Abstract

This presentation outlines methods used to calibrate the SEER SEM software evaluation model. The presentation will review basic measures and tools, all generally applicable.

This presentation aims to achieve consistency and agreement on a basic and essential task, model calibration. An approach towards software project data collection and refinement will be outlined. Calibration measures will next be discussed, emphasizing the challenge of small samples. Efficacy measures are integral to this discussion. Throughout the talk, the need and means towards calibration credibility will be stressed.

Calibrating A Software Evaluation Model

Calibration (def: the systematic adjustment of a man-made device such that performance is improved against some external measure)

12th IFCSCM
October 9th 1997

Presented by
Lee Fischman
Galarath / SEER
www.gaseer.com
Why Calibrate?

*Model Makers Work Hard To Insure Their Models Are Delivered Pre-Calibrated...*

...But Users Often Want Enhanced Traceability

*If A Model Is Accurate...*

...Calibration May Increase Accuracy, IF DONE RIGHT!

---

Why We Should Calibrate Together

<table>
<thead>
<tr>
<th>Program Office</th>
<th>Contractor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Common Method</td>
</tr>
<tr>
<td></td>
<td>Happy Face</td>
</tr>
<tr>
<td></td>
<td>Neutral Face</td>
</tr>
<tr>
<td></td>
<td>Sad Face</td>
</tr>
</tbody>
</table>

---
Steps To Calibration

"Usation"

Calibration

Stratification

Estimation - Benchmarking

Data Correction

Data Collection

---

Data Collection

Good Actuals + Good Estimates = Good Calibration

What To Look For In Actuals
What phases are included?
Which kinds of labor are included?

What To Look For In Estimates
What kinds of people did the work?
Which tools and practices were used?
What manner of product did they produce?
What To Look For In An Actual - Phases

**Phase:** All activities may not be included.

<table>
<thead>
<tr>
<th>System Concept is excluded</th>
<th>Integration is excluded</th>
</tr>
</thead>
</table>

**Phases**
- System Concept
- System Req Analysis
- Detailed Design
- Software Test
- System Req Design
- Preliminary Design
- Coding / Unit Testing
- System Integration / OT&E

What To Look For In An Actual - Labor

**Labor:** All categories may not be included.

<table>
<thead>
<tr>
<th>Configuration Mgt. is excluded</th>
<th>Quality Assurance is excluded</th>
</tr>
</thead>
</table>

**Labor Categories**
- Management
- Design
- Data Preparation
- Configuration Mgt
- System Engineering
- Coding
- Test
- Quality Assurance
**What Else To Look For In An Actual**

**Forensics**

**Schedule.** Was it “stop and start”? Were there schedule constraints?

**Resources.** Were there hard-hitting resource constraints?

**Volatility.** Did requirements undergo extraordinary evolution?

**Manager’s Objectives.** Was it to complete the project in minimum time or at least cost?

**Effort.** Are effort figures actually derived from cost figures?

When creating the estimate --- adjustments for extraordinary conditions may be possible within the software estimating model.

---

**What To Obtain For An Estimate**

Beyond effort / cost / duration...

What information is available about each Project?

<table>
<thead>
<tr>
<th>People</th>
<th>Process</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysts' Experience</td>
<td>Development Tools</td>
<td>Size!</td>
</tr>
<tr>
<td>Programmers' Ability</td>
<td>Practices</td>
<td>Complexity</td>
</tr>
<tr>
<td></td>
<td>Requirements</td>
<td>Specification Level</td>
</tr>
<tr>
<td></td>
<td>Volatility</td>
<td></td>
</tr>
</tbody>
</table>

Use A Model’s Worksheet As A Guide To Factors Representing People, Process, Product.

Create best possible estimates; do not deviate from standard modelling practice.
A Parameter Settings Benchmark

A software model’s parameters also calibrate!

You must deliver a calibrated set of parameters together with other calibration factors.

Steps in parameter benchmarking:

For each estimate, achieve contractor/client buy-in for all parameter settings.
For each stratified category, isolate parameter settings that exhibit commonality.
For each delivered calibration factor, also deliver also a set of common parameter settings. Because of statistical concerns, you may be able to deliver more parameter “suites” than calibration factors.

Making An Estimate and Actual Comparable

What’s missing from the actual?

