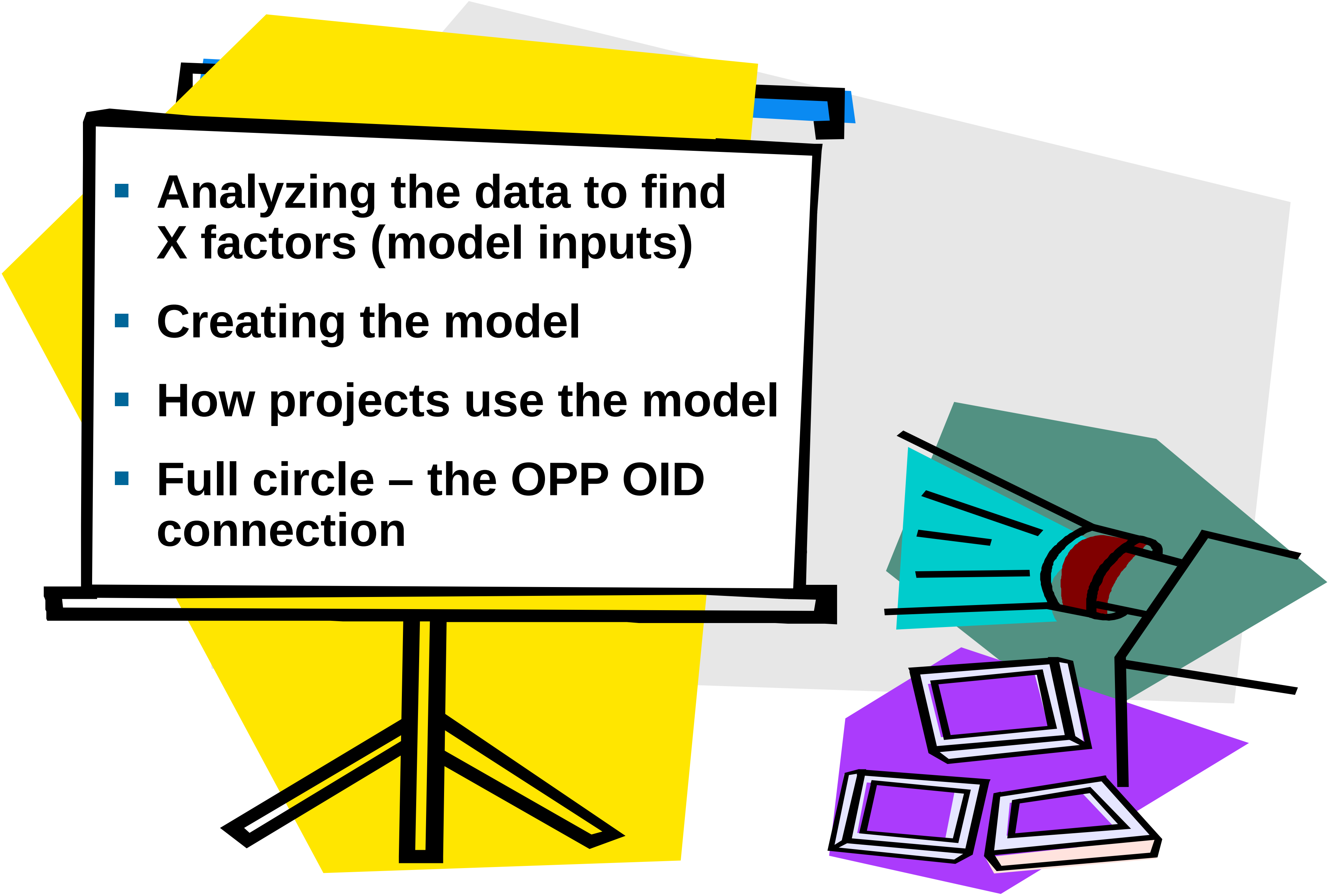
How I Created Our Peer Review Baselines and Models

CMMI Technology Conference
November 16-19, 2009

Diane Mizukami-Williams
Northrop Grumman Corporation

Agenda



- Analyzing the data to find X factors (model inputs)
- Creating the model
- How projects use the model
- Full circle – the OPP OID connection

Northrop Grumman Information Systems (IS) Sector

IS Sector

- $10 billion in sales in 2008
- 7,000 contracts
- 33,000 employees

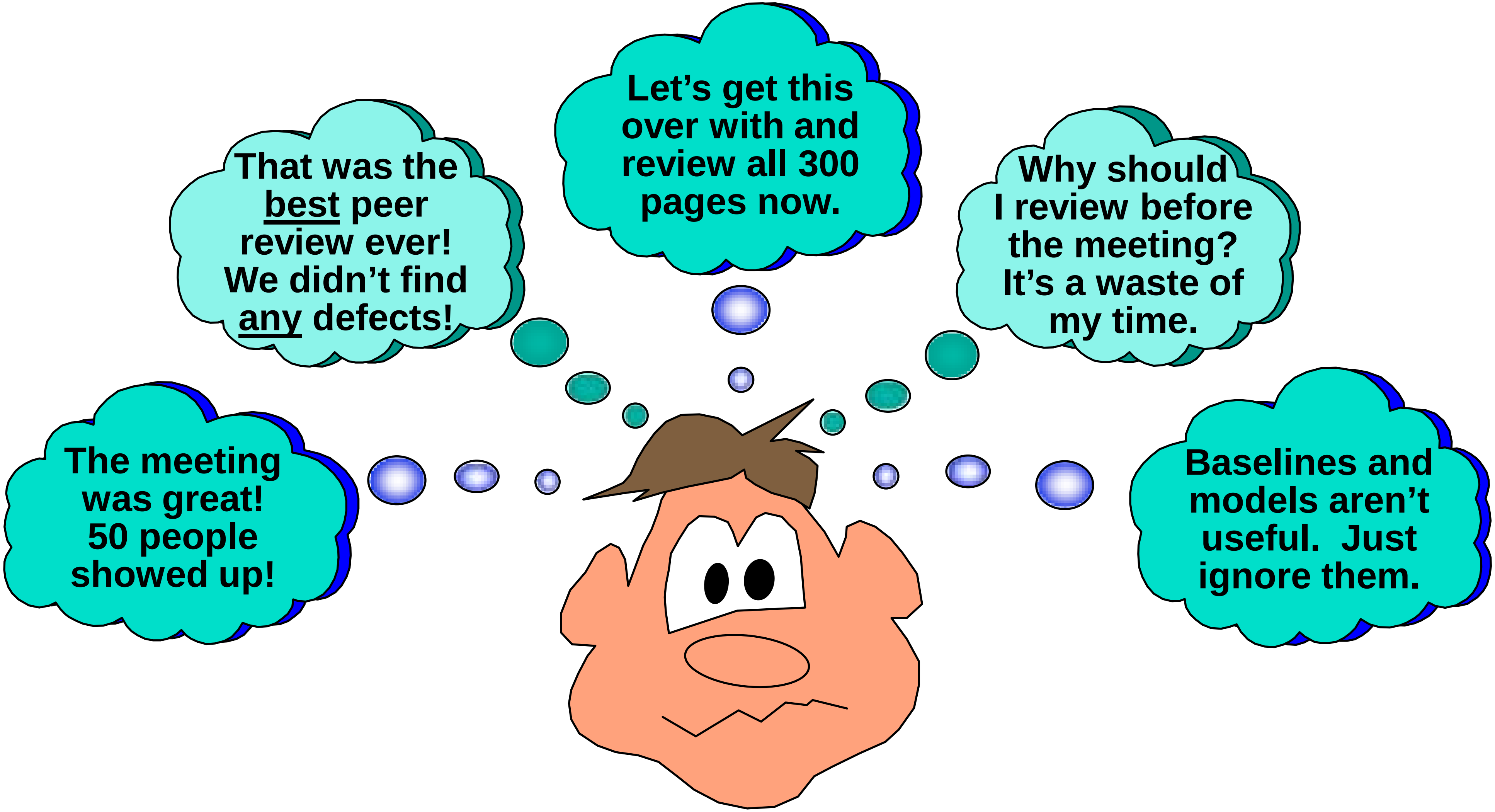
Products and Services

- Mission support
- Cybersecurity
- Command, control, and communications
- Enterprise applications
- IT & network infrastructure
- Management & engineering services
- Intelligence, surveillance, & reconnaissance

CMMI Appraisals

- Over 80 organizations (over 250 projects) appraised at Level 3 or higher

Why Was This Important to Us? (Goals)


That was the best peer review ever! We didn't find any defects!
Let's get this over with and review all 300 pages now.
Why should I review before the meeting? It's a waste of my time.
The meeting was great! 50 people showed up!
Baselines and models aren't useful. Just ignore them.

Goal was to change the way people think!

Peer Review Data

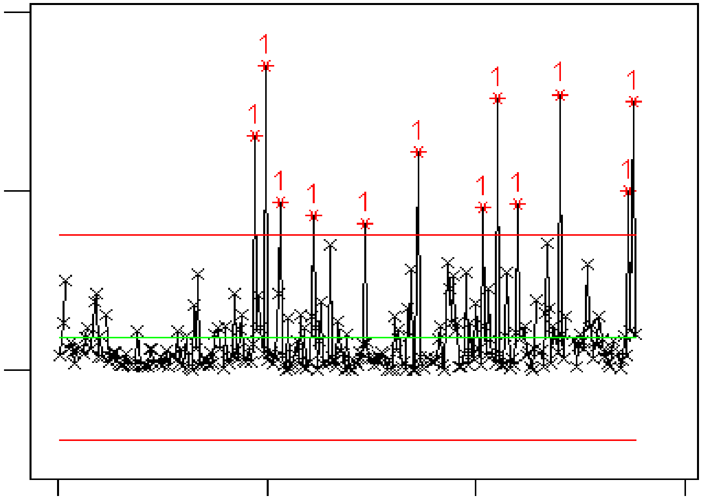
- 5 years of data from April 2003 through December 2008
- 1,860 peer reviews and 11,166 action items/defects

608	**Pages**
395	**Test Cases**
352	**Shalls**
276	**SLOCs**
123	None
85	VI
21	Nodes

