Can You Trust Your Data?
Measurement and Analysis Infrastructure Diagnosis

SEPG 2007
David Zubrow
SEI
Disclaimer

This is a work in progress
It is evolving frequently
Therefore,
  • Slides are not as clean as I would like
  • Ideas are still being fleshed out
  • This is still a draft
But, I think you will get something out of it

Here is your chance to escape........
Outline

The Need for a Measurement and Analysis Infrastructure Diagnostic (MAID)
  • Why measure?
  • Measurement errors and their impact

The MAID Framework
  • Reference Model: CMMI and ISO 15939
  • Measure and Analysis Infrastructure Elements

MAID Methods
  • Process Diagnosis
  • Data and Information Product Quality Evaluation
  • Stakeholder Evaluation

Summary and Conclusion
Measurements Are Used for Many Purposes

Aggregate Data
• corporate/industry comparisons
• business decisions

Manage Projects
• plan
• track

Describe Products
• qualify
• classify

Improve Processes
• understand
• control

Measurement Process
• communicate clearly
• use process consistently
Measurement Purposes

Characterize (baseline performance)
Evaluate (actual with regard to plan)
Predict (estimation and prediction)
Improve (process improvement)
Why Measure?

Characterize

• to understand the current process, product, and environment
• to provide baselines for future assessments

Evaluate

• to determine status so that projects and processes can be controlled
• to assess the achievement
Why Measure?  

**Predict**

- to understand the relationships between and among processes and products
- to establish achievable goals for quality, costs, and schedules

**Improve**

- to identify root causes and opportunities for improvement
- to track performance changes and compare to baselines
- to communicate reasons for improving
Purposes of Measurement are Understood

Source: CMU/SEI-2006-TR-009
Do you trust your data

What do you trust? Why?

What don’t you trust? Why?
Where do Measurement Errors come From

Differing Operational Definitions
- Project duration, defect severity or type, LOC definition, milestone completion

Not a priority for those generating or collecting data
- Complete the effort time sheet at the end of the month
- Inaccurate measurement at the source

Double Duty
- Effort data collection is for Accounting not Project Management.
  - Overtime is not tracked.
  - Effort is tracked only to highest level of WBS.

Lack of rigor
- Guessing rather than measuring
- Measurement system skips problem areas
  - “Unhappy” customers are not surveyed
- Measuring one thing and passing it off as another
Where do Measurement Errors come From?

Dysfunctional Incentives

- Rewards for high productivity measured as LoC/Hr.
- Dilbert-esque scenarios

Failure to provide resources and training

- Assume data collectors all understand goals and purpose
- Arduous manual tasks instead of automation

Lack of priority or interest

- No visible use or consequences associated with poor data collection or measurement
- No sustained management sponsorship

Missing data is reported as “0”.

- Can’t distinguish 0 from missing when performing calculations.
What is Measurement Error?

Deviation from the “true” value

• Distance is 1 mile, but your odometer measures it as 1.1 miles
• Effort really expended on a task is 3 hours, but it is recorded as 2.5

Variation NOT associated with process performance

• Aggregate impact on variation of the errors of individual measurement
• Good analogy is signal to noise ration

Error introduced as a result of the measurement process used

• Not as defined, but as practiced
Are documented processes used?

1852 Responses

- Frequently: 876 (47.3%)
- Occasionally: 559 (30.2%)
- Rarely: 269 (14.5%)
- Never: 87 (4.7%)
- I don't know: 18 (1.0%)
- N/A: 43 (2.3%)

Source: CMU/SEI-2006-TR-009
Impacts of Poor Data Quality

Inability to manage the quality and performance of software or application development

Poor estimation

Ineffective process change instead of process improvement

Improper architecture and design decisions driving up the lifecycle cost and reducing the useful life of the product

Ineffective and inefficient testing causing issues with time to market, field quality and development costs

Products that are painful and costly to use within real-life usage profiles

Bad Information leading to Bad Decisions
Cost of Poor Data Quality to an Enterprise

TYPICAL ISSUES:
- Inaccurate data: 1–5% of data fields are errored
- Inconsistencies across databases
- Unavailable data necessary for certain operations or decisions

TYPICAL IMPACTS:
Operational Impacts:
- Lowered customer satisfaction
- Increased cost: 8–12% of revenue in the few, carefully studied cases
  For service organizations, 40–60% of expense
- Lowered employee satisfaction

Typical Impacts:
- Poorer decision making: Poorer decisions that take longer to make
- More difficult to implement data warehouses
- More difficult to reengineer
- Increased organizational mistrust

Strategic Impacts:
- More difficult to set strategy
- More difficult to execute strategy
- Contribute to issues of data ownership
- Compromise ability to align organizations
- Divert management attention

Source: Redman, 1998
What we are not addressing with MAID

Development process instability

• Separate issue
• Detection fairly robust against measurement error

Development process performance

• Poor performance not a function of measurement, but detecting it is

Deceit in reporting

• Could result in measurement error, but focus here is on infrastructure design and implementation and how to characterize measurement and analysis infrastructure quality

