Controlling Peer Reviews During Software Development

A 5-Year Longitudinal Case Study

19 November 2008

Richard L. W. Welch, PhD
Associate Technical Fellow
Steve D. Tennant
SEPG Lead
Northrop Grumman Corporation

ISER-MLB-PR-08-137

Topics

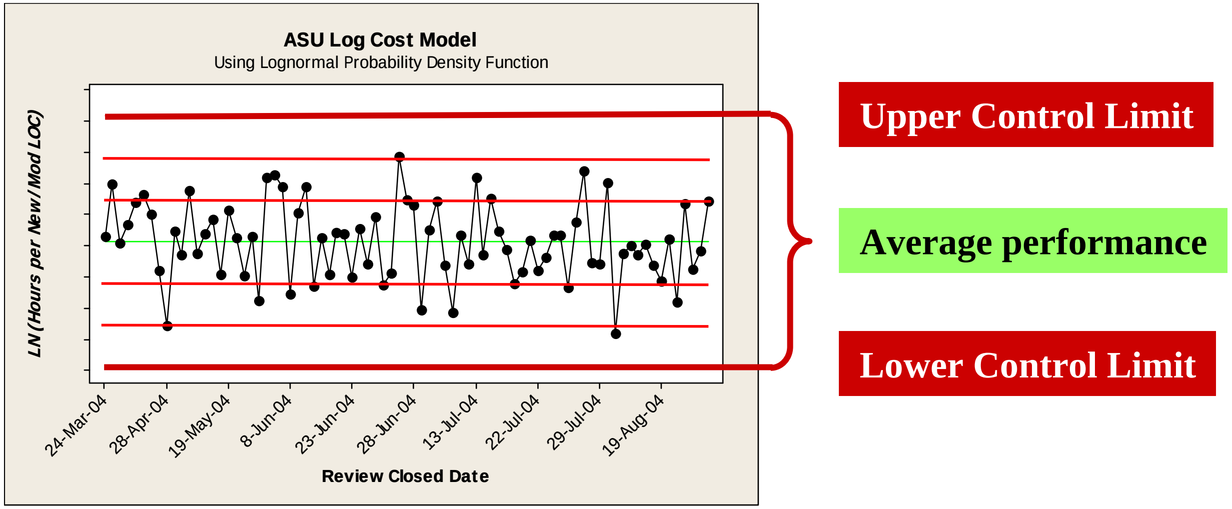
- Background
- Getting started (2004)
  - Overcoming technical issues
- Handling human issues and institutionalizing the process (2005)
- Growing the benefits (2006)
  - More processes, projects, disciplines
- Increasing our effectiveness (2007-2008)
  - Exploring new techniques
- Future pathways

Why Peer Reviews?

NORTHROP GRUMMAN

- Ubiquity
  - Many work products reviewed throughout software development life cycle
    - System & software design artifacts
    - Source code
    - Test plan, procedures & reports
- Frequency
  - High data rates
- Influence
  - Approximately 10% of the software development effort is spent on peer reviews and inspections
  - Code walkthroughs represent biggest opportunity & most advantageous starting point
- Serendipity
  - All engineering disciplines peer review their design products
  - Techniques & lessons learned have demonstrated extensibility

Why Statistical Process Control?
Successful Quantitative Project Management



A stable process operates within the control limits 99.7% of the time

- Analysis of special cause variation focuses on recognizing & *preventing* deviations from this pattern
- Analysis of common cause variation focuses on *improving* the average and tightening the control limits
- SPC offers opportunities for systematic process improvement that NGC & industry benchmarks indicate will yield an *ROI averaging between 4:1 & 6:1*

Case Study Essentials

- Data represent software-related peer reviews conducted at Northrop Grumman's Integrated Systems Eastern Region – Melbourne, Florida facility between March 2004 and October 2008
- One peer review process (now standard for Integrated Systems)
  - Covers the entire system life cycle from system requirements analysis & architecture through maintenance
  - Requires the peer review of all major systems, software & test artifacts
  - Uses an automated data base tool that integrates data quality & process control features
- All peer review records captured in the data base
  - > 5,700 source code peer reviews
  - > 1,100 other software-related peer reviews
- CMMI Level 5 appraisals
  - CMMI-SE/SW (V1.1) in 2005
  - CMMI-SE/SW/IPPD/SS (V1.1) in 2006
  - CMMI-DEV+IPPD (V1.2) scheduled for 2009

NORTHROP GRUMMAN

DEFINING THE FUTURE

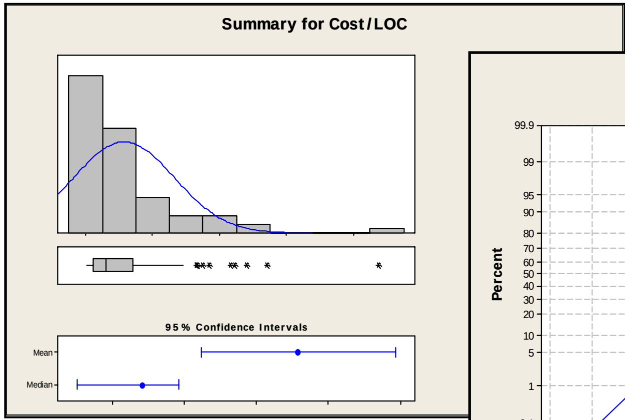
Getting Started - 2004

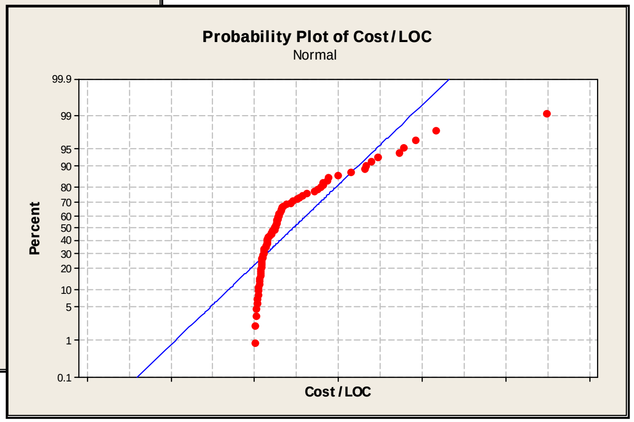
Overcoming Technical Difficulties & Learning To Love Logarithms

SW-CMM Level 4 Prior State (1998-2003)

- Software development baseline characterized by life cycle phase
  - SW Requirements-Design-Code & Verification-SW Integration-Software Test
  - 10+ year process improvement record resulted in costs reduced by over 67%
- But we had no CMMI “Gestalt”
  - No insight into the statistical behavior of lower level elements
  - No “above the shop floor” experience with statistical sub-process control
  - No insight into downstream behavior
- We wanted to control product quality, but were thwarted by issues with our process quality
  - Inconsistent data
  - Superficial results
- Root cause analysis traced this to indifferent attention paid to *managing* peer reviews
  - We realized we had to control the *efficiency* of our peer reviews in terms of the effort spent (peer review cost), based on classic industry guidelines that efficient reviewers operate in a “sweet spot” of about 200 lines of code per review hour