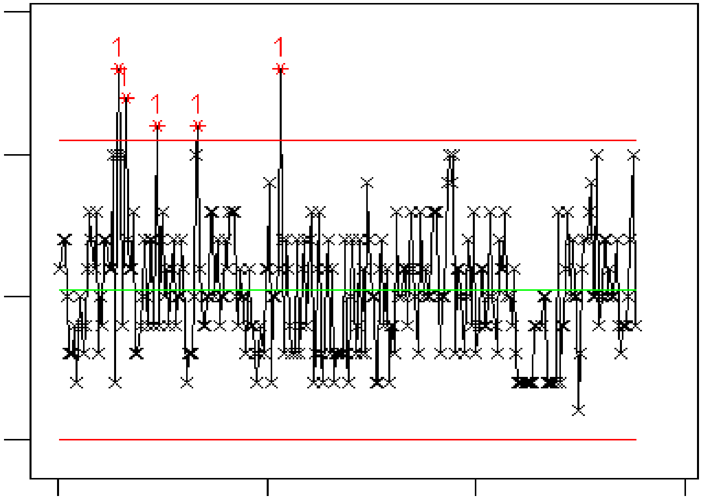
Created baselines and models for requirements (shalls), design (pages), code (SLOCs), and test (test cases); however, this presentation only focuses on SLOCs

Deleted Out of Control Points

Number of SLOCs (Size)



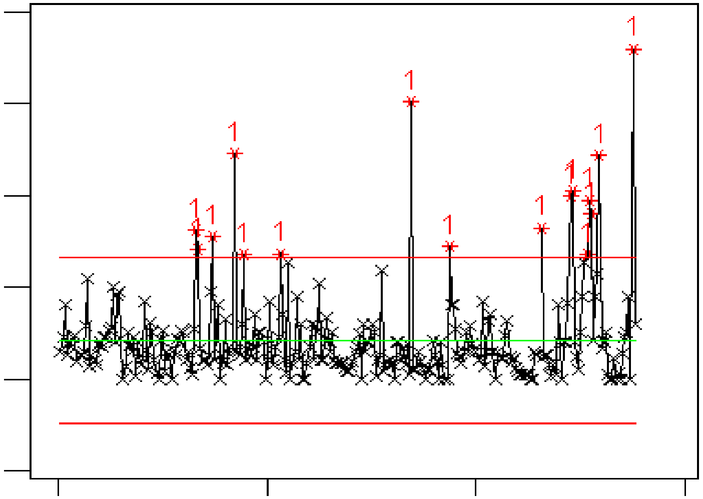
Number of Attendees



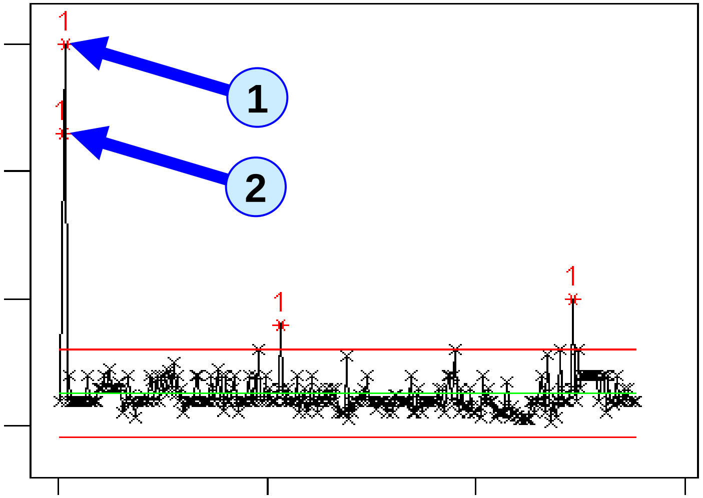
Remove invalid data, not necessarily out of control data, or they will corrupt the regression equation for the model. Only data that are clearly invalid were removed.

Deleted 2 out of 276 SLOC peer reviews.

Pre-Review Hours



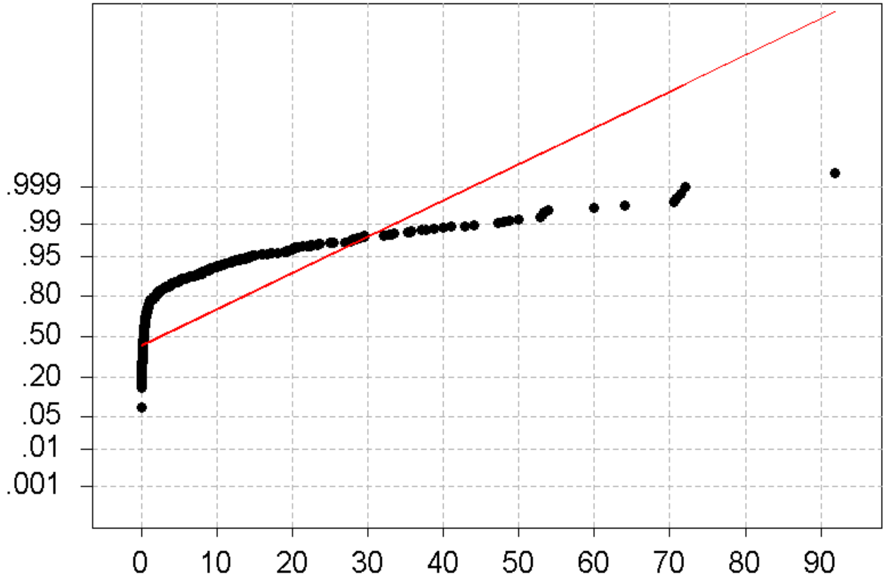
Meeting Hours



Converted to Lognormal Data

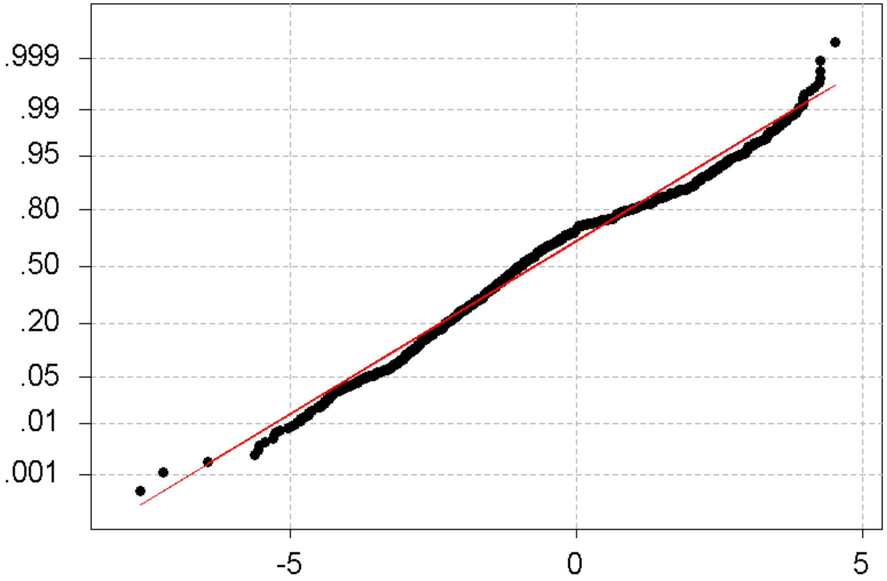
- Used **Normality Tests** to verify whether the data is normal. Data must be normal for regression equations (models).
- When data is not normal, convert to lognormal data using **LN(Data)**

Data is not normal if the Normality Test shows points are not on the line.



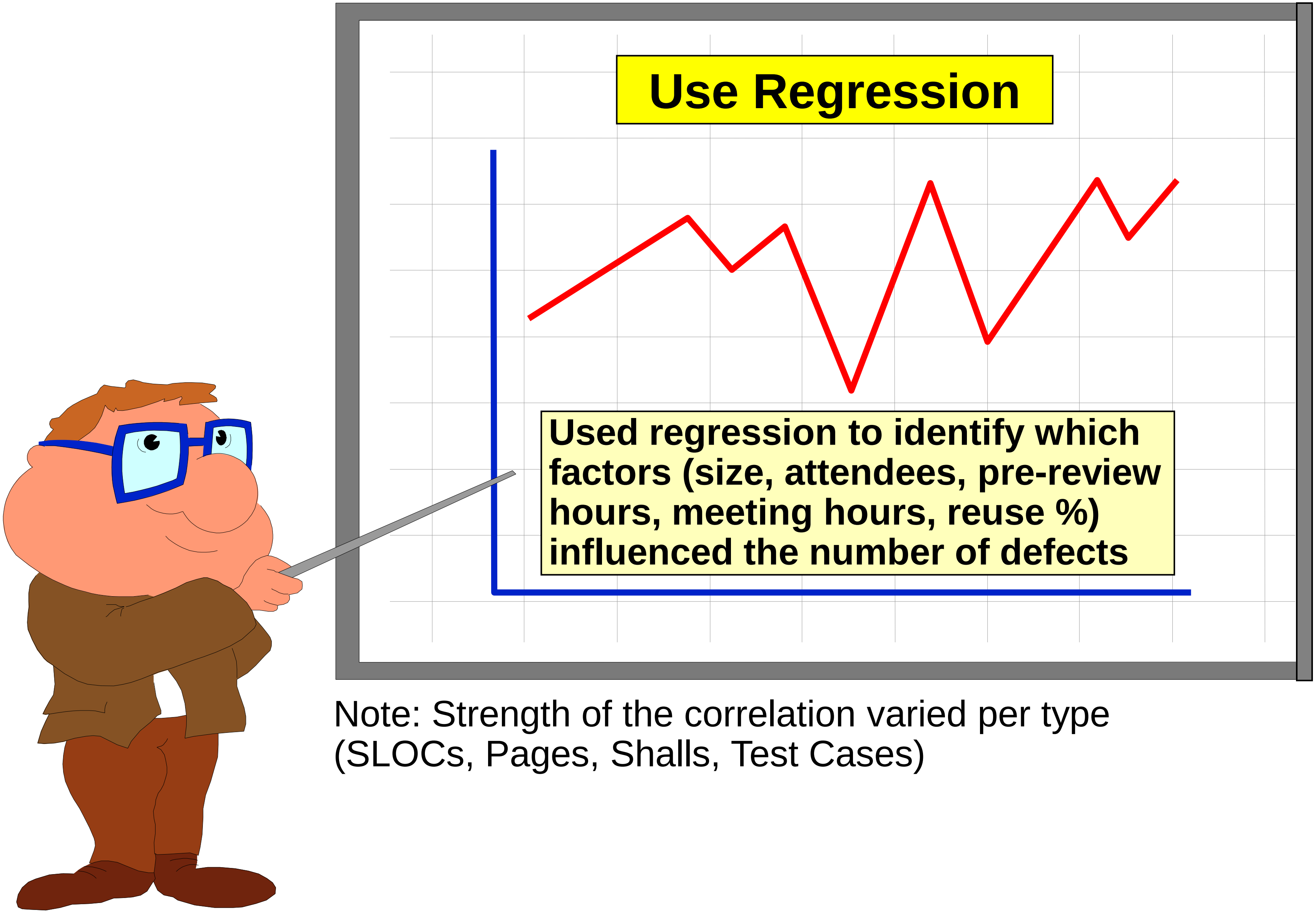
Actual Defect Density Data

After data is converted, the Normality Test shows points are on the line.