Models often break out activities and labor... allowing you to exclude these from the estimate.

Management is excluded from actual.
Concept is excluded from actual.
CM/QA is excluded from actual.

THE ACTUAL

You should exclude...
...management from the estimate.
...concept from the estimate.
...CM/QA from the estimate.
Breaking Out An Estimate

Pulling Out Labor & Activities Using A SEER-SEM Report...

<table>
<thead>
<tr>
<th>Activity</th>
<th>Management</th>
<th>Systems Eng</th>
<th>Design</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Concept</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>System Requirements Design</td>
<td>0.30</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Total Requirements Analysis</td>
<td>2.31</td>
<td>3.19</td>
<td>4.25</td>
<td>1.81</td>
</tr>
<tr>
<td>Preliminary Design</td>
<td>2.31</td>
<td>6.37</td>
<td>26.12</td>
<td>7.05</td>
</tr>
<tr>
<td>Detailed Design</td>
<td>11.52</td>
<td>10.75</td>
<td>44.06</td>
<td>12.39</td>
</tr>
<tr>
<td>Code &amp; CSS I Test</td>
<td>11.06</td>
<td>5.06</td>
<td>10.11</td>
<td>83.73</td>
</tr>
<tr>
<td>CSS Integration &amp; Test</td>
<td>16.23</td>
<td>4.06</td>
<td>6.11</td>
<td>78.10</td>
</tr>
<tr>
<td>CSS Test</td>
<td>1.89</td>
<td>8.47</td>
<td>9.33</td>
<td>9.13</td>
</tr>
<tr>
<td>Systems Integrative Testing</td>
<td>1.99</td>
<td>2.32</td>
<td>4.44</td>
<td>11.09</td>
</tr>
<tr>
<td>Maintenance</td>
<td>0.00</td>
<td>0.06</td>
<td>0.06</td>
<td>0.00</td>
</tr>
</tbody>
</table>

| Development Total         | 81.29      | 42.75       | 98.05  | 224.83 |
| Life Cycle Total          | 81.29      | 42.75       | 98.05  | 224.83 |

<table>
<thead>
<tr>
<th>Activity</th>
<th>Data Prep</th>
<th>Test</th>
<th>IOH</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Concept</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>System Requirements Design</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Stratifying Data

Calibration factors that are more specific can be more accurate
But patterns may not be evident in samples that are too small.

Categories Concerning
- Development Organization (team, contractor, government...)
- Application Type (command, control, signal, MIS...)
- Platform (device, ground, PC, client-server...)
- Estimating hierarchy (unit, program, project)
- Chronology (recent, ancient)

Beware: With more subcategories...
-Samples get smaller -- and statistical validity decreases.
Bridging Estimates And Actuals

Introducing the Estimate Ratio

Estimate Ratios usefully encapsulate... variation between actuals and estimates.

![Diagram showing actuals and estimates with a ratio of 1.5]

A single estimate ratio

\[
\frac{30K}{20K} = 1.5
\]

A potential calibration factor

\[
\frac{\text{sum of actuals}}{\text{sum of estimates}} = 1.5
\]

Calibration Factors

There are two classes of calibration factors:

Effort and Schedule. Linear adjustments to model outputs: these adjustments are made independently.

Adjusted Estimate = Original Estimate \times \text{Calibration Factor}

Technology and Complexity Adjustments. Linear adjustments to key (internal) model drivers: nonlinear effects on both schedule and effort estimates.

However, due to complexity and other factors, these adjustments are not linear.
### Linear Calibration Factors

Various measures are available:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Actuals</td>
<td>Performs well against a large asymmetric sample.</td>
</tr>
<tr>
<td>Sum of Estimates</td>
<td></td>
</tr>
<tr>
<td>Average (Estimate Ratios)</td>
<td>Almost as good as above, against asymmetric samples.</td>
</tr>
<tr>
<td>Median (Estimate Ratios)</td>
<td>Good against very small samples, handles outliers well.</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>Best against large symmetric samples. (Regression of actuals on estimates without coefficient.)</td>
</tr>
<tr>
<td>Least Absolute Deviation Regression</td>
<td>Results in a factor which minimizes deviation while controlling against the effect of outliers.</td>
</tr>
</tbody>
</table>

*Based on empirical analyses.*

### Least Absolute Deviation Regression

**Standard regression:**

Choose a constant factor that minimizes the sum of squared deviations between actuals and adjusted estimates.

\[
\min \sum (y_i - ax)^2
\]

**Least absolute deviation regression:**

Choose a constant factor that minimizes the sum of absolute deviations between actuals and adjusted estimates.

\[
\min \sum |y_i - ax|
\]
Nonlinear Calibration Factors

Technology and Complexity calibrations change the basis of inputs to the model.