This is about the Measurement and Analysis Infrastructure
Why a Measurement and Analysis Infrastructure Diagnostic

Quality of data is important

- Basis for decision making and action
- Erroneous data can be dangerous or harmful
- Need to return value for expense

Cannot go back and correct data once it is collected – opportunity/information lost

Need to get the quality information to decision makers in an appropriate form at the right time

Measurement practices should be piloted and then evaluated periodically

- But what are the criteria for evaluation?
- How should the evaluation be done?
Outline

The Need for a Measurement and Analysis Infrastructure Diagnostic (MAID)
  • Why measure?
  • Measurement errors and their impact

The MAID Framework
  • Reference Model: CMMI and ISO 15939
  • Measure and Analysis Infrastructure Elements

MAID Methods
  • Process Diagnosis
  • Data and Information Product Quality Evaluation
  • Stakeholder Evaluation

Summary and Conclusion
MAID Objectives

Provide information to help improve an organization’s measurement and analysis activities.

• Are we doing the right things in terms of measurement and analysis?
• How well are we doing things?
• How good is our data?
• How good is the information we generate?
• Are we providing value to the organization and stakeholders?

Looking to the future

• Are we preparing for reaching higher maturity?
• Many mistakes made in establishing M&A at ML2 and 3 that do not create a good foundation for ML4 and 5
MAID Framework: Sources

CMMI Measurement and Analysis Process Area Goals

- Align measurement and analysis activities
  - Align objectives
  - Integrate processes and procedures
- Provide measurement results
- Institutionalize a managed process

ISO 15939 Measurement Process

- Plan the measurement process
- Perform the measurement process
- Establish and sustain measurement commitment
- Evaluate measurement
MAID Framework: Sources

Six Sigma
- Measurement system evaluation
- Practical applications of statistics

Basic Statistical Practice
- Types of measures and appropriate analytical techniques
- Modeling and hypothesis testing techniques
Basic Support Process Areas

MA = Measurement and Analysis
CM = Configuration Management
PPQA = Process and Product Quality Assurance
ISO 15939 Measurement Process

Requirements for Measurement

Technical and Management Processes

Information Needs

Information Products

Measurement User Feedback

Plan the Measurement Process (5.2)

Perform the Measurement Process (5.3)

Core Measurement Process

Plan the Measurement Process (5.2)

Perform the Measurement Process (5.3)

Evaluation Measurement (5.4)

Measurement Experience Base

Establish & Sustain Measurement Commitment (5.1)

Source: ISO/IEC 15939, 2002

Legend

Activity
Flow
Data Storage
Information Needs

ISO 15939 Information Model

Entity

Attribute

Measurement Method

Base Measure

Data collection

Source: ISO/IEC 15939, 2002
Elements of the Measurement and Analysis Infrastructure

Planning for Measurement and Analysis

- Measurement plans
- Data definitions – indicator templates, measurement constructs
- Data collection and storage procedures
- Data analysis and reporting procedures

Performing Measurement and Analysis

- Data collected – base measures
- Analyses performed – derived measures, models
- Reports produced – indicators, interpretations

Institutionalizing Measurement and Analysis

- Tools used
- Staffing
- Training
- QA activities
- Improvement activities
Criteria for Evaluation: Measurement Planning Criteria
(ISO 15939)

Measurement Objectives and Alignment

• business and project objectives
• prioritized information needs and how they link to the business, organizational, regulatory, product and/or project objectives
• necessary organizational and/or software process changes to implement the measurement plan
• criteria for the evaluation of the measurement process and quality assurance activities
• schedule and responsibilities for the implementation of measurement plan including pilots and organizational unit wide implementation
Measurement Planning Criteria\textsubscript{2} (ISO 15939)

Measurement Process

- definition of the measures and how they relate to the information needs
- responsibility for data collection and sources of data
- schedule for data collection (e.g., at the end of each inspection, monthly)
- tools and procedures for data collection
- data storage
- requirements for data verification and verification procedures
- confidentiality constraints on the data and information products, and actions/precautions necessary to ensure confidentiality
- procedures for configuration management of data, measurement experience base, and data definitions
- data analysis plan including frequency of analysis and reporting
Criteria for Evaluation: Measurement Processes and Procedures

Measurement Process Evaluation

- Availability and accessibility of the measurement process and related procedures
- Defined responsibility for performance
- Expected outputs
- Interfaces to other processes
  - Data collection may be integrated into other processes
- Are resources for implementation provided and appropriate
- Is training and help available?
- Is the plan synchronized with the project plan or other organizational plans?
Criteria for Evaluation: Data Definitions

Data Definitions (meta data)

• Completeness of definitions
  — Lack of ambiguity
  — Clear definition of the entity and attribute to be measures
  — Definition of the context under which the data are to be collected
• Understanding of definitions among practitioners and managers
• Validity of operationalized measures as compared to conceptualized measure (e.g., size as SLOC vs FP)
Validity

Definition: Extent to which measurements reflect the “true” value

Observed Value = True Value + error

Compliment to Measurement Reliability – another characterization of measurement error

Various strengths of validity based on evidence and demonstration

Practical perspective – How well does our approach to measuring really match our measurement objective?