After LN(Defect Density) Conversion

Checked Strength of Correlation

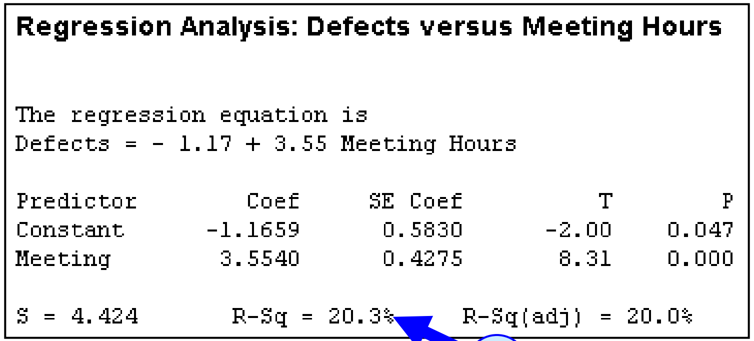


Use Regression

Used regression to identify which factors (size, attendees, pre-review hours, meeting hours, reuse %) influenced the number of defects

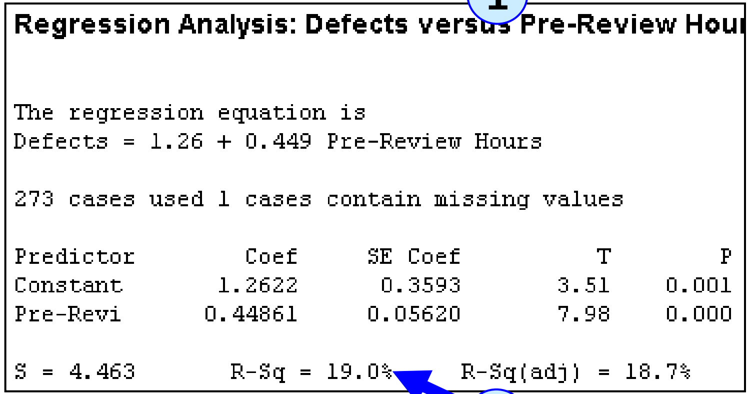
Note: Strength of the correlation varied per type
(SLOCs, Pages, Shalls, Test Cases)

Strength of Correlation



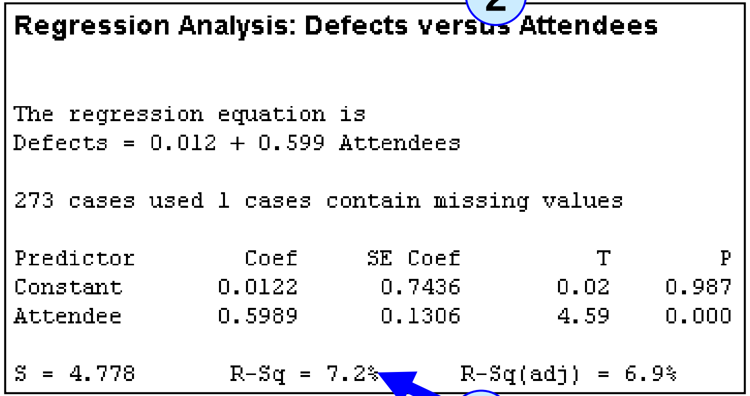
```
Regression Analysis: Defects versus Meeting Hours


The regression equation is
Defects = - 1.17 + 3.55 Meeting Hours

Predictor          Coef     SE Coef          T        P
Constant        -1.1659      0.5830      -2.00    0.047
Meeting          3.5540      0.4275       8.31    0.000

S = 4.424         R-Sq = 20.3%      R-Sq(adj) = 20.0%
```

1



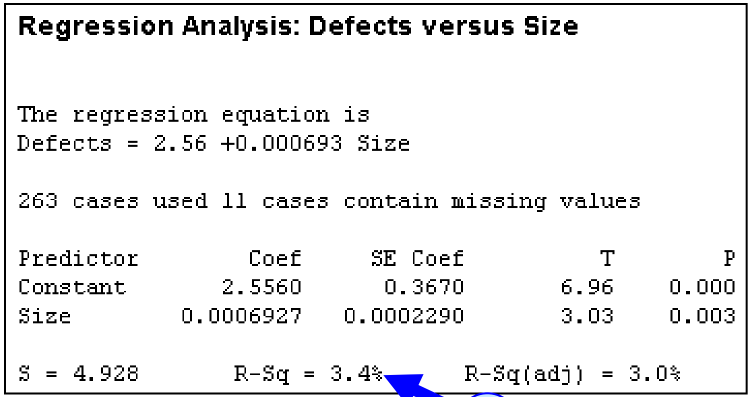
```
Regression Analysis: Defects versus Pre-Review Hour


The regression equation is
Defects = 1.26 + 0.449 Pre-Review Hours

273 cases used 1 cases contain missing values

Predictor          Coef     SE Coef          T        P
Constant         1.2622      0.3593       3.51    0.001
Pre-Revi        0.44861     0.05620       7.98    0.000

S = 4.463         R-Sq = 19.0%      R-Sq(adj) = 18.7%
```

2



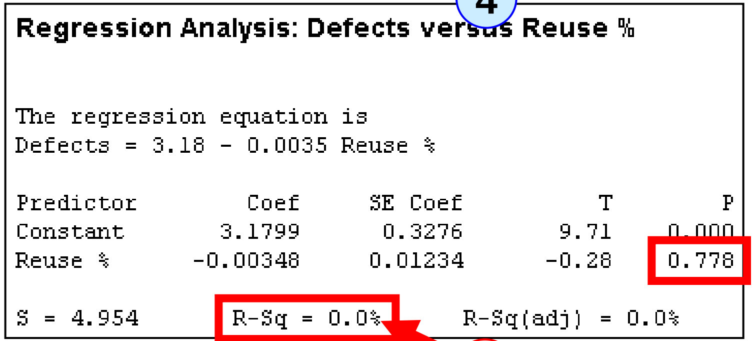
```
Regression Analysis: Defects versus Attendees


The regression equation is
Defects = 0.012 + 0.599 Attendees

273 cases used 1 cases contain missing values

Predictor          Coef     SE Coef          T        P
Constant         0.0122      0.7436       0.02    0.987
Attendee         0.5989      0.1306       4.59    0.000

S = 4.778         R-Sq = 7.2%       R-Sq(adj) = 6.9%
```

3



```
Regression Analysis: Defects versus Size


The regression equation is
Defects = 2.56 +0.000693 Size

263 cases used 11 cases contain missing values

Predictor          Coef     SE Coef          T        P
Constant         2.5560      0.3670       6.96    0.000
Size          0.0006927   0.0002290       3.03    0.003

S = 4.928         R-Sq = 3.4%       R-Sq(adj) = 3.0%
```

4



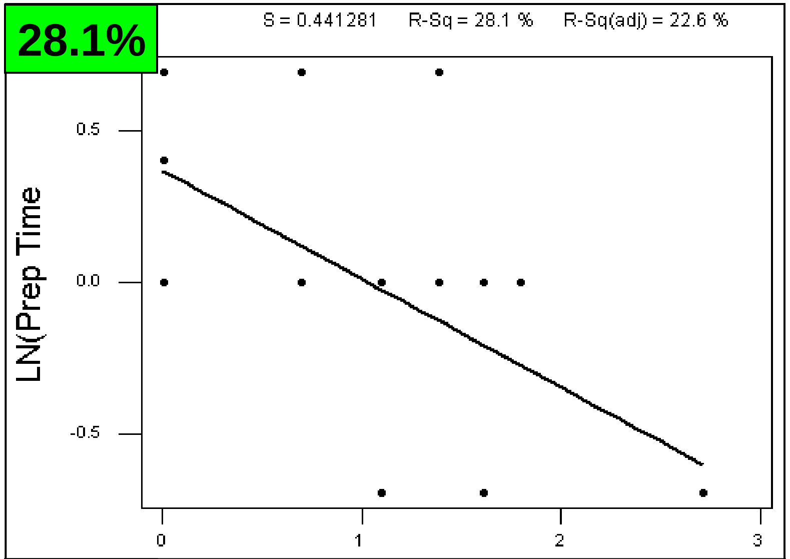
```
Regression Analysis: Defects versus Reuse %


The regression equation is
Defects = 3.18 - 0.0035 Reuse %

Predictor          Coef     SE Coef          T        P
Constant         3.1799      0.3276       9.71    0.000
Reuse %        -0.00348     0.01234      -0.28    0.778

S = 4.954         R-Sq = 0.0%       R-Sq(adj) = 0.0%
```