Our Problem



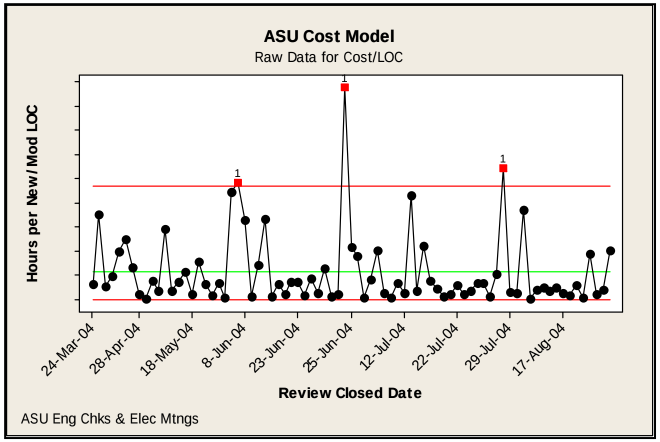


Data Characteristics

- Anderson-Darling test p < 0.005
- Data non-normality & asymmetry violated probability model assumptions

Control Chart Difficulties

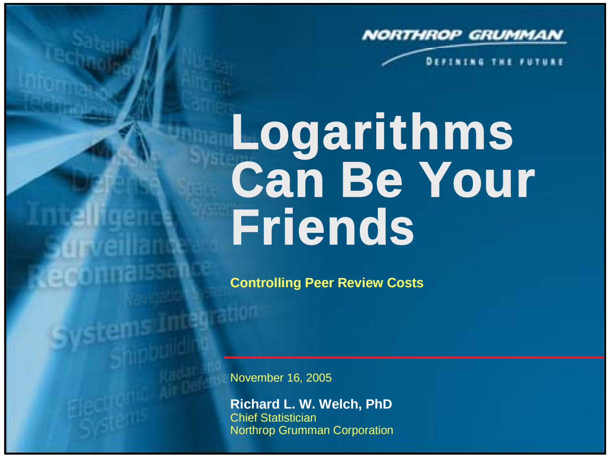
- 11% false alarm rate (Chebyshev's inequality)
  - *Penalized due diligence in reviewing code*
- No meaningful lower control limit
  - *Did not flag superficial reviews*
- Arithmetic mean distorted the central tendency
  - *Apparent cost did not meet budget*



Could We Control Our Peer Reviews?

Stabilizing the Data

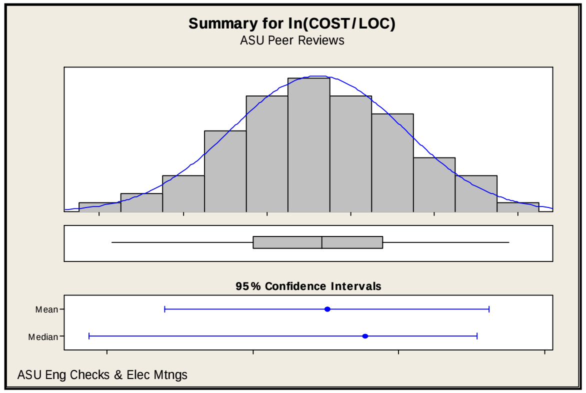
- Senior author’s presentation at 2005 CMMI℠ Technology Conference demonstrated how a log-cost model can successfully control software code inspections

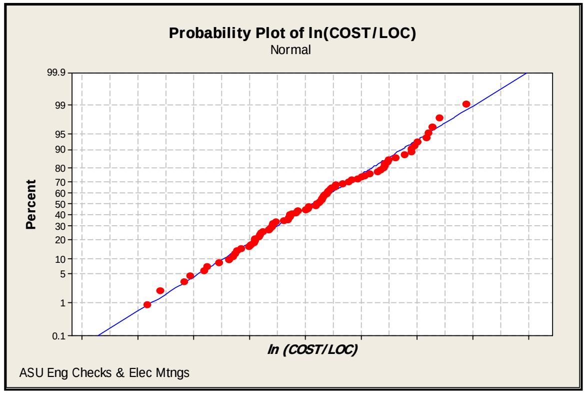


  - Peer review unit costs (hours per line of code) behave like commodity prices in the short term
  - Short term commodity price fluctuations follow a lognormal distribution
  - As a consequence, commodity prices follow a lognormal distribution
  - Therefore, taking the natural logarithm of a sequence of peer review costs transforms the sequence to a normally distributed series

Notes:
- Details on the log-cost model, “one of the most ubiquitous models in finance,” can be found at riskglossary.com (http://www.riskglossary.com/articles/lognormal_distribution.htm)
- Prior CMMI Technology Conference & User Group papers are published on-line at: http://www.dtic.mil/ndia/

