Presented at the 2007 ISPA/SCEA Joint Annual International Conference and Workshop - www.iceaaonline.com

NORTHROP GRUMMAN

DEFINING THE FUTURE

Sensitivity Analysis of Commercial Cost Estimating Tools to the CSCI Structure

SCEA 2007
June 12-15
New Orleans, LA

Abbey Turnau
Operations Researcher
Northrop Grumman Corporation

SCEA
The Society of Cost Estimating and Analysis

Presented at the 2007 ISPA/SCEA Joint Annual International Conference and Workshop - www.iceaaonline.com

Outline

- Background
- Purpose
- Experimental Design
- Analysis
- Results
- Conclusions

Abbey.Turnau@ngc.com June 2007

NORTHROP GRUMMAN

Background

- **Software development effort is often estimated using commercial tools such as PRICE and SEER**
- **Obtaining accurate estimates is an important key to program planning and control**
- **Size inputs are organized into Computer Software Configuration Items (CSCIs) typically based on the architecture designed by the systems engineering team**
- **In many situations the CSCI structure is unknown to cost estimators**
  - **If the estimate is performed early on in the development lifecycle, the CSCI structure may not yet be defined**
  - **Often times when a third party is performing an independent cost estimate, access to information at the CSCI level is not provided**
    - **Without this information, the third party analysts must make assumptions about the architecture**

NORTHROP GRUMMAN

Presented at the 2007 ISPA/SCEA Joint Annual International Conference and Workshop - www.iceaaonline.com

Purpose

- **To evaluate the sensitivity of two commercial software cost estimating tools to the Computer Software Configuration Item (CSCI) structure**
  - **Is the output effort (hours or cost) sensitive to the structure of the inputs?**
    - **Fixed total amount of code**
    - **Varying number of CSCIs**
  - **How sensitive is it?**
  - **What assumptions play a critical role?**
    - **If there are multiple CSCIs, does the user (or the tool) assume they start development concurrently?**
  - **If an estimate must be done but nothing is known about the software structure, what is a good default?**
  - **IN NO WAY IS THIS STUDY INTENDED TO DETERMINE OR IMPLY THAT ONE COMMERCIAL TOOL IS BETTER THAN ANOTHER**

NORTHROP GRUMMAN

Presented at the 2007 ISPA/SCEA Joint Annual International Conference and Workshop - www.iceaaonline.com