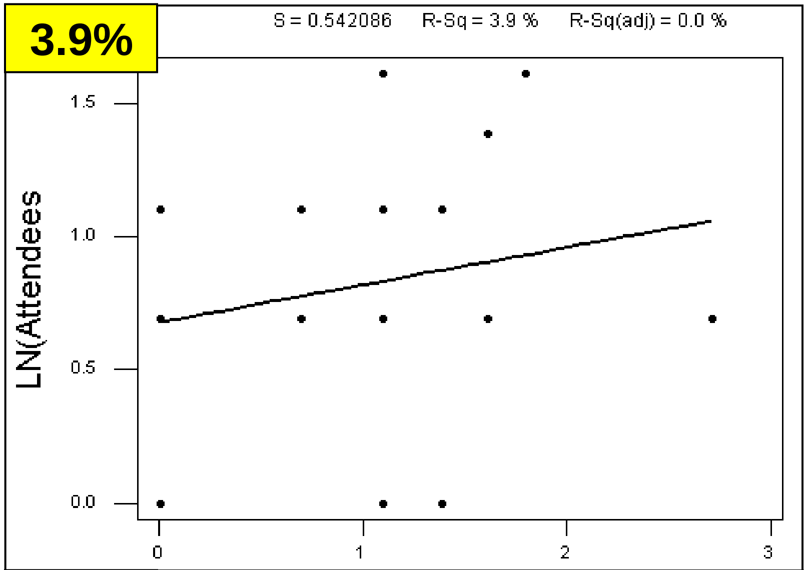
5

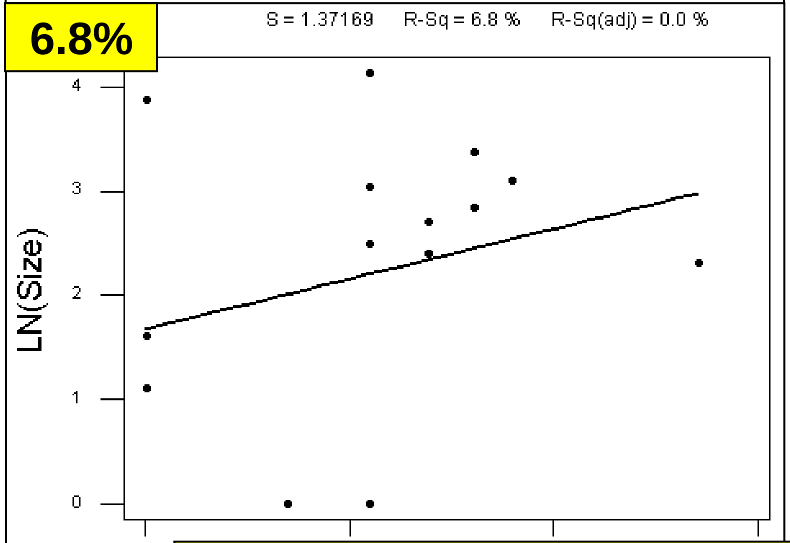
Conclusion: No correlation for Reuse %

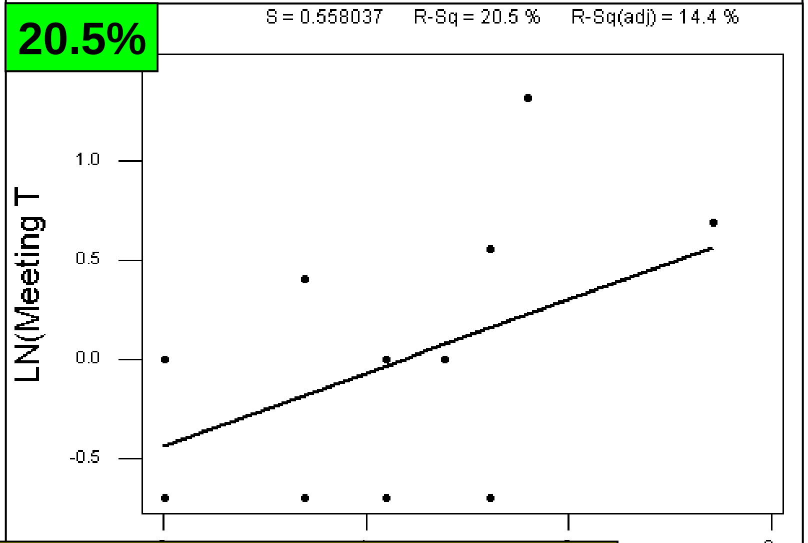
1. Meeting Hours	20.3%
2. Pre-Review Hours	19.0%
3. Attendees	7.2%
4. Size	3.4%
5. Reuse	0.0% (also P-value is high)

Strength of Correlation (another organization)









Conclusion: Consistent results even with other organizations. Strongest correlation is Meeting Hours and Pre-Review Hours.

Strength of Correlation Summary

- Green = P-value = 0.00 Strong correlation
- Red = P-value > 0.05 No correlation

	SLOCs	Pages	Shalls	Test Cases

Size	R-Sq=3.4%
	P-value=0.003
		R-Sq=1.2%
		P-value=0.006
			R-Sq=0.0%
			P-value=0.898
				R-Sq=1.5%
				P-value=0.020

Attendees	R-Sq=7.2%
	P-value=0.000
		R-Sq=1.2%
		P-value=0.006
			R-Sq=11.2%
			P-value=0.000
				R-Sq=8.8%
				P-value=0.000

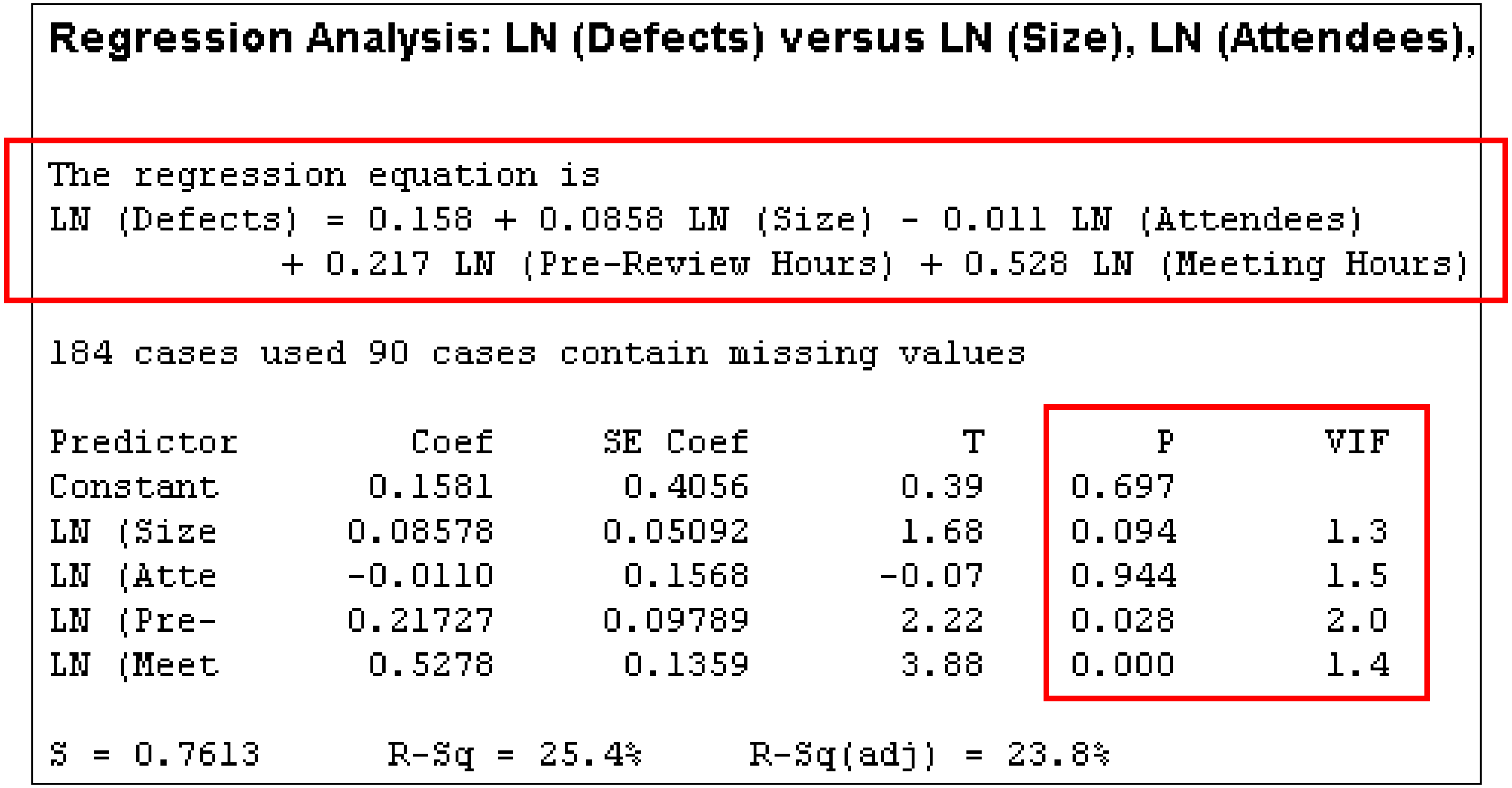
Pre-Review Hours	R-Sq=19.0%
	P-value=0.000
		R-Sq=3.6%
		P-value=0.000
			R-Sq=0.0%
			P-value=0.778
				R-Sq=3.9%
				P-value=0.000

Meeting Hours	R-Sq=20.3%
	P-value=0.000
		R-Sq=3.3%
		P-value=0.000
			R-Sq=9.4%
			P-value=0.000
				R-Sq=17.6%
				P-value=0.000

Reuse %	R-Sq=0.0%
	P-value=0.778
		R-Sq=1.0%
		P-value=0.013
			R-Sq=0.0%
			P-value=0.735
				R-Sq=0.5%
				P-value=0.169

Conclusion: Table easily shows which X factors should be used for the SLOCs, Pages, Shalls, and Test Cases models and which should be discarded. Don’t include Reuse % just because your gut instinct tells you to.

Regression Equation for Model



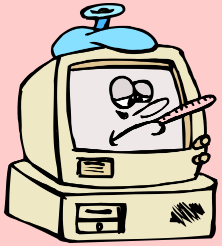
Regression Analysis: LN (Defects) versus LN (Size), LN (Attendees),

```
The regression equation is
LN (Defects) = 0.158 + 0.0858 LN (Size) - 0.011 LN (Attendees)
           + 0.217 LN (Pre-Review Hours) + 0.528 LN (Meeting Hours)

184 cases used 90 cases contain missing values

Predictor        Coef     SE Coef          T        P       VIF
Constant       0.1581      0.4056       0.39    0.697
LN (Size      0.08578     0.05092       1.68    0.094       1.3
LN (Atte      -0.0110      0.1568      -0.07    0.944       1.5
LN (Pre-      0.21727     0.09789       2.22    0.028       2.0
LN (Meet       0.5278      0.1359       3.88    0.000       1.4

S = 0.7613      R-Sq = 25.4%     R-Sq(adj) = 23.8%
```

Note: VIF > 5 means if you include that factor in the equation, it will distort the results, i.e., inflate the results

Conclusion: **Attendees had a large P-value; however, Variance Inflation Factor (VIF) is < 5 so using all the X factors should be okay in the regression equation.**

Final X Factors and Y Outcome



Defects

Y

Number of defects and defect density

Size

X

Choose how much to peer review, e.g., choose to peer review 200 SLOCs

Attendees

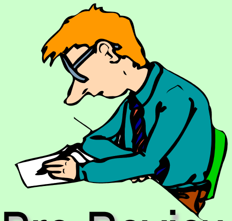
X

Choose how many people to invite to the peer review, e.g., choose to only invite 3 people