• Does number of lines of code really reflect software size? How about the amount of effort?
• Does the number of paths through the code really reflect complexity? Size of vocabulary and length (Halstead)? Depth of inheritance?
• Does the number of defects really reflect quality?

Often becomes an exercise in logic (which is ok)
Criteria for Evaluation: Data Collection

Data collection

- Is implementation of data collection consistent with definitions?
- Reliability of data collection (actual behavior of collectors)
- Reliability of instrumentation (manual/automated)
- Training in data collection methods
- Ease/cost of collecting data
- Storage
  - Raw or summarized
  - Period of retention
  - Ease of retrieval
Criteria for Evaluation: Data

Quality

- Data integrity and consistency
- Amount of missing data
  - Performance variables
  - Contextual variables
- Accuracy and validity of collected data
- Timeliness of collected data
- Precision and reliability (repeatability and reproducibility) of collected data
- Are values traceable to their source (meta data collected)

Audits of Collected Data
Criteria for Evaluation: Data Analysis

Data analysis

- Data used for analysis vs. data collected but not used
- Appropriateness of analytical techniques used
  - For data type
  - For hypothesis or model
- Analyses performed vs reporting requirements
- Data checks performed
- Assumptions made explicit
Criteria for Evaluation: Reporting

Reporting

- Evidence of use of the information
- Timing of reports produced
- Validity of measures and indicators used
- Coverage of information needs
  - Per CMMI
  - Per Stakeholders
- Inclusion of definitions, contextual information, assumptions and interpretation guidance
Criteria for Evaluation: Stakeholder Satisfaction

Stakeholder Satisfaction

- Survey of stakeholders regarding the costs and benefits realized in relation to the measurement system
- What could be approved
  - Timeliness
  - Efficiency
  - Defect containment
  - Customer satisfaction
  - Process compliance

Adapted from ISO 15939.
Outline

The Need for a Measurement and Analysis Infrastructure Diagnostic (MAID)
  • Why measure?
  • Measurement errors and their impact

The MAID Framework
  • Reference Model: CMMI and ISO 15939
  • Measure and Analysis Infrastructure Elements

MAID Methods
  • Process Diagnosis
  • Data and Information Product Quality Evaluation
  • Stakeholder Evaluation

Summary and Conclusion
Methods Overview

SCAMPI C Artifact Review – Are we doing the right things?

Measure System Evaluation – Are we do things right?

Interviews, Focus Groups – How do stakeholders perceive and experience the measurement system?
## Measurement and Analysis Infrastructure Diagnostic Elements and Evaluation Methods

<table>
<thead>
<tr>
<th>Elements</th>
<th>Method</th>
<th>Process Assessment</th>
<th>Measurement System Evaluation</th>
<th>Survey, Interview, Focus Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Plans, Data and Process Definitions</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Data Collection</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Analyses, Reports</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Stakeholder Ratings</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Measurement and Analysis Process Diagnosis: Are we doing the right things?

Use a SCAMPI C approach to look at planning and guidance documents as well as elements of institutionalization.

Elements to Address

- Plans, Process Definitions, Data definitions
- Data Collection Processes
- Data Analysis and Reporting Process
- Stakeholder Evaluation

Infrastructure for measurement support

- People and skills for development of measures
- Data repositories
- Time for data generation and collection
- Processes for timely reporting
Establishing Measurement Objectives: Basic Project Management Process Areas

PMC = Project Monitoring and Control
PP = Project Planning
SAM = Supplier Agreement Management
Establishing Measurement Objectives: Advanced Project Management Process Areas
Establishing Measurement Objectives: Basic Process Management Process Areas

Senior management

Organization's business objectives

OT

Training for projects and support groups in standard process and assets

Training needs

Standard process and other assets

OPF

OPD+IPPD

Resources and coordination

Improvement information (e.g., lessons learned, data, and artifacts)

Process-improvement proposals; participation in defining, assessing, and deploying processes

Project Management, Support, and Engineering process areas

OFF = Organizational Process Focus
OT = Organizational Training
OPD+IPPD = Organizational Process Definition (with the IPPD addition)
Establishing Measurement Objectives: Engineering Process Areas

Diagram:

- REQM: Requirements
- RD: Requirements Development
- TS: Technical Solution
- PI: Product Integration
- VER: Verification
- VAL: Validation
- Customer

Arrows indicate flow:
- Requirements to Product and product component requirements
- Alternative solutions to Product components
- Product components to Product
- Product to Customer

Customer needs:
- Product components, work products, verification and validation reports
Establishing Measurement Objectives: Basic Support Process Areas

MA - Measurement and Analysis
CM - Configuration Management
PPQA - Process and Product Quality Assurance
Establishing Information Needs: Advanced Support Process Areas

Review a sample of analyses associated with PIPs and formal evaluations

CAR = Causal Analysis and Resolution
DAR = Decision Analysis and Resolution
Documenting Measurement Objectives, Indicators, and Measures