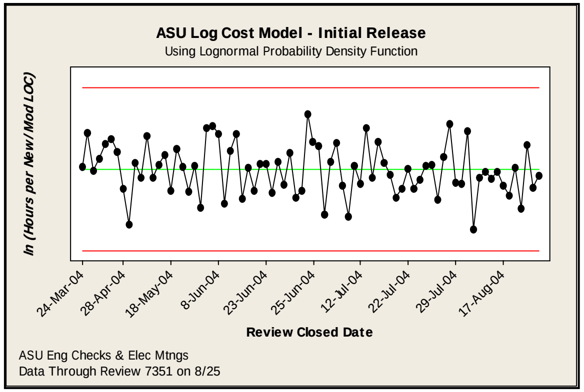
Our Data on Logs





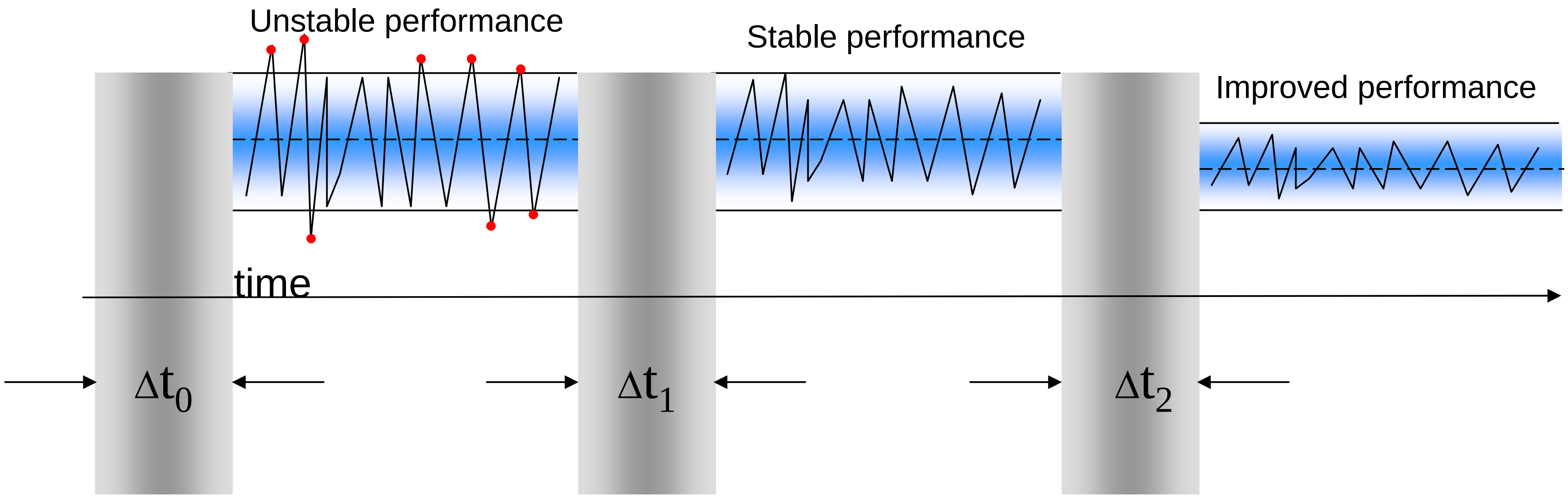
Anderson-Darling test $p < 0.759$

- Impacts
  - False alarms minimized
  - Meaningful lower control limit
  - Geometric mean preserves the budget
    - *OK, you still have to find the antilog*
- Demonstrated utility & applicability
  - > 6,800 peer reviews over 5 years provide large sample validation



A Textbook Demonstration of an In-control, Stable Process

The High Maturity Data Dilemma
Why Management Can't Have It Tomorrow



Δt_0 :
- Process selection
- Analysis of suitability for SPC

Δt_1 & Δt_2 :
- Identify improvement proposals
- Evaluate & prioritize proposals
- Select improvement
- Pilot improvement
- Deploy improvement

We can minimize $\Delta t_{0,}$ Δt_1 & Δt_2 by careful management, but the length of the data runs will depend on the periodicity of the process itself

Handling Human Issues - 2005

And Institutionalizing the Process

2005 Challenges

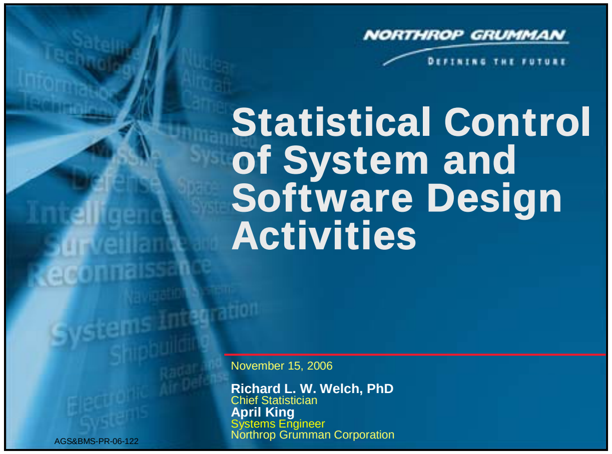
- First demonstration of CMMI(SM) Level 4 and 5 capabilities focused on code inspections and parallel effort to control peer reviews of software test plans, procedures and reports
  - High data rates inherent in these back-end processes helped us to understand and overcome statistical difficulties
  - We gained practical lessons learned on the obstacles that had to be overcome
- Desire to introduce successful SPC techniques for quantitative project management in the front-end system and software design phases
- When coding starts
  - Product development is one-half over
  - Opportunities to recognize and correct special & common cause variation in the design process are gone

First-year Decisions Determine up to 70% of Total Life Cycle Cost on DoD Programs. Early, Effective Statistical Control Offers Great Practical Benefit

Practical Difficulties at Level 4

NORTHROP GRUMMAN

- Getting started
  - Selecting good candidates for statistical management
- Statistical innumeracy
  - Discipline needs to own the right skill set
- Little historical data & inherently low data rates
  - Personnel need familiarity with robust statistical procedures



- Cautionary note: you must also take care of the basics (CMMI℠ Level 3)
  - Budget and charter
    - Project impacts
  - Metrics infrastructure across engineering
    - Metric definitions
    - Data collection mechanisms
    - Consistency of processes across projects

–“Outstanding Presentation for High Maturity”

–“Conference Winner”

Note: Prior CMMI Technology Conference & User Group papers are published on-line at: http://www.dtic.mil/ndia/

Getting Started
Process Selection for Statistical Management

- Statistical control is imposed on sub-processes at an elemental level in the process architecture
- Processes are selected based on their
  - Business significance – “sufficient conditions”
  - Statistical suitability – “necessary conditions”
- Business checklist
  - Is the candidate sub-process a component of a project’s defined key process?
    - Is it significant to success of a business plan goal?
    - Is it a significant contributor to an important estimating metric in the discipline?
  - Is there an identified business need for predictable performance?
    - Cost, schedule or quality
  - How does it impact the business?
    - Need to map sub-process ↔ process ↔ business goal
- Statistical checklist (table)

PRINCIPAL FACTORS INVOLVED IN SUB-PROCESS SELECTION FOR STATISTICAL PROCESS CONTROL

PRIMARY QUESTION	SUPPORTING QUESTION	DEMONSTRATED INDICATOR
Are data collected?	Does the data collection system require update or redeployment?	Data collection system ready for deployment
What is the data rate? That is, how often will the process be repeated on the project?	Will the process be repeated frequently enough to develop control limits if such limits do not exist from baseline historical performance?	At least 20-30 data points exist or will be produced
Are there historical performance data?	Stable performance: Will the process be performed in roughly the same manner as on previous projects?	A documented procedure or training materials are used by those performing the process
Has a control metric been identified?	Has statistical analysis of past project performance identified measures that are indicative of overall process performance?	Control metric can be computed from the collected data
Does a baseline exist for the control metric?	What are the average and statistical variation of previous performance?	Performance excursions outside the control limits can be identified & attributed to their root causes
Do specification limits exist for process performance? (Optional)	Do limits exist beyond which process performance is deemed unacceptable?	Specification limits are documented

Overcoming Statistical Innumeracy
Success Factors

- Minitab
- “Dark green belt” training
    - Curriculum tailored to focus on applied statistical techniques and Minitab familiarity
    - Deming principle applied in the class room
        - In God we trust, *all others bring data*
    - Lean and process management training covered in other courses
- Green belt community of practice
- Chief statistician