Meeting Hours

X

Choose how long to schedule the meeting, e.g., choose a 1 hour meeting



Pre-Review Hours

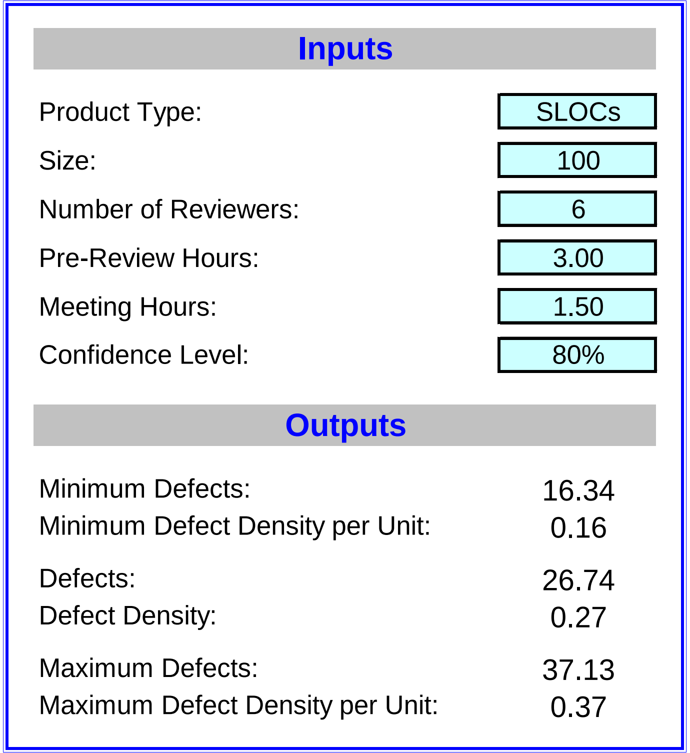
X

Choose minimum hours to review prior to the meeting (most hours spent by a reviewer, not the total number of hours)

Peer Review Model

- Model is deterministic, i.e., provides a single value, and probabilistic, i.e., provides a range of values (80% confidence interval)
- Confidence intervals in Excel are very complicated

Keep it Simple Stupid
(even a child can understand it)



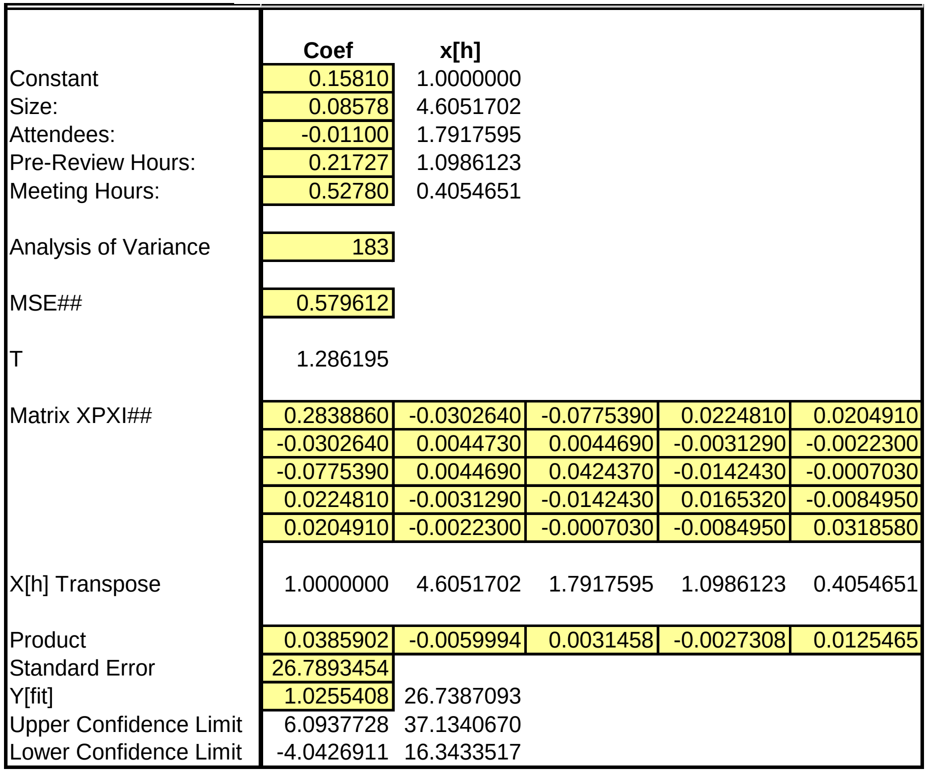
Inputs

Product Type:	SLOCs
Size:	100
Number of Reviewers:	6
Pre-Review Hours:	3.00
Meeting Hours:	1.50
Confidence Level:	80%

Outputs

Minimum Defects:	16.34
Minimum Defect Density per Unit:	0.16
Defects:	26.74
Defect Density:	0.27
Maximum Defects:	37.13
Maximum Defect Density per Unit:	0.37

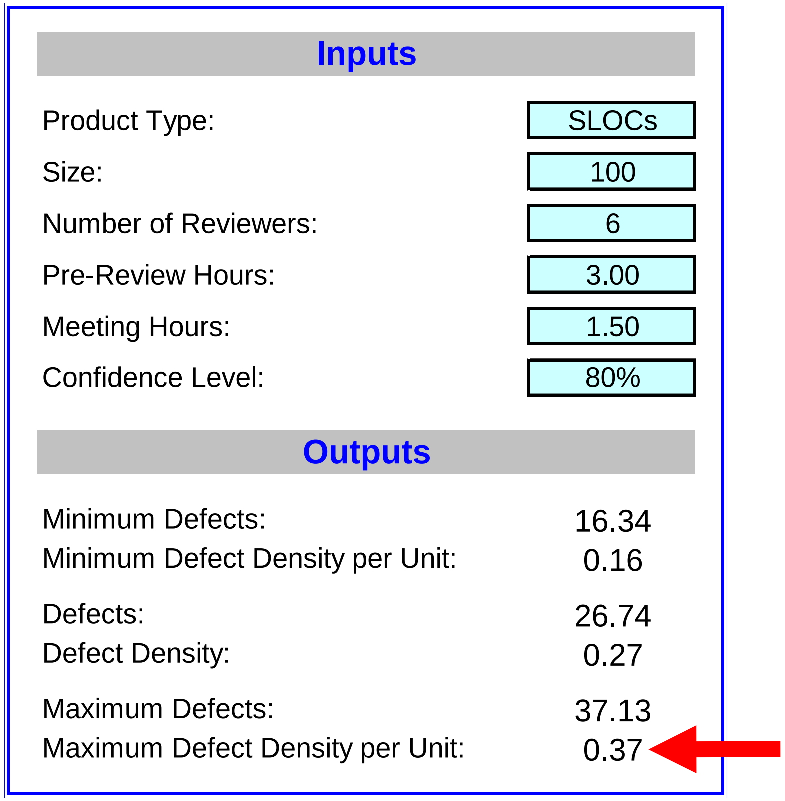
Hide the Intelligence
(hide complexity from the user)



	Coef	x[h]			
Constant	0.15810	1.0000000			
Size:	0.08578	4.6051702			
Attendees:	-0.01100	1.7917595			
Pre-Review Hours:	0.21727	1.0986123			
Meeting Hours:	0.52780	0.4054651			
Analysis of Variance	183				
MSE##	0.579612				
T	1.286195				
Matrix XPXI##	0.2838860	-0.0302640	-0.0775390	0.0224810	0.0204910
	-0.0302640	0.0044730	0.0044690	-0.0031290	-0.0022300
	-0.0775390	0.0044690	0.0424370	-0.0142430	-0.0007030
	0.0224810	-0.0031290	-0.0142430	0.0165320	-0.0084950
	0.0204910	-0.0022300	-0.0007030	-0.0084950	0.0318580
X[h] Transpose	1.0000000	4.6051702	1.7917595	1.0986123	0.4054651
Product	0.0385902	-0.0059994	0.0031458	-0.0027308	0.0125465
Standard Error	26.7893454				
Y[fit]	1.0255408	26.7387093			
Upper Confidence Limit	6.0937728	37.1340670			
Lower Confidence Limit	-4.0426911	16.3433517			

How Projects Should Use the Model

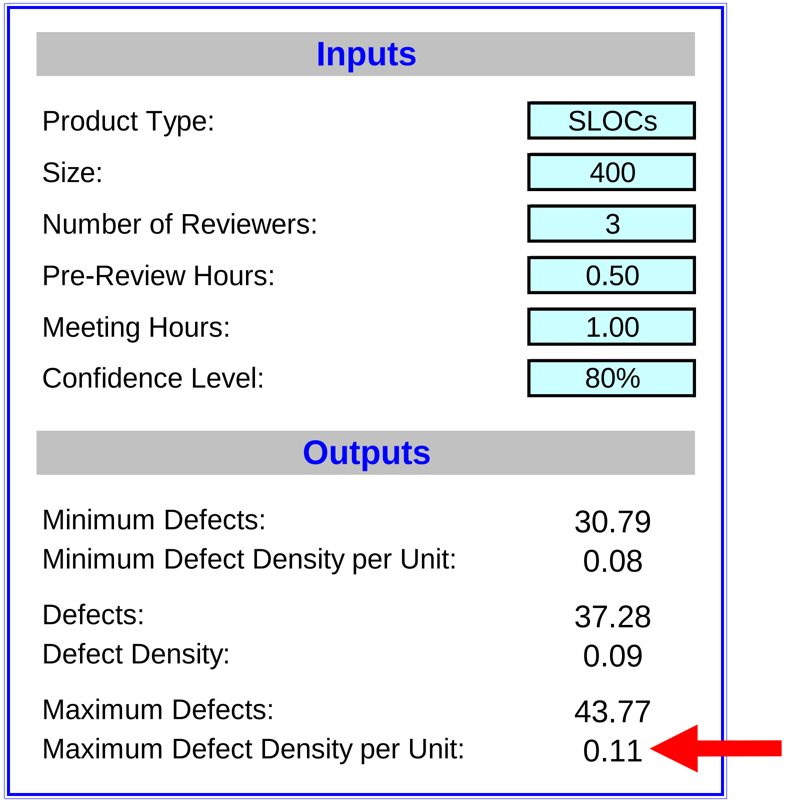
Effective Review



Inputs	
Product Type:	SLOCs
Size:	100
Number of Reviewers:	6
Pre-Review Hours:	3.00
Meeting Hours:	1.50
Confidence Level:	80%

Outputs	
Minimum Defects:	16.34
Minimum Defect Density per Unit:	0.16
Defects:	26.74
Defect Density:	0.27
Maximum Defects:	37.13
Maximum Defect Density per Unit:	0.37

Not as Effective Review



Inputs	
Product Type:	SLOCs
Size:	400
Number of Reviewers:	3
Pre-Review Hours:	0.50
Meeting Hours:	1.00
Confidence Level:	80%

Outputs	
Minimum Defects:	30.79
Minimum Defect Density per Unit:	0.08
Defects:	37.28
Defect Density:	0.09
Maximum Defects:	43.77
Maximum Defect Density per Unit:	0.11

- **Peer Review Planning**
  Do “what-if” analysis with the controllable factors to determine optimal settings. Use different settings depending on cost and schedule constraints, critical high risk products, etc.

- **After Peer Review is Completed**
  Enter actual data and see if results are > minimum. If < minimum, consider another peer review if the peer review was ineffective.