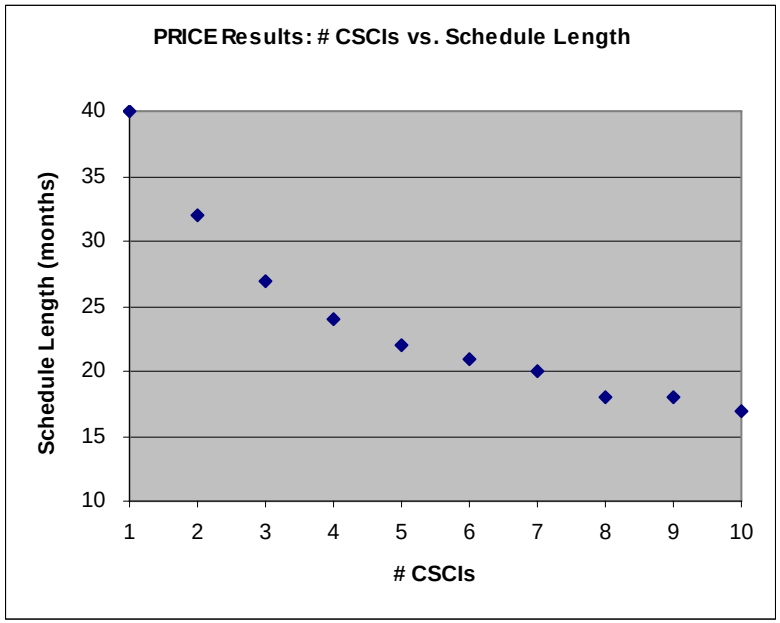
Experimental Design

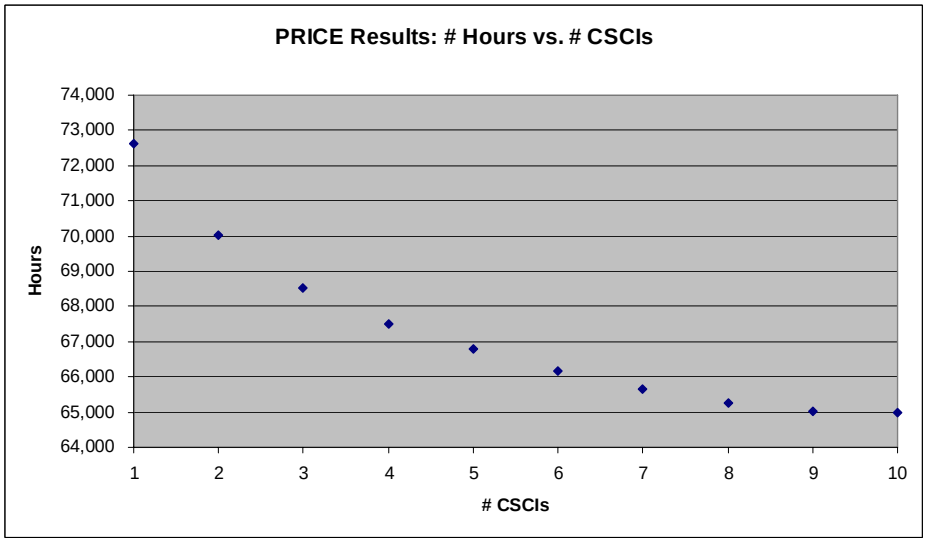
- **Commercial tools used: PRICE S and SEER SEM**
- **Constant total software size of 100,000 Equivalent Lines of Code (ELOC), spread equally among the CSCIs**
- **Number of CSCIs varying from 1 to 10**
- **Optimal schedule**
- **Spiral development**
- **All other tool inputs are set to nominal (factory defaults) and held constant**

Results

- **In both tools, it was found that a software structure with only 1 CSCI resulted in the greatest effort**
- **PRICE results**

# CSCIs	Total Hours
1	72,641
2	70,006
3	68,546
4	67,495
5	66,786
6	66,178
7	65,673
8	65,241
9	65,023
10	64,981



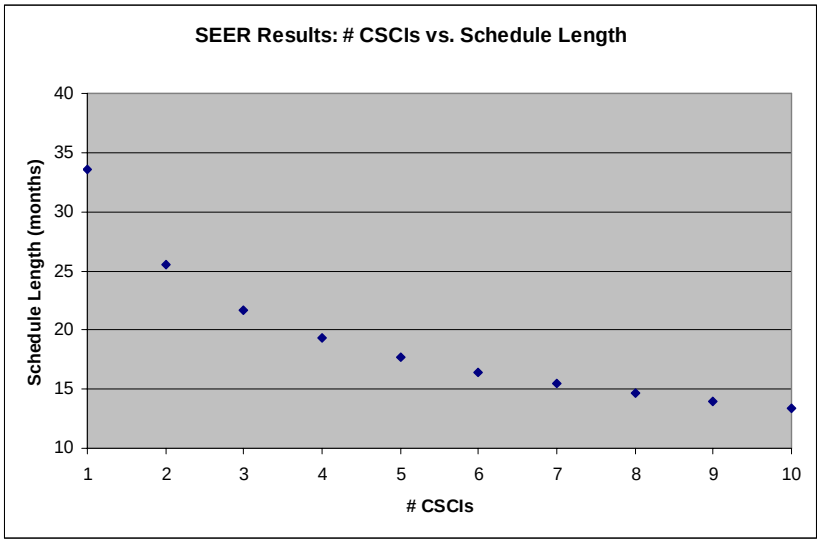


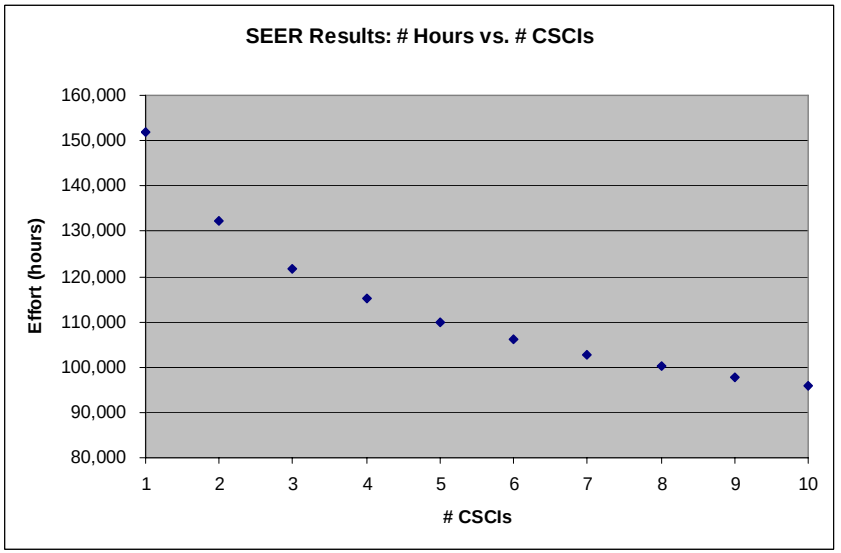
NORTHROP GRUMMAN

Results (cont'd)