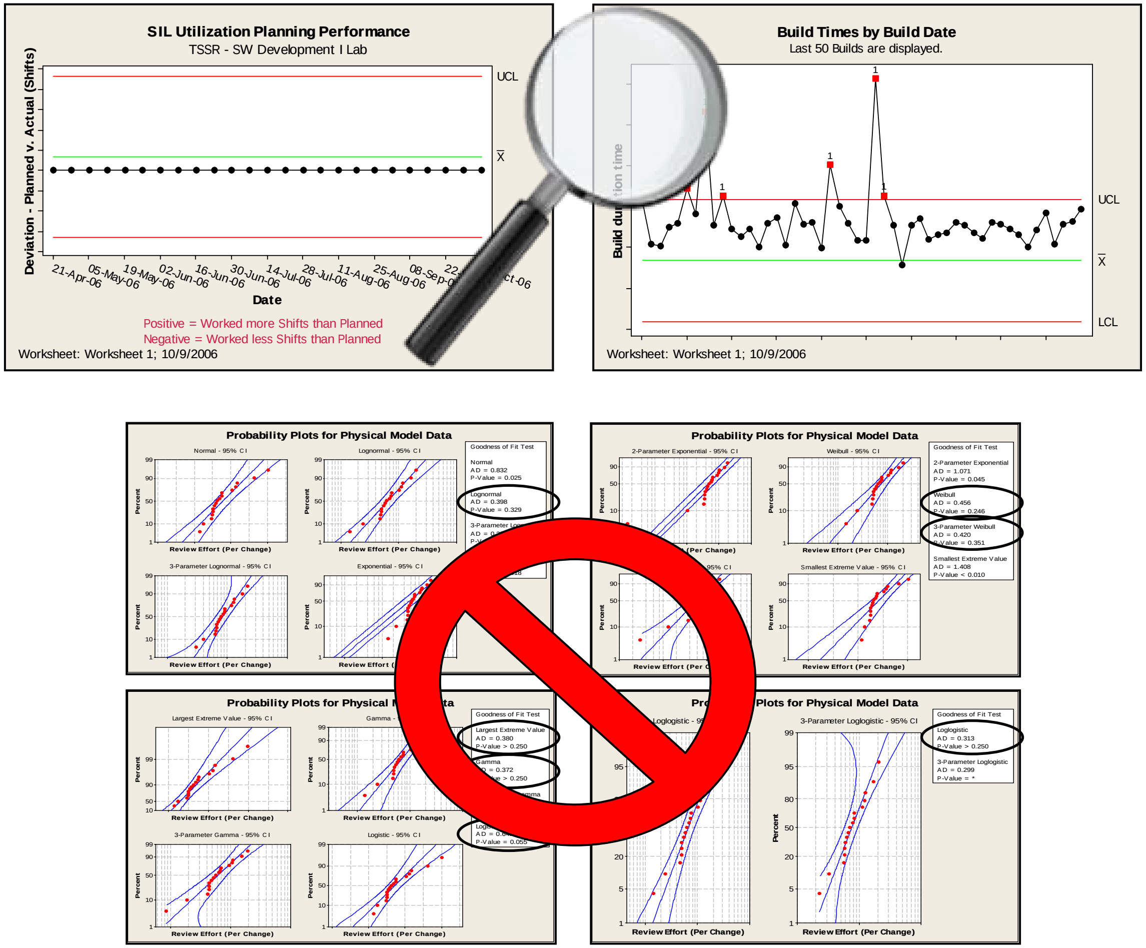
Key Success Factors: Management Recognition & Support for the Investment

Post 2005 Follow-up

- Sector standards for certifying Green Belts, Black Belts & Master Black Belts (2006)
  - Training
  - Project portfolio
- Green Belt certification (2006)
- Black Belt cadre (2007-2008)
- Future Master Black Belt cadre (2009+)

Success Creates a Continuing Need to Grow the Infrastructure

Pitfalls



Dealing With What Was Taught – but Not Learned – in Green Belt Class

NORTHROP GRUMMAN

DEFINING THE FUTURE

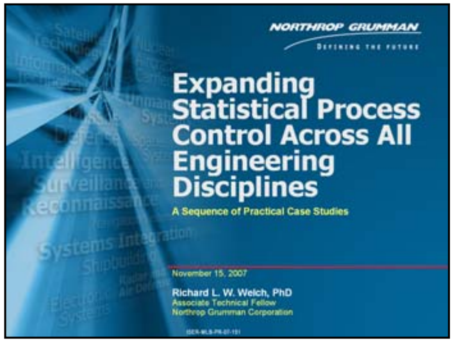
Growing the Benefits - 2006

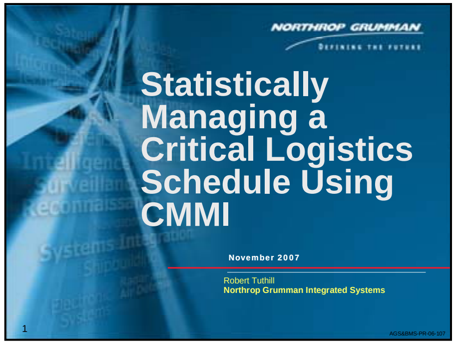
More Processes, Projects, Disciplines

Growth

NORTHROP GRUMMAN

- After our initial success, we aggressively expanded the use of SPC techniques in all engineering projects
  - Led by senior management
  - Clear expectations of significant benefit to the business
  - Particular focus on our hardware and logistics disciplines
- By year-end 2006, we had gone from the original 4 sub-processes under control in 2 Engineering homerooms to 30 sub-processes under control in 6 Engineering homerooms
  - Expect ~45 sub-processes that are significant to our business under active control by year-end 2008





–"Outstanding Presentation for High Maturity"

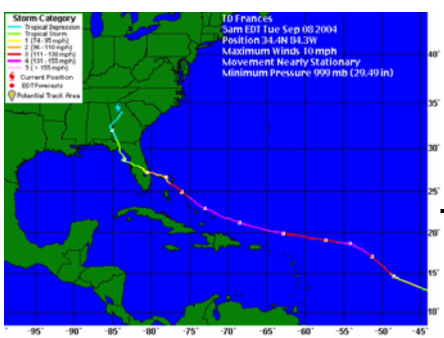
–"Conference Winner"

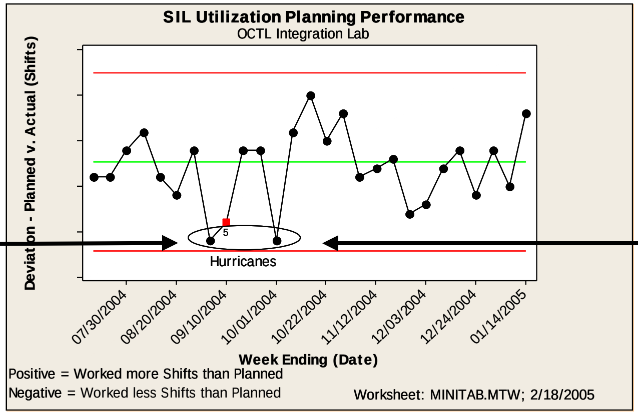
Note: Prior CMMI Technology Conference & User Group papers are published on-line at: http://www.dtic.mil/ndia/

A Humorous Sidebar
Identifying Special Causes

- As part of this effort, our Test & Evaluation personnel analyzed some 2004-2005 baseline data, and asked what has become one of our favorite statistical questions:

  Can isolated points be considered as special cause points, and be deleted from a data set as outliers, even though they don't fall outside of the 3-sigma control limits?