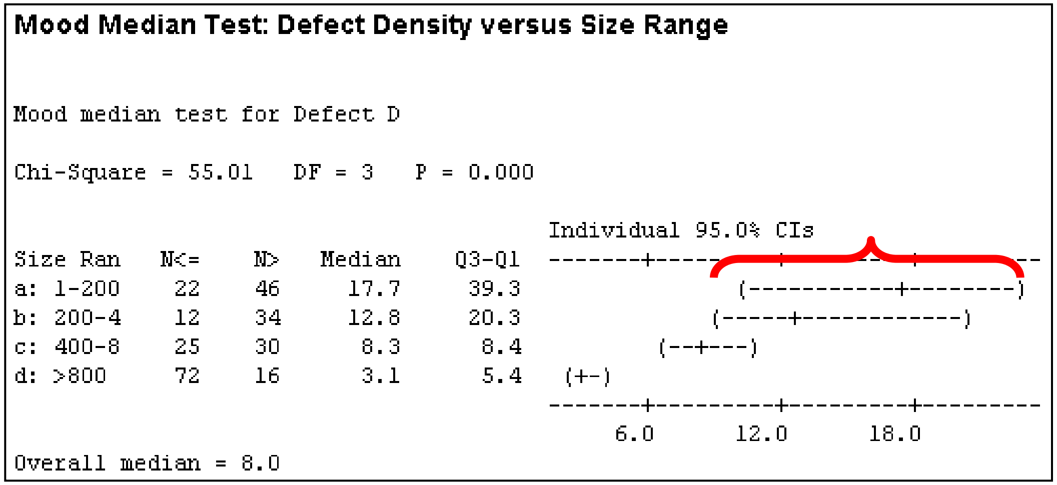
Full Circle - Used OPP for OID

- OPP analysis uncovered “sweet spots” where peer reviews were more effective, i.e., Defect Density was higher
- Identified “Sweet spots” for:
  - Size
  - Attendees
  - Meeting Hours
  - Pre-Review Hours
- "Best Kept Secrets of Peer Code Review" textbook by Jason Cohen, "LOC under review should be under 200; not to exceed 400.”
- Determine whether constraining peer reviews to the “sweet spots” will consistently result in higher quality peer reviews
- If Defect Density is consistently higher, modify the standard process to recommend the “sweet spots”

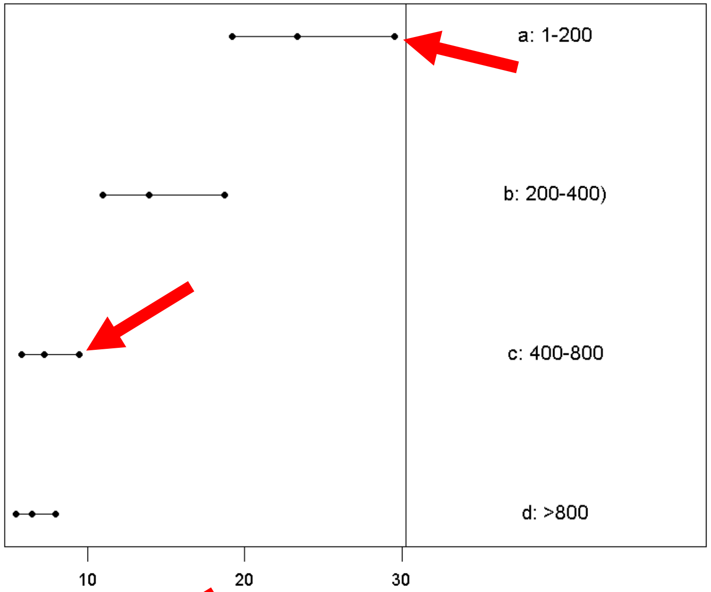
Quality and Process Performance Goals

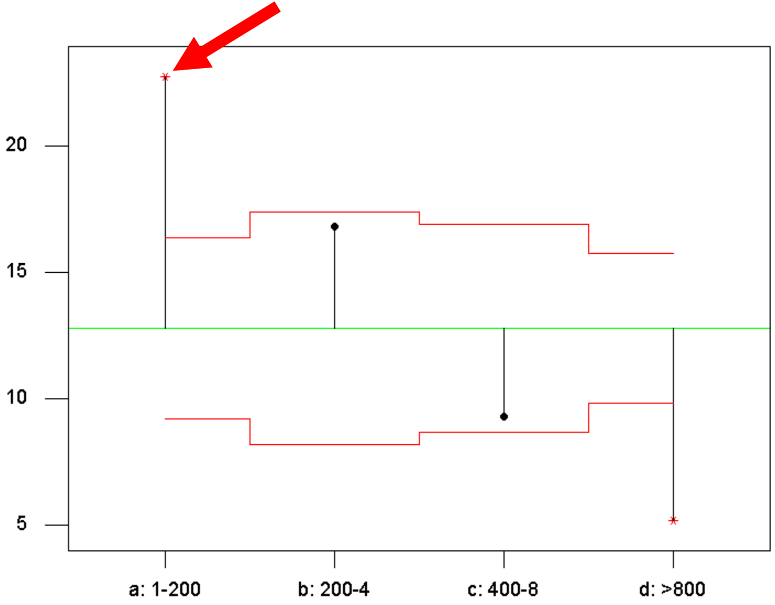
- Goal for process performance is to improve the efficiency of code peer reviews, i.e., more cost effective
  - Too many reviewers do not improve Defect Density
  - Long meetings do not improve Defect Density
- Goal for quality performance is to improve Defect Density
  - Less SLOCs increases Defect Density
  - Adequate preparation increases Defect Density

What is the “Sweet Spot” for Size



```
Mood Median Test: Defect Density versus Size Range

Mood median test for Defect D

Chi-Square = 55.01   DF = 3   P = 0.000

                                               Individual 95.0% CIs
Size Ran   N<=   N>   Median   Q3-Q1   -------+---------+---------+---------
a: 1-200    22   46     17.7    39.3              (-----------+--------)
b: 200-4    12   34     12.8    20.3            (-----+------------)
c: 400-8    25   30      8.3     8.4        (--+---)
d: >800     72   16      3.1     5.4    (+-)
                                       -------+---------+---------+---------
                                            6.0      12.0      18.0

Overall median = 8.0
```



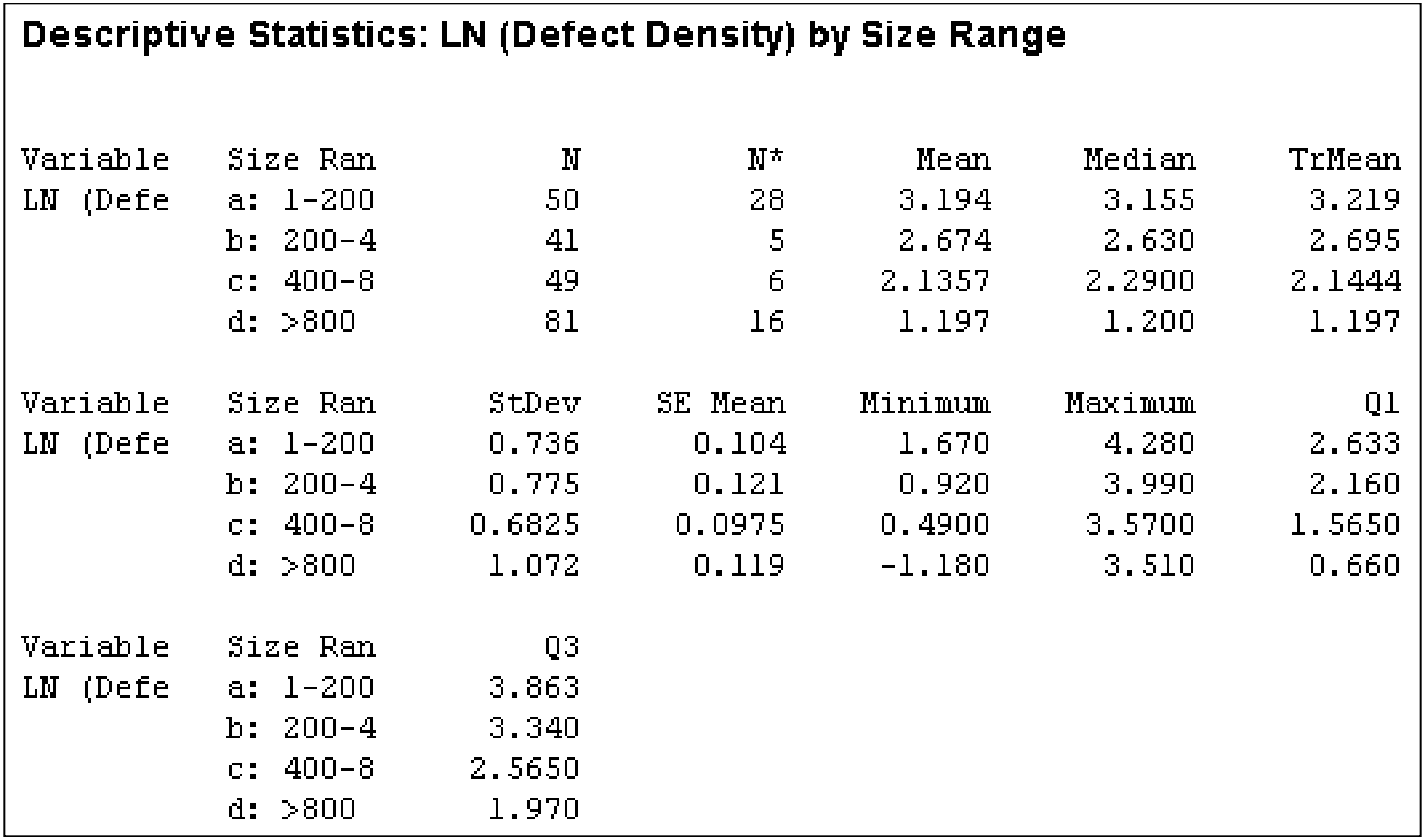


Conclusion: Recommend 1 to 400 SLOCs, preferably 1 to 200 SLOCs. Never review >= 400 SLOCs.

Textbook is correct !!!

Did the same “sweet spot” analysis for attendees, pre-review hours, and meeting hours.