- Establish Measurement Objectives
- Specify Data Collection Procedures
- Collect Data
- Communicate Results
- Specify Measures
- Data Storage
  - Where
  - How
  - Security
- Algorithm
- Assumptions
- Interpretation
- Probing Questions
- Analysis
- Evolution
- Feedback Guidelines
- X-reference
- Store Data & Results
- Specify Analysis Procedures
- Analyze Data

- Data Elements
  - Definitions

- Input(s)
- Objective Questions
- Visual Display
- Perspective
- Communicate Results
- Indicators Name/Title
- Data Reporting
- How
- By Whom
- Form(s)
- Responsibility for Reporting
- By/To Whom
- How Often
Schedule Predictability—1

**Indicator Name:** Schedule Predictability

**Objective:** To monitor trends in the predictability of meeting schedules as input toward improvements at the technical unit level and across the enterprise.

**Questions:**
- Are we improving our schedule estimates in small, medium, and large projects?
- How far are our schedule plans from actual effort, cost, & dates?

**Visual Display:**

![Graph showing percent deviation over time for different project effort categories (Small, Medium, Large)](chart.png)
Schedule Predictability—2

**Input:** Data is to be segregated into three project effort categories (small, medium, and large) and only submitted for projects completed during the quarter.

**Data Elements:**
There are two types of input data:

1. **Organizational reference information,** which includes
   - name of organization
   - reporting period
   - contact person
   - contact phone number

2. **Schedule predictability metric data** for each project completed during the period, which includes
   - actual date of the end of the design phase
   - planned ship date
   - project end date
   - effort category (small, medium, or large)
Schedule Predictability—3

Project Phases

<table>
<thead>
<tr>
<th>Feasible Study</th>
<th>Alternative Analysis</th>
<th>Functional Specification</th>
<th>Design</th>
<th>Code &amp; Unit Test</th>
<th>Integration Test</th>
<th>UAT</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initiation</td>
<td>Definition</td>
<td>Design</td>
<td>Build</td>
<td>Verification</td>
<td></td>
<td></td>
<td>Implementation</td>
</tr>
</tbody>
</table>

Start date  
End of design (Start of construction)  
End date (Ship date)  
Planned  
Actual

Project End Date: Actual calendar date the project ends; when the user formally signs off the UAT.

Graphic included to ensure no misunderstanding.
Schedule Predictability—4

Responsibility for Reporting:
The project manager is responsible for collecting and submitting the input.

Forms:
Forms to record the required data can be designed and maintained at the organization level.

Algorithm:
The deviation from the planned schedule is calculated based on the number of calendar days the project end date deviates from the planned ship date, expressed as a percentage of the planned duration.

The percent deviation is calculated for each effort category according to the following formula:

\[
\text{Percent Deviation} = \frac{\text{absolute value (project end date - planned end date)}}{\text{(Planned end date - start date)}} \times 100
\]
Algorithm: (continued)

The average percent deviation for each effort grouping is plotted for each quarter.

Assumptions: Schedule deviation is undesirable regardless of whether it is a slip in delivery date or a shipment earlier than planned. The goal of project schedule estimations is accuracy so that others may plan their associated tasks with a high degree of confidence. (A shipment of software a month early may just sit for a month until UAT personnel are free to begin testing.)

- Measurements are based on elapsed calendar days without adjustment for weekends or holidays.
- The value reported for planned ship date is the estimate of planned ship date made at the end of the design phase (start of construction).
Schedule Predictability—6

**Probing Questions:**
- Is there a documented process that specifies how to calculate the planned ship date?
- Does the planning process take into account historic data on similar projects?
- Has the customer successfully exerted pressure to generate an unrealistic plan?
- How stable have the requirements been on projects that have large deviation?
- Do delivered projects have the full functionality anticipated or has functionality been reduced to stay within budget?
**Evolution:** The breakdown based on project effort (small, medium, or large) can be modified to look at projects based on planned duration (e.g., all projects whose planned duration lies within a specified range). This may lead to optimization of project parameters based on scheduling rules.

**Historical data** can be used in the future to identify local cost drivers and to fine tune estimation models in order to improve accuracy. **Confidence limits** can be placed around estimates, and root cause analysis can be performed on estimates falling outside these limits in order to remove defects from the estimation process.
## Schedule Predictability—8

**Definitions:**  

**Project Effort Categorization:** The completed projects are grouped into the three effort categories (small, medium, large) according to the criteria described in the table below.