Francis

Jeanne

NORTHROP GRUMMAN
DEFINING THE FUTURE

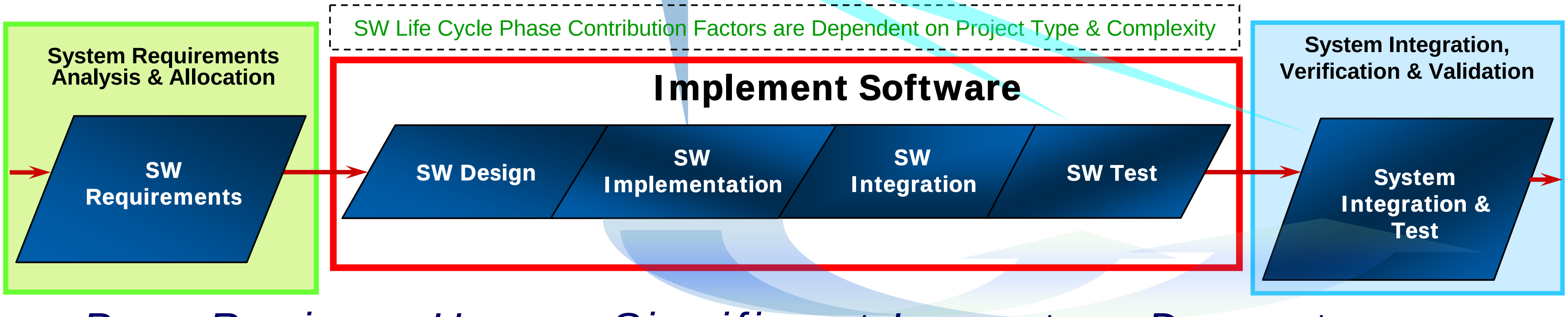
Increasing Our Effectiveness – 2007-2008

Exploring New Techniques

What Is the Benefit?

NORTHROP GRUMMAN

- Organizational Impact
  - 100% of delivered code is peer reviewed
  - Average of 105 reviews completed per month
  - This activity affects all major development & test activities after software design
    - SW Implementation
    - Software Test
    - System Test
    - Ground & Flight Test Support
  - Code review effort constitutes a significant portion of the earned value credit in these phases
- Benefits
  - Increased early defect detection
  - Fewer delivered defects
  - Increased code maintainability, reduced cost on future sustaining programs

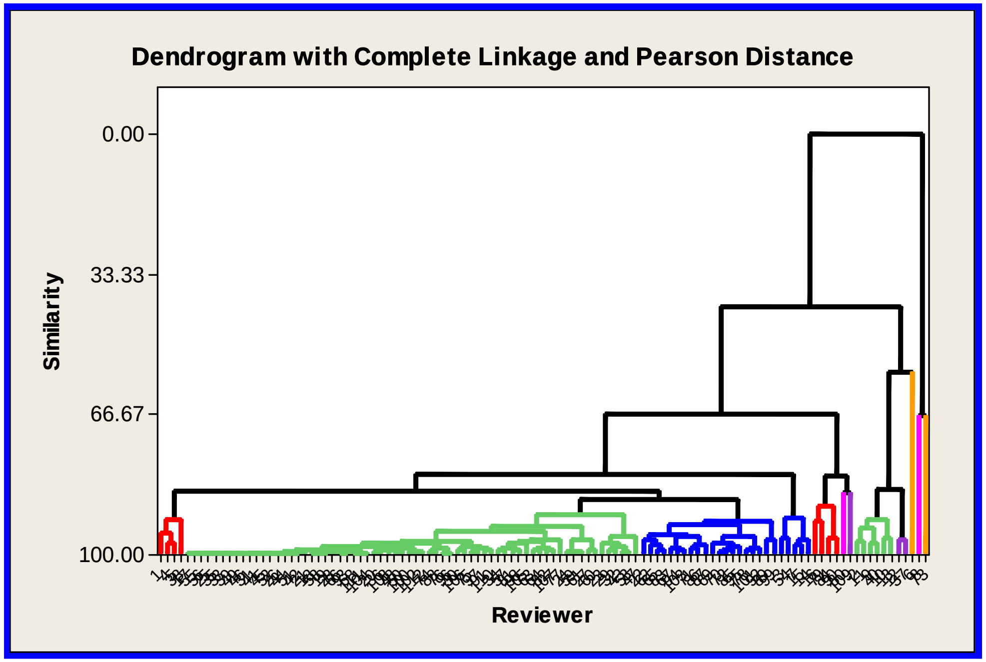


Peer Reviews Have a Significant Impact on Downstream Product Quality and Development Costs

Analytical Approach

NORTHROP GRUMMAN

- Use accumulated data to explore factors related to reviewer performance and experience
- Use a multivariate clustering procedure (agglomerative hierarchical method) to identify groups of reviewers with similar performance characteristics (initially not known)
- Decide how many groups are logical for the data and classify accordingly



- Three reviewer performance categories support the needed level of insight
  - Group 1 reviewers have lots of review experience, review at the best rates, & identify the most defects
  - Group 2 reviewers are newer & less experienced (reflected by the number of reviews they have completed), with a wide range of rates and discovered defects
  - Group 3 reviewers have lots of experience, review at fast rates, but identify significantly fewer defects

A Serious Sidebar
Measuring Individual Performance

NORTHROP GRUMMAN

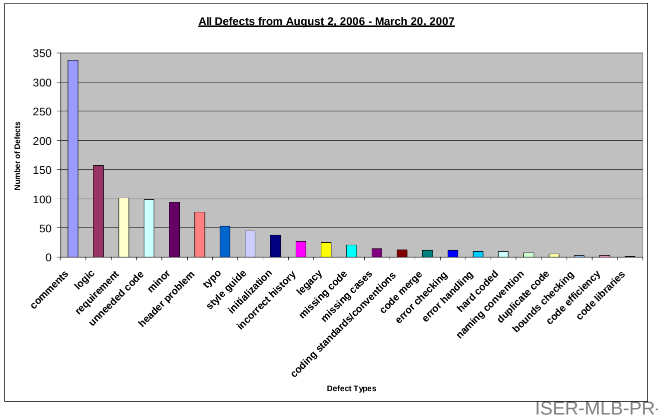
- It violates the peer review process to use numbers of defects to measure the Author's performance ("killing the goose that lays the golden egg")
  - Need reviewers free to report any issues they find, even if they are not totally sure that the item is truly a defect
  - Even the very best and most conscientious engineers create defects – the primary objective of the review is to find and remove any defects
- Peer review database design enables study of individual reviewer performance – *with the express goal of increasing skills through vital, focused training*
  - Good reviewers provide an essential contribution both for the author and the company – reviewer diligence should be encouraged, recognized and rewarded
  - Cumulative data on reviewer performance provides valuable insight - similar to measurements applied in sports