- **SEER Results**

# CSCIs	Total Hours
1	151,806
2	132,155
3	121,861
4	115,048
5	110,026
6	106,086
7	102,866
8	100,155
9	97,823
10	95,783





NORTHROP GRUMMAN

Presented at the 2007 ISPA/SCEA Joint Annual International Conference and Workshop - www.iceaaonline.com

Analysis

- **Each of the resulting scatter plots appears to be a power curve**
- **To perform regressions on this data using Excel, a linear transformation must be performed**
  - **Power Curve: $y = ax^b$**
  - **Linear Equivalent: $\ln(y) = \ln(a) + b \ln(x)$**
  - **Therefore, the natural log of each x and y data point was used in the regression, as opposed to using each x and y data point if the scatter plots appeared to be linear**

Abbey.Turnau@ngc.com June 2007

Analysis (cont'd)

- **PRICE analysis**
  - **Optimal schedule**

# CSCIs	Total Hours
1	72,641
2	70,006
3	68,546
4	67,495
5	66,786
6	66,178
7	65,673
8	65,241
9	65,023
10	64,981

Take the natural log of each data point

ln(# CSCIs)	ln(Total Hours)
0.00	11.19
0.69	11.16
1.10	11.14
1.39	11.12
1.61	11.11
1.79	11.10
1.95	11.09
2.08	11.09
2.20	11.08
2.30	11.08

Perform regression using Excel

Regression Statistics	
Multiple R	1.00
R Square	1.00
Adjusted R Square	1.00
Standard Error	0.00
Observations	10

Linear Form: $\ln(y) = 11.19 - 0.05 \ln(x)$

Power Form: $y = 72{,}464x^{-0.05}$

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.01	0.01	1814.18	0.00
Residual	8	0.00	0.00		
Total	9	0.01			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	11.19	0.00	5758.64	0.00	11.19	11.20	11.19	11.20
X Variable 1	-0.05	0.00	-42.59	0.00	-0.05	-0.05	-0.05	-0.05

NORTHROP GRUMMAN

Analysis (cont'd)

- **SEER analysis**
  - **Optimal schedule**

Note: the regressions in this paper produce a near zero P value because they are not a fit of *random* data, but rather a fit of *non-random* data, in other words, OLS is being used for curve fitting, not as a regression in the usual sense.

# CSCIs	Total Hours
1	151,806
2	132,155
3	121,861
4	115,048
5	110,026
6	106,086
7	102,866
8	100,155
9	97,823
10	95,783

Take the natural log of each data point

ln(# CSCIs)	ln(Total Effort)
0.00	11.93
0.69	11.79
1.10	11.71
1.39	11.65
1.61	11.61
1.79	11.57
1.95	11.54
2.08	11.51
2.20	11.49
2.30	11.47

Perform regression using Excel

Regression Statistics	
Multiple R	1.00
R Square	1.00
Adjusted R Square	1.00
Standard Error	0.00
Observations	10

Linear Form: $\ln(y) = 11.93 - 0.20 \ln(x)$

Power Form: $y = 151{,}806x^{-0.20}$

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.19	0.19	28248366713.16	0.00
Residual	8	0.00	0.00		
Total	9	0.19			

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	11.93	0.00	6029368.64	0.00	11.93	11.93	11.93	11.93
X Variable 1	-0.20	0.00	-168072.50	0.00	-0.20	-0.20	-0.20	-0.20