<table>
<thead>
<tr>
<th>Categories</th>
<th>SMALL</th>
<th>MEDIUM</th>
<th>LARGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development Effort (hours)</td>
<td>&lt; 200 hrs</td>
<td>200 – 1800 hrs</td>
<td>&gt; 1800 hrs</td>
</tr>
</tbody>
</table>
# Milestone Definition Checklist

<table>
<thead>
<tr>
<th>Start &amp; End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Start Date</strong></td>
</tr>
<tr>
<td>✓ Sign-off of user requirements that are detailed enough to start functional specification</td>
</tr>
<tr>
<td>✓ Kick-off meeting</td>
</tr>
<tr>
<td><strong>Project End Date</strong></td>
</tr>
<tr>
<td>✓ Actual UAT sign-off by customer</td>
</tr>
<tr>
<td><strong>Estimation Start Date</strong></td>
</tr>
<tr>
<td>✓ Start of code construction</td>
</tr>
</tbody>
</table>
Are we doing things right? Quality Assessment

Use Six Sigma Measurement System Evaluation and Statistical Methods Review

Focus on Artifacts of the Measurement and Analysis Infrastructure

- Data
- Analyses
- Reports

Assess for quality
Measurement System Evaluation

Data Evaluation: Basic Data Integrity Analysis

- Single variable
- Multiple variables

Data and Data Collection Evaluation: Measurement Validity and Reliability Analysis

- Accuracy and Validity
- Precision and Reliability

Data Definitions

- Fidelity between operational definitions and data collection

Data Analysis and Reporting Evaluation

- Appropriate Use of Analytical Techniques
- Usability of reports
Basic Data Integrity: Tools and Methods

Single Variable

1. Inspect univariate descriptive statistics for accuracy of input
   - Out of range values
   - Plausible central tendency and dispersions
   - Coefficient of variation

2. Evaluate number and distribution of missing data

3. Identify and address outliers
   - Univariate
   - Multivariate

4. Identify and address skewness in distributions
   - Locate skewed variables
   - Transform them
   - Check results of transformation

5. Identify and deal with nonlinearity and heteroscedasticity

6. Evaluate variable for multicollinearity and singularity

Tabachnick and Fidel, 1983
Data Integrity: Tools and Methods

Histograms or frequency tables
- Identify valid and invalid values
- Identify proportion of missing data
- Nonnormal distributions

Run charts
- Identify time oriented patterns

Multiple Variables
Checking sums
Crosstabulations and Scatterplots
- Unusual/unexpected relationships between two variables

Apply the above to particular segments (e.g., projects, products, business units, time periods, etc…)
Example: Histogram and Descriptive Stats

Summary for Mtg_Time

- Non-normal distribution
- Outliers
- 95% Confidence Intervals
  - Mean: 0.62030 ± 0.79280
  - Median: 0.50000 ± 0.54057
  - StDev: 0.60675 ± 0.72930

Anderson-Darling Normality Test
- A-Squared: 12.62
- P-Value < 0.005

Non-normal distribution

- Mean: 0.70655
- StDev: 0.66237
- Variance: 0.43873
- Skewness: 2.9288
- Kurtosis: 13.9671
- N: 229

Minimum: 0.00000
1st Quartile: 0.30000
Median: 0.50000
3rd Quartile: 1.00000
Maximum: 5.50000

95% Confidence Interval for Mean: 0.62030 ± 0.79280
95% Confidence Interval for Median: 0.50000 ± 0.54057
95% Confidence Interval for StDev: 0.60675 ± 0.72930
Example: Boxplot

Boxplot of Mtg_Time

Outliers
Example: Frequency Table

<table>
<thead>
<tr>
<th>Mtg_Time</th>
<th>Count</th>
<th>Mtg_Time</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>10</td>
<td>1.00</td>
<td>28</td>
</tr>
<tr>
<td>0.05</td>
<td>1</td>
<td>1.20</td>
<td>4</td>
</tr>
<tr>
<td>0.10</td>
<td>5</td>
<td>1.25</td>
<td>2</td>
</tr>
<tr>
<td>0.15</td>
<td>3</td>
<td>1.40</td>
<td>2</td>
</tr>
<tr>
<td>0.20</td>
<td>17</td>
<td>1.50</td>
<td>8</td>
</tr>
<tr>
<td>0.25</td>
<td>16</td>
<td>1.70</td>
<td>2</td>
</tr>
<tr>
<td>0.30</td>
<td>22</td>
<td>1.75</td>
<td>1</td>
</tr>
<tr>
<td>0.40</td>
<td>15</td>
<td>2.00</td>
<td>2</td>
</tr>
<tr>
<td>0.45</td>
<td>3</td>
<td>2.10</td>
<td>1</td>
</tr>
<tr>
<td>0.50</td>
<td>37</td>
<td>2.50</td>
<td>1</td>
</tr>
<tr>
<td>0.55</td>
<td>2</td>
<td>2.60</td>
<td>1</td>
</tr>
<tr>
<td>0.60</td>
<td>6</td>
<td>2.75</td>
<td>2</td>
</tr>
<tr>
<td>0.70</td>
<td>5</td>
<td>3.00</td>
<td>2</td>
</tr>
<tr>
<td>0.75</td>
<td>9</td>
<td>3.50</td>
<td>1</td>
</tr>
<tr>
<td>0.80</td>
<td>8</td>
<td>5.50</td>
<td>1</td>
</tr>
<tr>
<td>0.85</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