Reviewer Knowledge and Skill Are Key–
Knowing What to Look for & How to Find It…

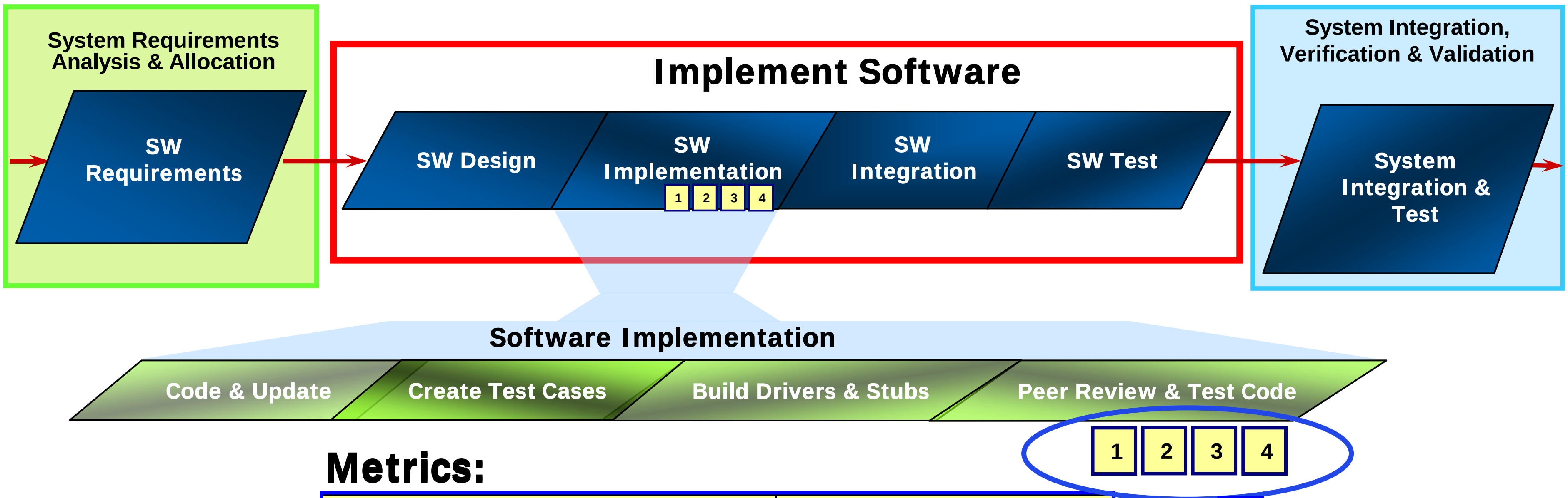
Causal Analysis & Resolution

- After the fact analysis by Group 1 reviewers indicated Group 2 and Group 3 reviewers consistently miss defects
- A retrospective study focused on the common types of defects being discovered
- An improvement team identified ways to increase the skills of Group 2/3 reviewers
  - Pair Group 2/3 reviewers with Group 1 mentors
  - Review and update coding standards to clarify descriptions or address missing elements
  - Develop and deliver a technical-level review training course to provide refreshed or deeper insight into ‘problematic’ programming issues
    - ‘Problematic’ programming issues were identified based on team member experience and results from the retrospective study
  - Enhance checklists

Common Issues Emerge When We Examine Defect Types & Frequencies



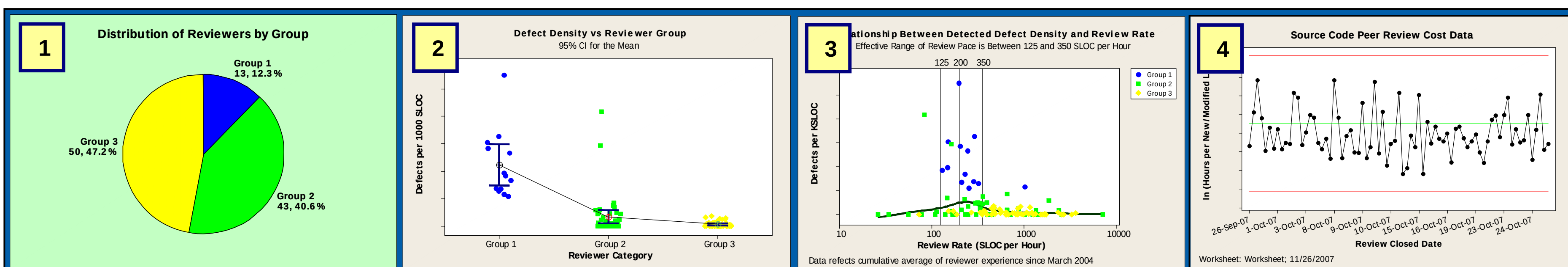
Peer Review Effectiveness Metrics



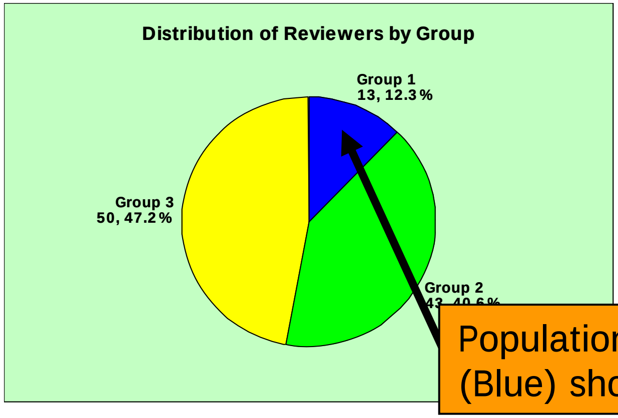
Metrics:

	Peer Review Effectiveness Metrics	Units
1	Percentage of Group 1 Reviewers	Percent of Total Reviewers
2	Average Detected Defect Density	Defects per Thousand SLOC
3	Average Review Rate	SLOC per Review Hour
4	Log Cost Model	Log(Total Hours per SLOC)

Note: SW Test & System Test are decomposed similarly

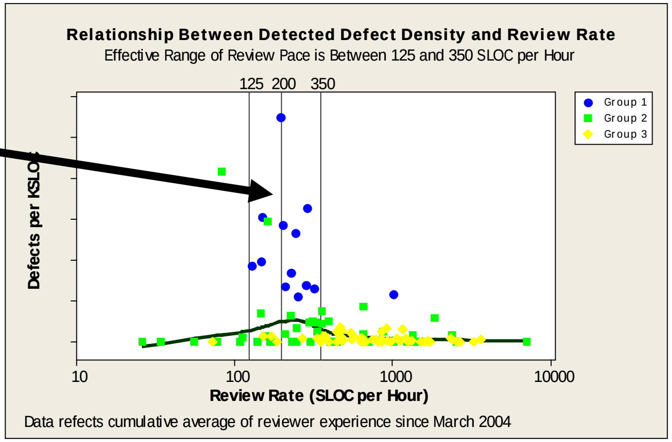


Forecasting the Outcome

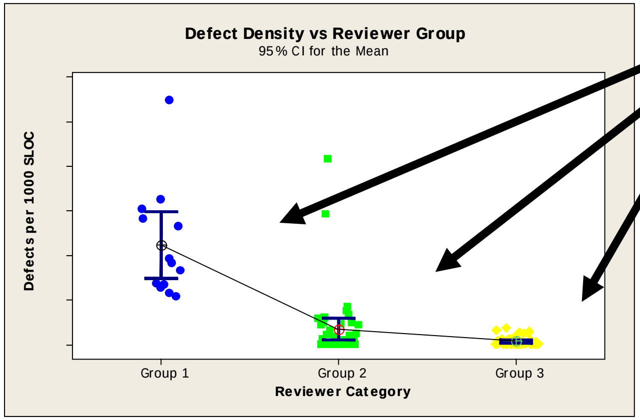


Population should cluster around ideal review rate of 200 LOC/Hr (industry std)