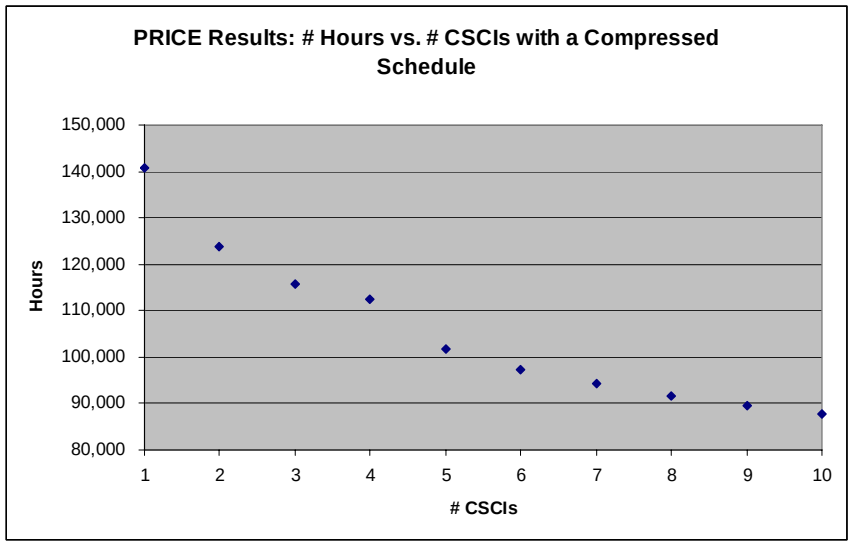
NORTHROP GRUMMAN

Presented at the 2007 ISPA/SCEA Joint Annual International Conference and Workshop - www.iceaaonline.com

Analysis (cont'd)

- **It appears that both the cost and the schedule length decrease on a power curve as the number of CSCIs increase. Is the schedule the true driver here?**
- **The same data was run through PRICE with a fixed 12 month schedule**
- **The compressed schedule results:**

# CSCIs	Total Hours
1	140,838
2	123,683
3	115,664
4	112,350
5	101,650
6	97,263
7	94,285
8	91,591
9	89,565
10	87,821



Abbey.Turnau@ngc.com June 2007

NORTHROP GRUMMAN

Analysis (cont'd)

- **Does schedule length drive the effort power curve?**
- **No. Running the data through PRICE with a fixed 12 month schedule showed that effort decreases on a power curve as the number of CSCIs increases**

# CSCIs	Total Hours
1	140,838
2	123,683
3	115,664
4	112,350
5	101,650
6	97,263
7	94,285
8	91,591
9	89,565
10	87,821

Take the natural log of each data point

ln(# CSCIs)	ln(Total Hours)
0.00	11.86
0.69	11.73
1.10	11.66
1.39	11.63
1.61	11.53
1.79	11.49
1.95	11.45
2.08	11.43
2.20	11.40
2.30	11.38

Perform regression using Excel

Regression Statistics	
Multiple R	0.99
R Square	0.98
Adjusted R Square	0.98
Standard Error	0.02
Observations	10

Linear Form: ln(y) = 11.88 – 0.21 ln(x)

Power Form: $y = 143{,}655x^{-0.21}$

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.22	0.22	524.48	0.00
Residual	8	0.00	0.00		
Total	9	0.22			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	11.88	0.02	771.11	0.00	11.84	11.91	11.84	11.91
X Variable 1	-0.21	0.01	-22.90	0.00	-0.23	-0.19	-0.23	-0.19

NORTHROP GRUMMAN

Presented at the 2007 ISPA/SCEA Joint Annual International Conference and Workshop - www.iceaaonline.com

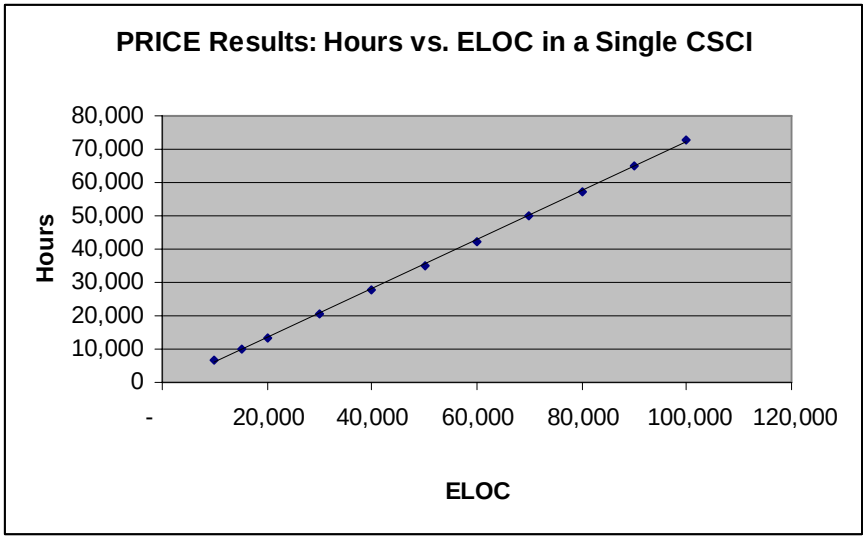
Analysis (cont'd)