Best tactic: Start with scatter plots; iterate through settings until R-squared (estimates vs. actuals) is optimally high.

Evaluating Efficacy

Calibrations are conducted with best possible estimates against best possible actuals. Calibration instruments make systematic adjustments to estimates; don’t further “tskew” estimates!

1. Strainy sample - too small is too bad.
2. Reserve a control group for each stratification - this may be a luxury in small samples.
3. Calibrate for each stratification - which measure is most appropriate?
4. Evaluate the efficacy of calibrations, using the control groups if possible. If samples are too small, evaluate before and after effects.

Use multiple statistical measures.
**Statistical Measures of Efficacy**

- **Mean Magnitude of Relative Error (MRE)**: Percentage variation between actual and estimated with actual as divisor. Closer to zero the better.
- **Average Estimation Ratio**: The average of (actual / estimate) ratios. Closer to one the better.
- **Median Estimation Ratio**: The median of (actual / estimate) ratios. Closer to one the better.
- **Standard Deviation**: The standard deviation of estimate vs. actual variances. Closer to zero the better.
- **Prediction (R)**: The percent of estimates falling within 95% of actual outcomes. Closer to 100%, the better.
- **R-squared**: Goodness of fit for nonlinear calibration only. (The goal is to purely fit the data and so this is a good measure.) Closer to one the better.
- **Wilcoxon Signed-Ranks Test**: A means test that performs well against small samples. Probability is calculated for smallest samples; tends towards normal distribution as sample size grows.

**Creating A Custom Knowledge Base**

- Create a new project – without any work elements.
- Calibration factors are located in the Adjustment Factors category; near the bottom of the Parameter View.
- Save as a Class knowledge base.
Creating A Custom Knowledge Base

A knowledge base can use more than calibration factors...

<table>
<thead>
<tr>
<th>Parameter View</th>
<th>CSCI</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personnel Capability &amp; Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysts Application Experience</td>
<td>Low</td>
<td>High</td>
<td>Very High</td>
<td></td>
</tr>
<tr>
<td>Programmer Capabilities</td>
<td>Low</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Programmers Language Experience</td>
<td>Low</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Test System Experience</td>
<td>Low</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Legal System Experience</td>
<td>Low</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
</tbody>
</table>

**DEVELOPMENT SUPPORT ENVIRONMENT**

<table>
<thead>
<tr>
<th>Parameter View</th>
<th>CSCI</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEVELOPMENT SUPPORT ENVIRONMENT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development Practice Use</td>
<td>Low</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Automated Task Use</td>
<td>Non</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Logon to Practice Transition</td>
<td>Non</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Time to Practice Transition</td>
<td>Non</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Planning Dedication</td>
<td>Non</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Resource and Support Location</td>
<td>Non</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Host System Viability</td>
<td>Non</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Process Viability</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>PRODUCT DEVELOPMENT REQUIREMENTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requirements Viability Change</td>
<td>Non</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
<tr>
<td>Specification Level</td>
<td>Reliable</td>
<td>Non</td>
<td>Non</td>
<td></td>
</tr>
</tbody>
</table>

Within a category, which parameters exhibit systematic tendencies?

Supporting Documentation

Calibrations help build traceability - with adequate documentation.

- Items to include in supporting documentation:
  - What calibration measures were used?
  - What was the size of the sample used for calibration?
  - How were actuals cleaned?
  - How were estimates modified to suit actuals?
  - What results (efficacy) was obtained?
  - Were outliers excluded from the sample?
  - What other parameters were modified and why?
Knowledge Bases In Use

With calibration factors PLUS parameter benchmarks...
There is enhanced

**ACCURACY**

and

**CONSISTENCY**
Future deviation from agreed parameter settings are grounds for skepticism.

Happy Calibrating!