Provided Baselines for Size



Descriptive Statistics: LN (Defect Density) by Size Range

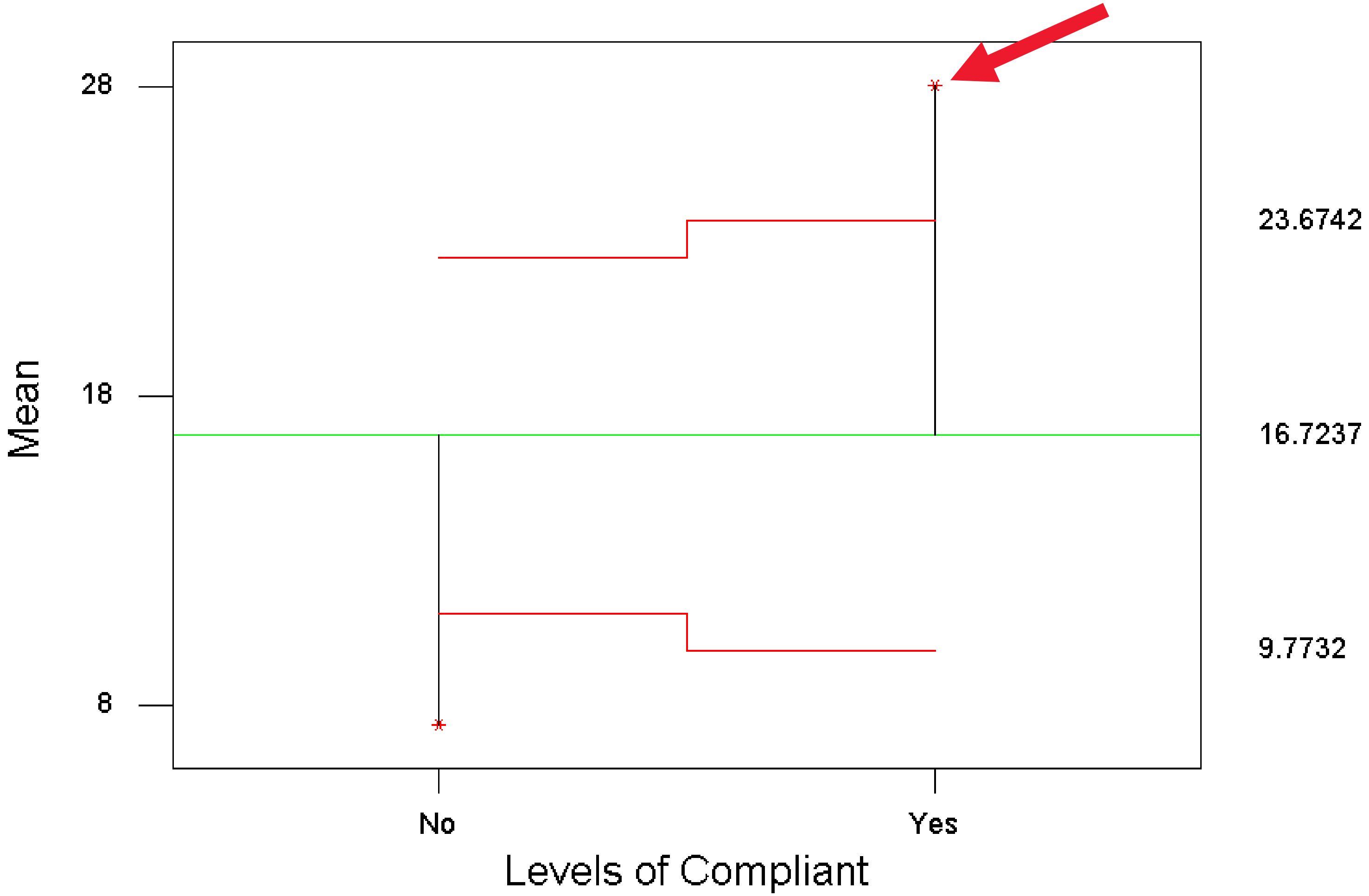
```
Variable    Size Ran        N       N*     Mean   Median   TrMean
LN (Defe    a: 1-200       50       28    3.194    3.155    3.219
            b: 200-4       41        5    2.674    2.630    2.695
            c: 400-8       49        6   2.1357   2.2900   2.1444
            d: >800        81       16    1.197    1.200    1.197

Variable    Size Ran    StDev  SE Mean  Minimum  Maximum       Q1
LN (Defe    a: 1-200    0.736    0.104    1.670    4.280    2.633
            b: 200-4    0.775    0.121    0.920    3.990    2.160
            c: 400-8   0.6825   0.0975   0.4900   3.5700   1.5650
            d: >800     1.072    0.119   -1.180    3.510    0.660

Variable    Size Ran       Q3
LN (Defe    a: 1-200    3.863
            b: 200-4    3.340
            c: 400-8   2.5650
            d: >800     1.970
```

Provided the same baselines for attendees, pre-review hours, and meeting hours.

OID for Constrained Peer Reviews Pilot

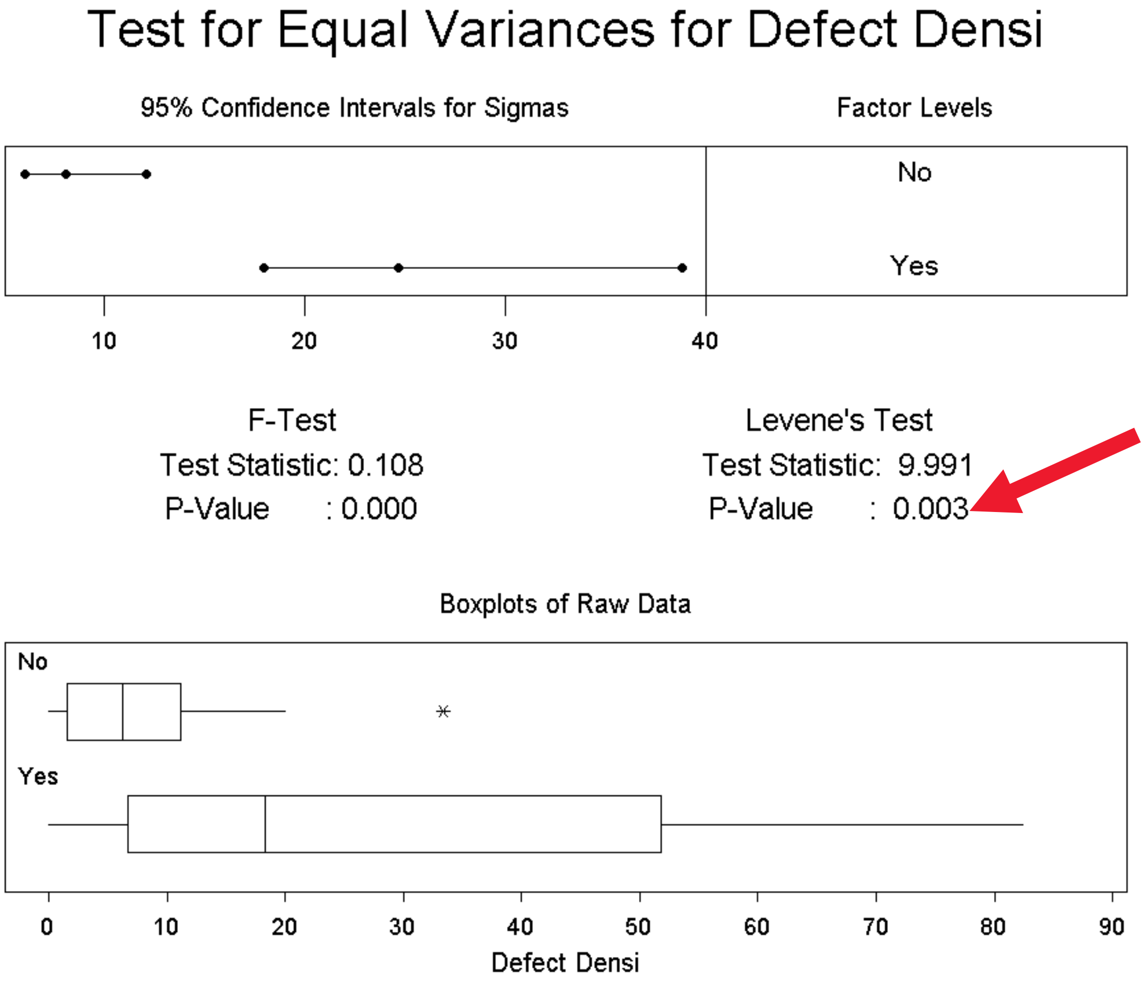
- Briefed all software projects on the “sweet spots”
- “Sweet spots” were provided for size, meeting hours, pre-review hours, and number of attendees
- 19 peer reviews were 100% constrained, i.e., used all “sweet spots”
- 23 peer reviews did whatever they felt was appropriate, and did not use all “sweet spots”
- Used multiple Hypothesis Tests to compare Defect Density for constrained (19 peer reviews) versus non-constrained (23 peer reviews)

Defect Density Mean



Conclusion: Defect Density mean for constrained peer reviews was statistically significantly higher. A set of constrained peer reviews will always have a higher Defect Density mean.

Defect Density Variation



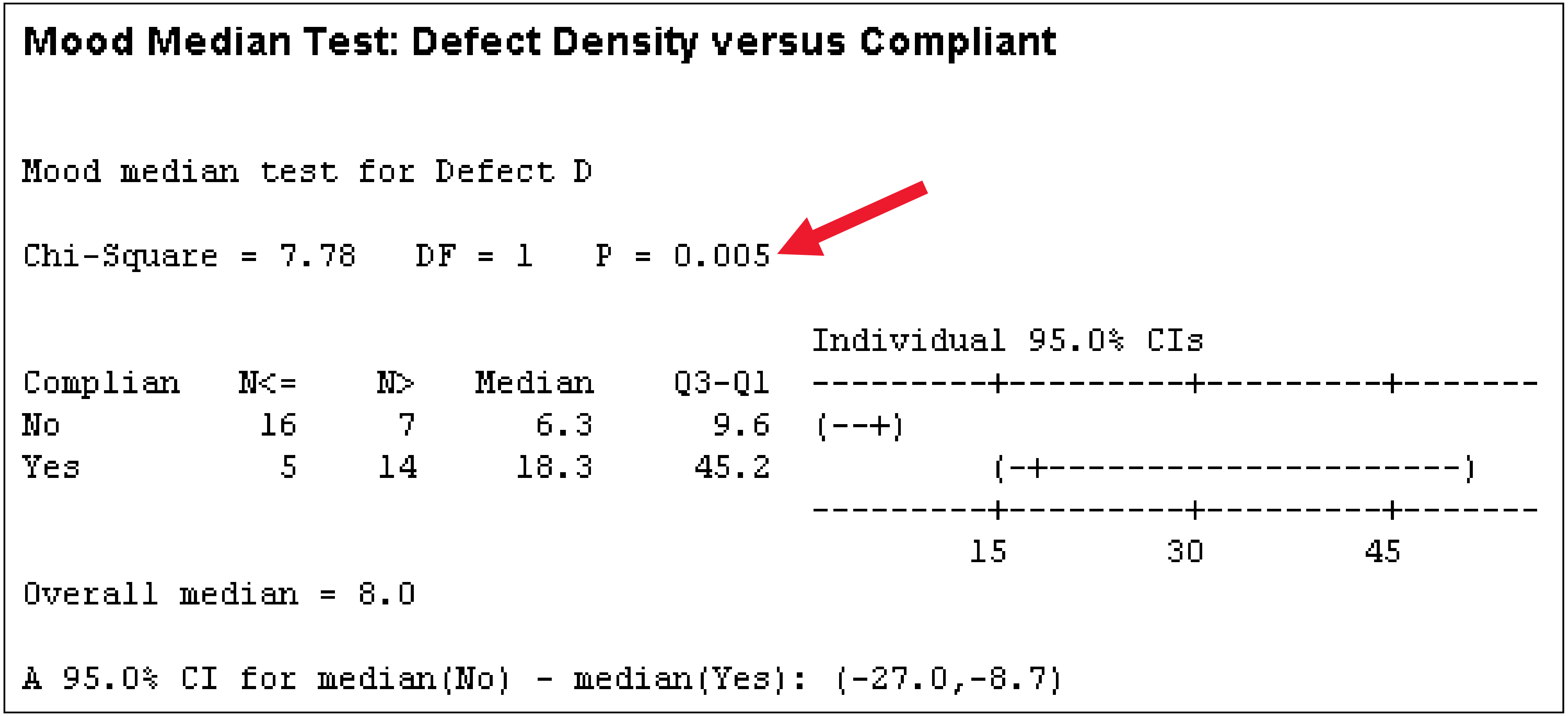
Test for Equal Variances for Defect Densi

