVISTA initiated a pilot study starting in February 2010 to study the effects of reducing review sizes wherever possible to five or fewer requirements.

Some of the results that are expected include:

1. The Mean and Median of requirements per review will decrease. This will determine if the review size guidance is actually being followed.
2. There will be an increased upfront effort as more reviews will need to be setup.
3. There will be more total time spent reviewing per requirement (Figure 2) showing that smaller reviews spent more time per requirement reviewing.
4. A greater percentage of defects will be caught by reviews (Figure 3 and 4). This should be observed by the defects per requirement increasing in this phase and a reduction of defects per requirement seen in quality assurance activities performed downstream of RBT.
    a. We expect this to further improve quality and actually reduce total lifecycle costs. The prediction of a reduction in total lifecycle costs is based on Barry Boehm's [2] observation that the cost of removing a software defect grows exponentially for each downstream phase of the development lifecycle in which it remains undiscovered.

5 Future Research

The obvious future research is to analyze the results of the pilot study. If these results support the hypothesis that reducing review sizes improves the detection of defects, the five requirement per review will be propagated to all RBT projects and we will continue to monitor defects per requirement rates for RBT reviews.

We will also look at review sizes in additional lifecycle stages, for instance LOCs in code reviews, and nodes for structural testing.

Additionally, we realize that all requirements are not of equivalent complexity – one requirement may be a simple “shall” statement, whereas another requirement may include a complex set of preconditions that must be setup, a complex set of expected results to observe, and a lengthy teardown procedure following the test. We are considering the development of automated tools to process requirements to produce metrics that would normalize requirement complexity. The result of this could mean that we produce a finer tuned metric for specifying RBT review sizes (and also help with other important software engineering tasks like cost and schedule estimation).

References

[1] Boehm, Barry and Basili, Victor, Software Defect Reduction Top 10 List, IEEE Computer, p135-137, v34, Jan 2001.

[2] Boehm, Barry, Software Engineering Economics, Prentice-Hall, 1981

APPENDIX A

Data Item Review Details

Data Item Review 5020

Project:	DEMO PROJECT 1	**Review Date:**	7/16/2007 10:51:47 AM
Job:	QPM - Group Demo Leader	**Review Method:**	PEER REVIEW
Sequence:	2	**Location:**	BIRCH
Total Preparation Time:	160 minutes	**Total Meeting Time:**	0 minutes
SQA Signoff:		**48-Hour Notification Waiver:**	Not Waived
SQA Signoff Date:			
SQA Signoff Time:			
Created By:	CJHALE	**Last Modified By:**	MCROWE
Creation Date:	7/12/2007 10:52:20 AM	**Update Date:**	3/17/2010 12:09:20 PM

Review Participants

Name	Role	Prep Time	Attended?	Signoff Timestamp
Daniel J Klein	PRODUCER	32 minutes	No	
Scott P Derby	REVIEWER	33 minutes	No	
Craig J Hale	REVIEWER	34 minutes	No	
Elisabeth J Farver	RECORDER	25 minutes	No	
Sultana Amin	MODERATOR, SQA	36 minutes	No	

Data Items

Item Name: BogusTestCase001.cs

Location: 0320 - AVISTA Quality Program\QMS_Data\TestsCases\

Review Type:	SOFTWARE TEST CASES/PROCEDURES	**Configuration:**	BASELINE
Total Size:	9	**Conclusion:**	REVISE, NO REVIEW
AVISTA Size:	9		

Prior Version:	0	**Reviewed Version:**	11	**Post Review Version:**	12

Reference Items

No Reference Items

Data Item Review Details

Data Item Review 5020

Action Items

Action Item 1

Review Items:

BogusTestCase001.cs

Originator:	Daniel J Klein	**Action Item Type:**	DEFECT
Responsibility:	Daniel J Klein	**Severity:**	HIGH SEVERITY
Line/Page Ref:	Page 1	**Injection Source:**	INTERNAL SOURCE
CR Reference:		**Injection Point:**	SOFTWARE RQMTS BASED TEST
Checklist Ref:	4	**Detection Point:**	SOFTWARE RQMTS BASED TEST

Description:

Test did not correctly test Rqmt X.

Resolution:		**Verification:**	
Dan fixed the test		Looked at test results and test and the fix is correct	
Resolved By:	Daniel J Klein	**Verified By:**	Craig J Hale
Time to Resolve:	5 minutes	**Time to Verify:**	2 minutes
Resolution Date:	7/12/2007 10:57:38 AM	**Verification Date:**	8/20/2008 4:00:32 PM