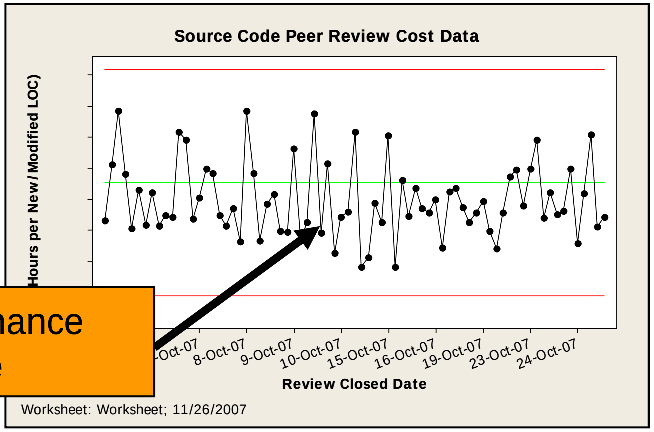
Population of Group 1 (Blue) should increase



We used four ways to measure changes from the initial March 2007 performance baseline. Demonstrating skill development in review effectiveness does not lend to routine control chart monitoring

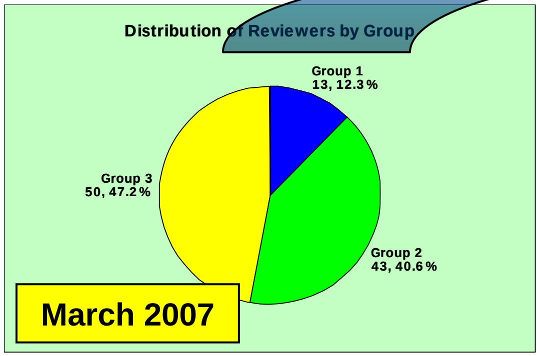


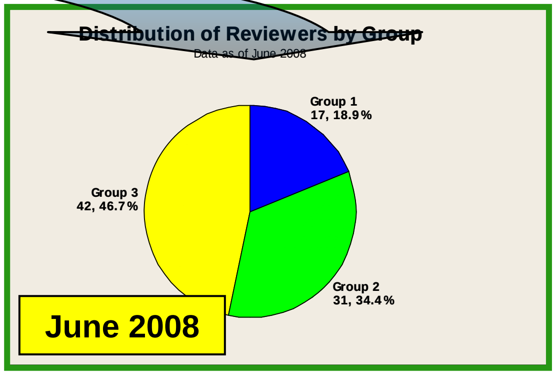
Detected defect density should increase or remain the same



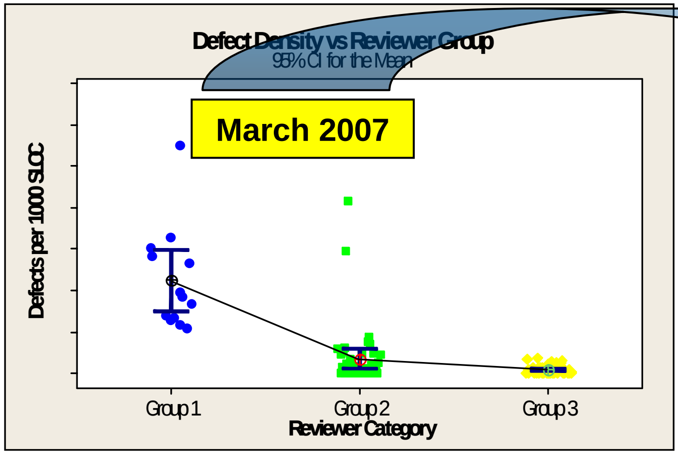
Process cost performance should remain stable