F-Test
Test Statistic: 0.108
P-Value : 0.000

Levene's Test
Test Statistic: 9.991
P-Value : 0.003

Conclusion: Test for Equal Variance hypothesis test shows the variation is statistically significantly different (P-Value < 0.05). Unconstrained peer reviews were consistently poorer.

Defect Density Median

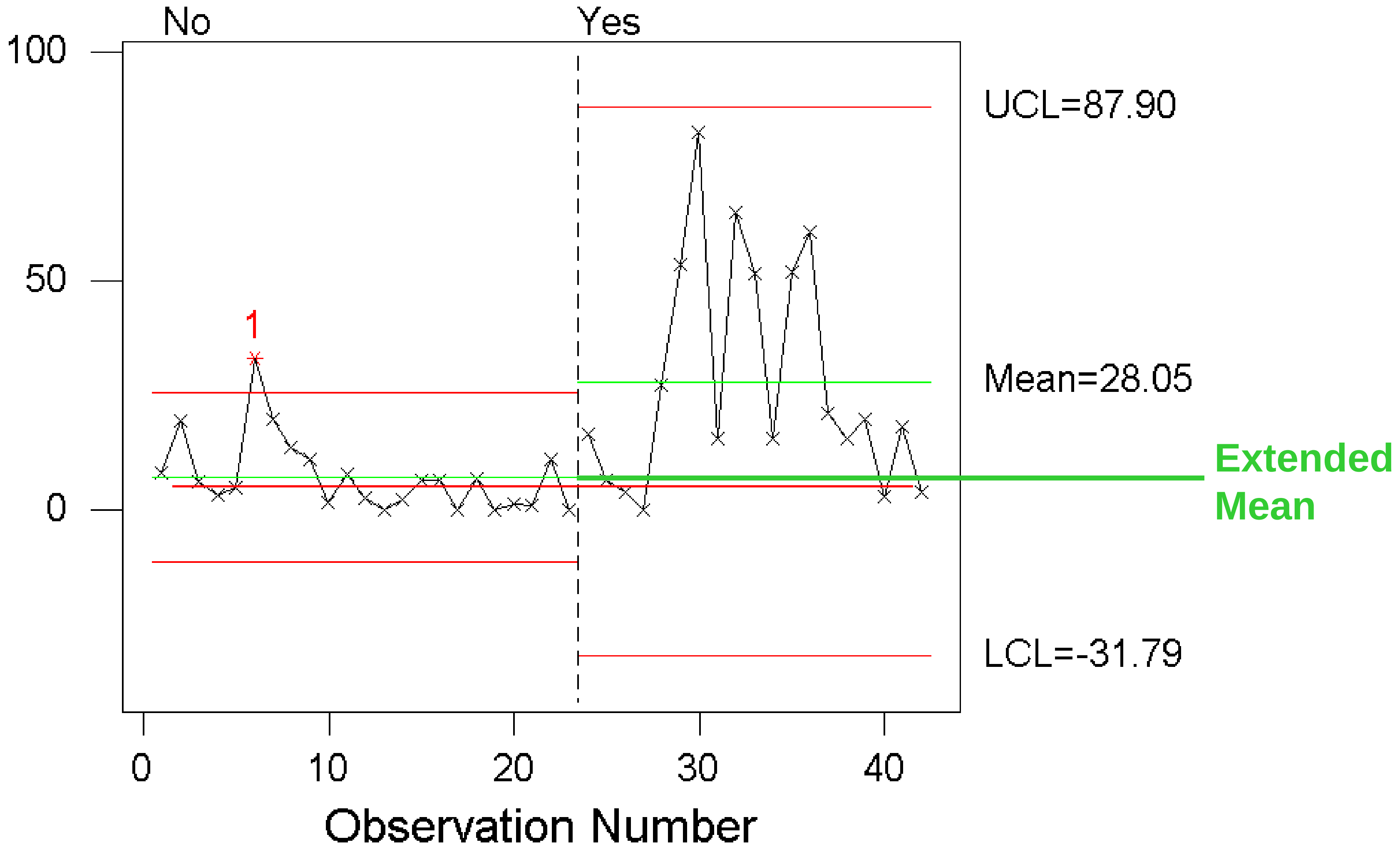


Mood Median Test: Defect Density versus Compliant

```
Mood median test for Defect D

Chi-Square = 7.78   DF = 1   P = 0.005

                                             Individual 95.0% CIs
Complian    N<=     N>    Median     Q3-Q1   ---------+---------+---------+-------
No           16      7       6.3       9.6   (--+)
Yes           5     14      18.3      45.2            (-+--------------------)
                                             ---------+---------+---------+-------
                                                     15        30        45

Overall median = 8.0

A 95.0% CI for median(No) - median(Yes): (-27.0,-8.7)
```

Conclusion: Mood Median hypothesis test shows the median is statistically significantly different (P-Value > 0.05). A set of constrained peer reviews will always have a higher Defect Density median.

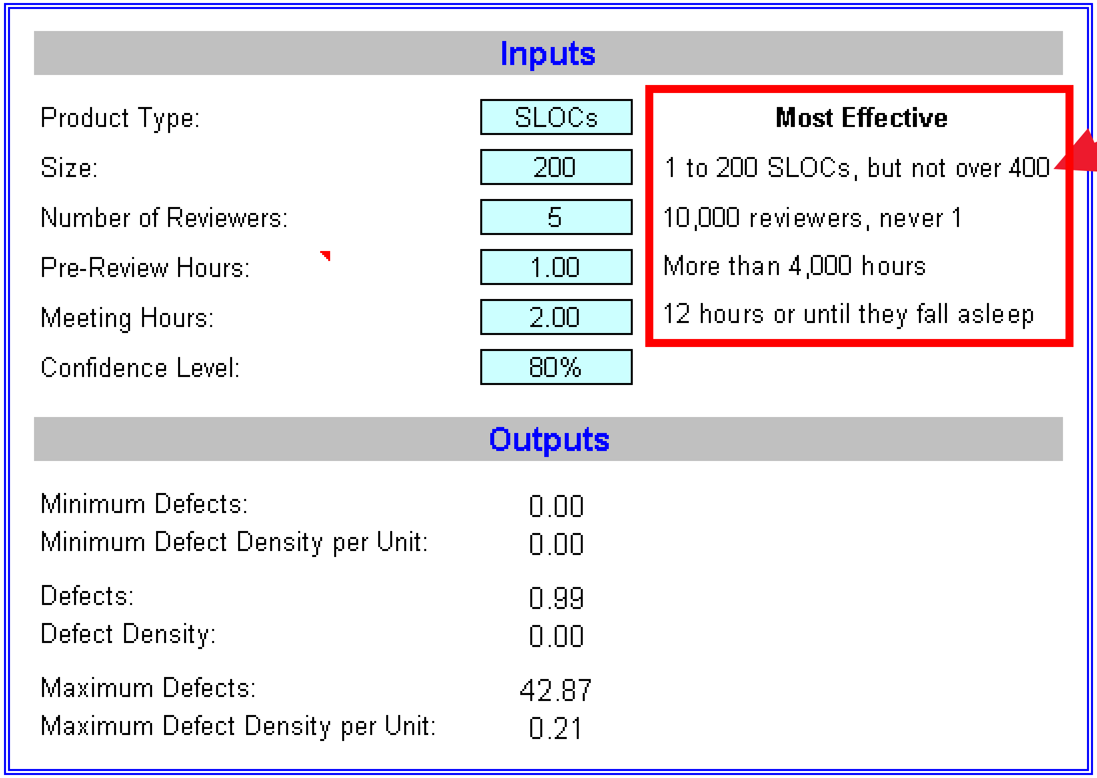
Defect Density Control Chart



Conclusion: Only 4 of the 19 constrained peer reviews were below the mean for the unconstrained peer reviews.

Full Circle OID Improvement Back to Projects

- Pilots showed conclusively that constraining peer reviews to “sweet spots” significantly improves Defect Density
- The model was modified to add “sweet spots” (Most Effective)
- OID was used to improve project performance using OPP analysis



Inputs

Product Type:	SLOCs	**Most Effective**
Size:	200	1 to 200 SLOCs, but not over 400
Number of Reviewers:	5	10,000 reviewers, never 1
Pre-Review Hours:	1.00	More than 4,000 hours
Meeting Hours:	2.00	12 hours or until they fall asleep
Confidence Level:	80%	

Outputs

Minimum Defects:	0.00
Minimum Defect Density per Unit:	0.00
Defects:	0.99
Defect Density:	0.00
Maximum Defects:	42.87
Maximum Defect Density per Unit:	0.21

Only real one, others are a joke

Summary

- Analyzing data can identify X factors to use and X factors to discard
- Projects need to understand how to use the model
- Use OID to improve the model and project performance
- Publisher approved writing a textbook that will be called, “Baselines and Models, Duh, I Don’t Get It” (taking the train to the airport from a previous conference presentation) or “Baselines and Models for CMMI Process Improvement Practitioners”. Manuscript is due to the Publisher by May 2010, for publishing later in 2010. Textbook will contain an expanded version of taking the train to the airport, an expanded version of this peer review presentation, and a different way of estimating hours that will help projects perform better.

Diane.Mizukami@ngc.com, 310-921-1939