60 min

15 – 20 min

<table>
<thead>
<tr>
<th>Mtg_Time</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>10</td>
</tr>
<tr>
<td>0.05</td>
<td>1</td>
</tr>
<tr>
<td>0.10</td>
<td>5</td>
</tr>
<tr>
<td>0.15</td>
<td>3</td>
</tr>
<tr>
<td>0.20</td>
<td>17</td>
</tr>
<tr>
<td>0.25</td>
<td>16</td>
</tr>
<tr>
<td>0.30</td>
<td>22</td>
</tr>
<tr>
<td>0.40</td>
<td>15</td>
</tr>
<tr>
<td>0.45</td>
<td>3</td>
</tr>
</tbody>
</table>

30 min

<table>
<thead>
<tr>
<th>Mtg_Time</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>37</td>
</tr>
<tr>
<td>0.55</td>
<td>2</td>
</tr>
<tr>
<td>0.60</td>
<td>6</td>
</tr>
<tr>
<td>0.70</td>
<td>5</td>
</tr>
</tbody>
</table>

45 min

<table>
<thead>
<tr>
<th>Mtg_Time</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>9</td>
</tr>
<tr>
<td>0.80</td>
<td>8</td>
</tr>
<tr>
<td>0.85</td>
<td>1</td>
</tr>
<tr>
<td>0.90</td>
<td>7</td>
</tr>
</tbody>
</table>
How would you get a sense of the measurement error associated with time spent in an inspection meeting?
# Missing Data: Analysis of Missing Build Indicator

<table>
<thead>
<tr>
<th>Build</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>N=</td>
<td>146</td>
</tr>
<tr>
<td>*=</td>
<td>83</td>
</tr>
</tbody>
</table>

36% missing

## Two-sample T for Mtg.Time

<table>
<thead>
<tr>
<th>Build</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing</td>
<td>83</td>
<td>0.90</td>
<td>0.837</td>
<td>0.092</td>
</tr>
<tr>
<td>Present</td>
<td>146</td>
<td>0.60</td>
<td>0.510</td>
<td>0.042</td>
</tr>
</tbody>
</table>

Difference = mu (0) - mu (1)

Estimate for difference: 0.306

95% CI for difference: (0.106, 0.506)

T-Test of difference = 0 (vs not =): T-Value = 3.03  P-Value = 0.003  DF = 117
Measurement System Evaluation: Magnitude of Measurement Error

What is Measurement System Evaluation (MSE)?

• A formal statistical approach to characterizing the accuracy and precision of the measurement system

What can MSE tell you?

• The accuracy of the measures
• The magnitude of variation in the process due to the measurement system vs true process variation
Accuracy (Bias)

Accuracy: The closeness of (average) reading to the correct value or accepted reference standard.

Compare the average of repeated measurements to a known reference standard (may use fault seeding for inspections and test processes).

Statistical tool: one-to-standard

\[
\begin{align*}
\text{Ho: } & \mu = \text{known value} \\
\text{Ha: } & \mu \neq \text{known value}
\end{align*}
\]

Accurate

Not accurate
Sources of Variation

\[
\sigma^2_{\text{Total}} = \sigma^2_{\text{Process}} + \sigma^2_{\text{MS}}
\]

How much variation can be attributed to the measurement system?

Measurement error = \[\frac{\sigma^2_{\text{MS}}}{\sigma^2_{\text{Total}}} : \]

Measurement error < 10% is acceptable

10% < Measurement error < 30% questionable

Measurement error > 30% unacceptable
Test of Meeting Time with Random Error Added

Paired T for Mtg_Time - newmtg2 (Random Error Added)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mtg_Time</td>
<td>229</td>
<td>0.7066</td>
<td>0.6624</td>
<td>0.0438</td>
</tr>
<tr>
<td>newmtg2</td>
<td>229</td>
<td>0.6777</td>
<td>1.1073</td>
<td>0.0732</td>
</tr>
<tr>
<td>Difference</td>
<td>229</td>
<td>0.0289</td>
<td>0.9052</td>
<td>0.0598</td>
</tr>
</tbody>
</table>

95% CI for mean difference: (-0.0890, 0.1467)

T-Test of mean difference = 0 (vs not = 0): T-Value = 0.48
P-Value = 0.630

Central tendency not affected, but variance is
Test of Variances: Meeting Time vs Meeting Time with Additional Random Error
Precision

Spread refers to the standard deviation of a distribution.

The standard deviation of the measurement system distribution is called the precision, \( \sigma_{MS} \). GRR is Gage Repeatability and Reproducibility.

\[
GRR = \frac{\sigma_{MS}}{\sigma_{Total}} \times 100 \% 
\]

Precision is made up of two sources of variation or components: repeatability and reproducibility.

\[
\sigma^{2}_{Measurement \ System} = \sigma^{2}_{RPD} + \sigma^{2}_{RPT}
\]
Repeatability

Repeatability is the inherent variability of the measurement system.

Measured by $\sigma_{\text{RPT}}$, the standard deviation of the distribution of repeated measurements.