- **Does increasing the number of CSCIs in a project change the cost per CSCI with constant code in a CSCI?**
  - **No. Using PRICE, ran:**
    - **Project A: 1 CSCI with 10,000 ELOC**
    - **Project B: 10 CSCIs, each with 10,000 ELOC**
    - **Compared the effort of a single CSCI from each and they were equal**

Analysis (cont'd)

- **Does increasing or decreasing the code within a single CSCI change the total effort by a power curve?**
  - **Test by running 10 data points ranging in size from 10,000 to 1,000,000 ELOC**

ELOC	Hours
10,000	6,474
20,000	13,323
30,000	20,414
40,000	27,639
50,000	34,966
60,000	42,376
70,000	49,857
80,000	57,399
90,000	64,996
100,000	72,641



  - **The scatter plot appears to be fairly linear but a closer examination of the data reveals that the hours/ELOC is rising**
    - **The rate at which the hours/ELOC rises slows as the ELOC grows**

ELOC	Hours	Hours/ELOC
10,000	6,474	0.647
20,000	13,323	0.666
30,000	20,414	0.680
40,000	27,639	0.691
50,000	34,966	0.699

ELOC	Hours	Hours/ELOC
60,000	42,376	0.706
70,000	49,857	0.712
80,000	57,399	0.717
90,000	64,996	0.722
100,000	72,641	0.726

Analysis (cont'd)

- **A regression was run on the data with the assumption that it is a power curve**
- **As shown below, as the amount of code in a single CSCI grows, the productivity decreases**
- **This is the true driver behind the sensitivity to CSCI structure**

ELOC	Hours
10,000	6,474
20,000	13,323
30,000	20,414
40,000	27,639
50,000	34,966
60,000	42,376
70,000	49,857
80,000	57,399
90,000	64,996
100,000	72,641

Take the natural log of each data point

ln(ELOC)	ln(Hours)
9.21	8.78
9.90	9.50
10.31	9.92
10.60	10.23
10.82	10.46
11.00	10.65
11.16	10.82
11.29	10.96
11.41	11.08
11.51	11.19

Regression Statistics	
Multiple R	1.00
R Square	1.00
Adjusted R Square	1.00
Standard Error	0.00
Observations	10

Linear Form: ln(y) = -0.91 + 1.05 ln(x)

Power Form: $y = 0.403x^{1.05}$

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	5.34	5.34	1150416.75	0.00
Residual	8	0.00	0.00		
Total	9	5.34			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.91	0.01	-86.28	0.00	-0.93	-0.88	-0.93	-0.88
X Variable 1	1.05	0.00	1072.57	0.00	1.05	1.05	1.05	1.05

NORTHROP GRUMMAN

Conclusions

- **Both PRICE and SEER are sensitive to the CSCI structure**
- **The sensitivity is significant and cost analysts should pay due diligence to ensure the development effort is modeled as accurately as possible**
- **Model runs for this study showed the effort required for 100,000 ELOC spread across ten CSCIs being only 60% as much effort as would be required for 100,000 ELOC all in one CSCI**
- **The sensitivity to the CSCI structure is actually driven by the relationship between productivity and CSCI size**
  - **As the amount of code in a single CSCI increases, the productivity decreases**
    - **The rate at which the productivity decreases slows as the ELOC grows**
- **The CSCI structure also drives the schedule length; this should be taken into consideration when doing schedule realism analysis**

NORTHROP GRUMMAN

Presented at the 2007 ISPA/SCEA Joint Annual International Conference and Workshop - www.iceaaonline.com

Conclusions (cont'd)

- If nothing is known about the CSCI structure and an estimate must be completed, what is a good default CSCI structure to use?
  - Let's take a look at some sample data:
    - Contractor A: 4 CSCIs with 6%, 7%, 36%, and 51% of the code, consistent for all releases
    - Contractor B: 13 CSCIs with 1%, 1%, 1%, 1%, 1%, 2%, 4%, 6%, 8%, 11%, 16%, 24%, and 24% of the code, fairly consistent for all releases
    - Contractor C: 5 CSCIs with 2%, 7%, 13%, 30%, and 48% of the code in the first release and the second and third release have only 1 CSCI each
  - Judging by the variability between contractors and programs, it seems there may be no "good default" CSCI structure, but it's probably better to err on the conservative side (less CSCIs means greater cost)
  - Ideally, the cost analyst would be able to talk to the software development team to ask about the structure
    - When gathering information about the software structure, be sure to ask if the CSCIs are developed concurrently

NORTHROP GRUMMAN