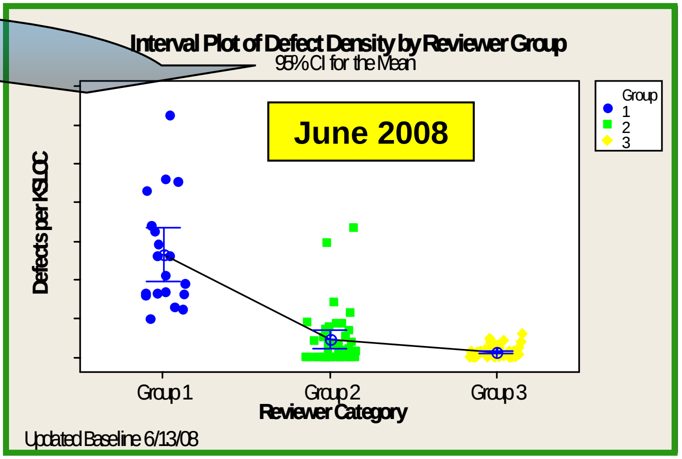
Verifying the Outcome





Group 1 Reviewers Increased by 53%

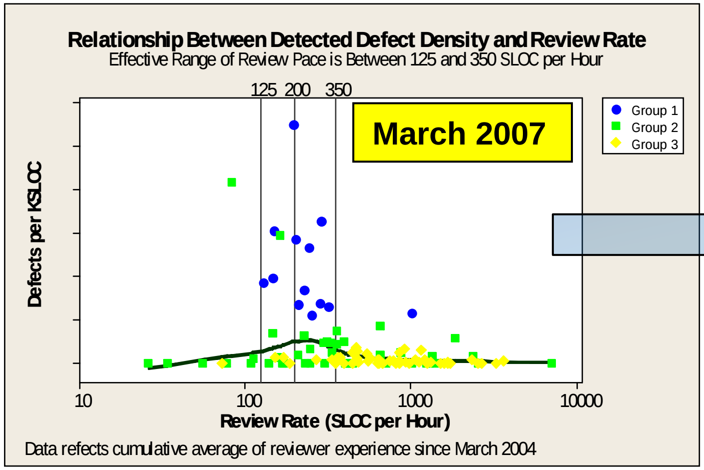


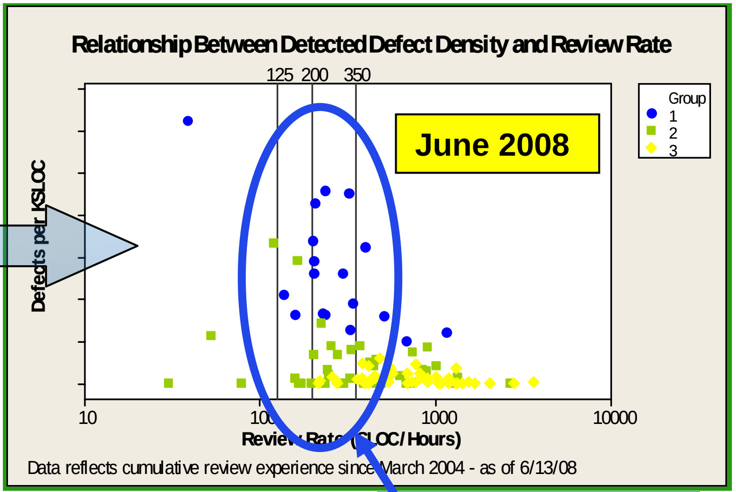


Overall Discovered Defect Density Increased by 56%

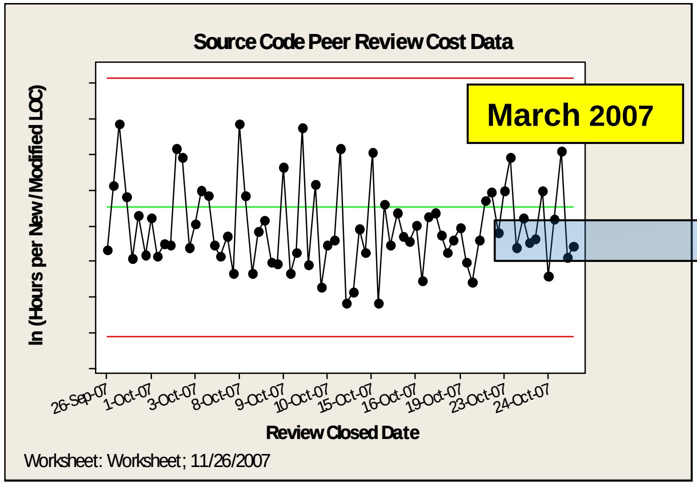
Checking the Control Variables

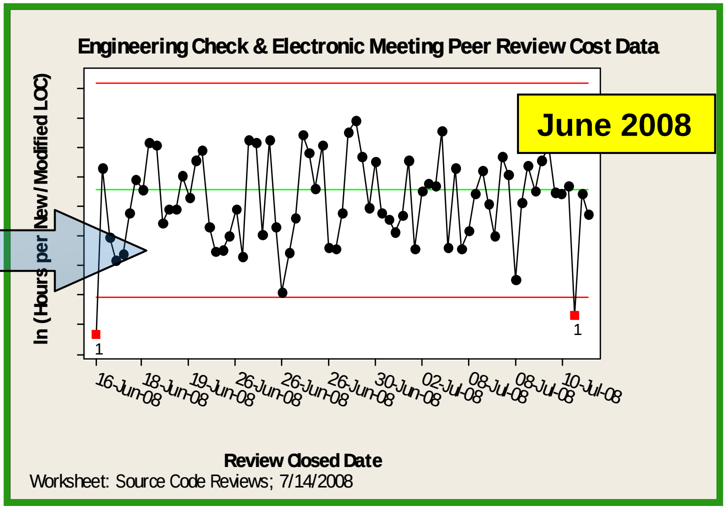
Stable Review Rate





Number of reviewers performing in the ideal range increased





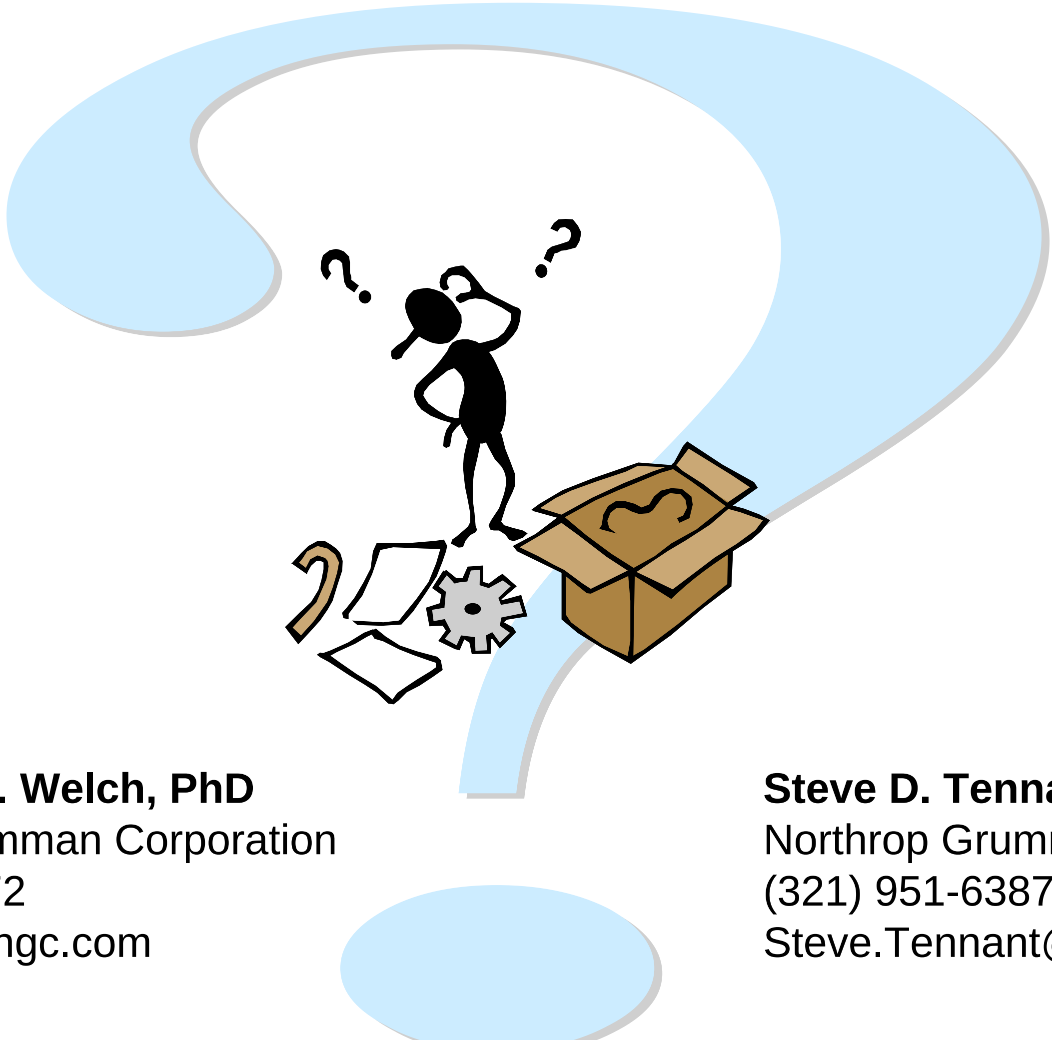
Predictable Process

Future Pathways

- ✓ Maintain strategic focus to sharpen skills
  - ✓ Continue to support inexperienced developers with mentoring
  - ✓ Maintain periodic skill enhancement training
- Continue the quest to remove impediments
  - Better integrated toolsets — In Progress
  - Improved coding standards — In Progress

Bottom Line Motivator:
The 2007-2008 Initiative Has Resulted in a 12% Reduction in the Number of Software Bugs per Release

Questions



Richard L. W. Welch, PhD
Northrop Grumman Corporation
(321) 951-5072
Rick.Welch@ngc.com

Steve D. Tennant
Northrop Grumman Corporation
(321) 951-6387
Steve.Tennant@ngc.com

NORTHROP GRUMMAN

DEFINING THE FUTURE