The variation that results when repeated measurements are made under identical conditions:

- same inspector, analyst
- same set up and measurement procedure
- same software or document or dataset
- same environmental conditions
- during a short interval of time
Reproducibility

Reproducibility is the variation that results when different conditions are used to make the measurement:

- different software inspectors or analysts
- different set up procedures, checklists at different sites
- different software modules or documents
- different environmental conditions;

Measured during a longer period of time.

Measured by \( \sigma_{\text{RPD}} \).
Types of Data—1

Nominal

Data set / observations placed into categories; may have unequal intervals.

Examples
- Defect type
- Job titles

Discrete
(aka, categorized, attribute)

Increasing information content

Continuous
(aka, variable)

What are some examples in your domain?
Types of Data—2

**Nominal**
- *Data set / observations placed into categories; may have unequal intervals.*
- Examples:
  - Defect type
  - Job titles

**Ordinal**
- *Data set with a > or < relationships among the categories; may have unequal intervals; integer values commonly used.*
- Examples:
  - Satisfaction ratings: unsatisfied, neutral, delighted
  - Risk estimates: low, med, high
  - CMMI maturity levels

**Continuous**
- (aka, variable)

What are some examples in your domain?
Types of Data—3

**Nominal**  
Data set / observations placed into categories; may have unequal intervals.  
Examples  
- Defect type  
- Job titles

**Ordinal**  
Data set with a > or < relationships among the categories; may have unequal intervals; integer values commonly used  
Examples  
- Satisfaction ratings: unsatisfied, neutral, delighted  
- Risk estimates: low, med, high  
- CMMI maturity levels

**Interval**  
Data set assigned to points on a scale in which the units are the same size; decimal values possible  
Examples  
- Degree F, C

What are some examples in your domain?
### Types of Data—4

#### Nominal
- **Data set / observations placed into categories; may have unequal intervals.**
- **Examples**
  - Defect counts by type
  - Job titles

#### Ordinal
- **Data set with a > or < relationships among the categories; may have unequal intervals; integer values commonly used**
- **Examples**
  - Satisfaction ratings: unsatisfied, neutral, delighted
  - Risk estimates: low, med, high
  - CMMI maturity levels

#### Ratio
- **Interval data set which also has a true zero point; decimal values possible**
- **Examples**
  - Time
  - Cost
  - Code size
  - Counts

---

**What are some examples in your domain?**
Assessment of Reliability for Continuous Data—1

- Have **10 objects** to measure (projects to forecast, modules of code to inspect, tests to run, etc…; variables data involved!).

- Have **3 appraisers** (different forecasters, inspectors, testers, etc…).

- Have each person repeat the measurement at least 2 times for each object.

- Measurements should be made independently and in random order.

- **Calculate the %GRR metric** to determine acceptability of the measurement system (see output next page).
Assessing Reliability for Continuous Data—2

Gage R&R

<table>
<thead>
<tr>
<th>Source</th>
<th>%Contribution</th>
<th>VarComp (of VarComp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Gage R&amp;R</td>
<td>0.09143</td>
<td>7.76</td>
</tr>
<tr>
<td>Repeatability</td>
<td>0.03997</td>
<td>3.39</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>0.05146</td>
<td>4.37</td>
</tr>
<tr>
<td>Operator</td>
<td>0.05146</td>
<td>4.37</td>
</tr>
<tr>
<td>Part-To-Part</td>
<td>1.08645</td>
<td>92.24</td>
</tr>
<tr>
<td>Total Variation</td>
<td>1.17788</td>
<td>100.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Study Var (SD)</th>
<th>%Study Var (%SV)</th>
<th>%Tolerance (SV/Toler)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Gage R&amp;R</td>
<td>0.30237</td>
<td>1.81423</td>
<td>27.86</td>
</tr>
<tr>
<td>Repeatability</td>
<td>0.19993</td>
<td>1.19960</td>
<td>18.42</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>0.22684</td>
<td>1.36103</td>
<td>20.90</td>
</tr>
<tr>
<td>Operator</td>
<td>0.22684</td>
<td>1.36103</td>
<td>20.90</td>
</tr>
<tr>
<td>Part-To-Part</td>
<td>1.04233</td>
<td>6.25396</td>
<td>96.04</td>
</tr>
<tr>
<td>Total Variation</td>
<td>1.08530</td>
<td>6.51180</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Reliability Calculations for Attribute Data—1

Conducting measurement system evaluation on attribute data is slightly different from the continuous data.

Two approaches for Attribute Data will be discussed:

- Quick rule of thumb approach
- Formal statistical approach, using Minitab
MSE Calculations for Attribute Data—2

Quick Rule of Thumb Approach for Pass/Fail Data

1. Randomly select 20 items to measure
   - Ensure at least 5-6 items barely meet the criteria for a “pass” rating.
   - Ensure at least 5-6 items just miss the criteria for a “pass” rating.

2. Select two appraisers to rate each item twice.
   - Avoid one appraiser biasing the other.

3. If all ratings agree (four per item), then the measurement error is acceptable, otherwise the measurement error is unacceptable.
Formal Statistical Approach

1. Use Minitab Attribute Agreement Analysis to measure error:
   - within appraisers
   - between appraisers
   - against a known rating standard

2. Select at least 20 items to measure.

3. Identify at least 2 appraisers who will measure each item at least twice.

4. View 95% Confidence Intervals on % accurate ratings (want to see 90% accuracy).

5. Use Fleiss’ Kappa statistic or Kendall’s coefficients to conduct hypothesis tests for agreement.
When should each formal statistical approach be used?

**Attribute data is on Nominal scale**
- Fleiss’ Kappa statistic
- e.g. Types of Inspection Defects,
  - Types of Test Defects, ODC Types,
  - Priorities assigned to defects,
  - Most categorical inputs to project forecasting tools,
  - Most human decisions among alternatives

**Attribute data is on Ordinal scale**
- Kendall’s coefficients
- (each item has at least 3 levels)
- e.g. Number of major inspection defects found,
  - Number of test defects found,
  - Estimated size of code to nearest 10 KSLOC,
  - Estimated size of needed staff,
  - Complexity and other measures used to evaluate architecture, design & code
MSE Calculations for Attribute Data—5

Interpreting results of Kappa’s or Kendall’s coefficients

<table>
<thead>
<tr>
<th>Result Range</th>
<th>Agreement Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result = 1.0</td>
<td>perfect agreement</td>
</tr>
<tr>
<td>Result &gt; 0.9</td>
<td>very low measurement error</td>
</tr>
<tr>
<td>0.70 &lt; Result &lt; 0.9</td>
<td>marginal measurement error</td>
</tr>
<tr>
<td>Result &lt; 0.7</td>
<td>too much measurement error</td>
</tr>
<tr>
<td>Result = 0</td>
<td>agreement only by chance</td>
</tr>
</tbody>
</table>

Interpreting the accompanying p value

**Null Hypothesis:** Consistency by chance; no association

**Alternative Hypothesis:** Significant consistency & association

*Thus, a p value < 0.05 indicates significant and believable consistency or association.*
Reliability Calculations for Attribute Data—6

Fleiss' Kappa Statistics

<table>
<thead>
<tr>
<th>Appraiser</th>
<th>Response</th>
<th>Kappa</th>
<th>SE Kappa</th>
<th>Z</th>
<th>P(vs &gt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Architecture</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Code</td>
<td>0.780220</td>
<td>0.316228</td>
<td>2.46727</td>
<td>0.0068</td>
<td></td>
</tr>
<tr>
<td>Design</td>
<td>0.523810</td>
<td>0.316228</td>
<td>1.65643</td>
<td>0.0488</td>
<td></td>
</tr>
<tr>
<td>Reqt</td>
<td>0.780220</td>
<td>0.316228</td>
<td>2.46727</td>
<td>0.0068</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.699248</td>
<td>0.223916</td>
<td>3.12281</td>
<td>0.0009</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Architecture</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Code</td>
<td>0.780220</td>
<td>0.316228</td>
<td>2.46727</td>
<td>0.0068</td>
<td></td>
</tr>
<tr>
<td>Design</td>
<td>0.393939</td>
<td>0.316228</td>
<td>1.24575</td>
<td>0.1064</td>
<td></td>
</tr>
<tr>
<td>Reqt</td>
<td>0.375000</td>
<td>0.316228</td>
<td>1.18585</td>
<td>0.1178</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.527559</td>
<td>0.230495</td>
<td>2.28881</td>
<td>0.0110</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Architecture</td>
<td>-0.052632</td>
<td>0.316228</td>
<td>-0.16644</td>
<td>0.5661</td>
</tr>
<tr>
<td>Code</td>
<td>0.797980</td>
<td>0.316228</td>
<td>2.52343</td>
<td>0.0058</td>
<td></td>
</tr>
<tr>
<td>Design</td>
<td>0.583333</td>
<td>0.316228</td>
<td>1.84466</td>
<td>0.0325</td>
<td></td>
</tr>
<tr>
<td>Reqt</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.626168</td>
<td>0.277383</td>
<td>2.25742</td>
<td>0.0120</td>
<td></td>
</tr>
</tbody>
</table>
Response is an ordinal rating. Thus, appraisers get credit for coming close to the correct answer!

How do you interpret these Kendall coefficients and p values?

<table>
<thead>
<tr>
<th>Appraiser</th>
<th>Coef</th>
<th>SE Coef</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duncan</td>
<td>0.89779</td>
<td>0.192450</td>
<td>4.61554</td>
<td>0.0000</td>
</tr>
<tr>
<td>Hayes</td>
<td>0.96014</td>
<td>0.192450</td>
<td>4.93955</td>
<td>0.0000</td>
</tr>
<tr>
<td>Holmes</td>
<td>1.00000</td>
<td>0.192450</td>
<td>5.14667</td>
<td>0.0000</td>
</tr>
<tr>
<td>Montgomery</td>
<td>1.00000</td>
<td>0.192450</td>
<td>5.14667</td>
<td>0.0000</td>
</tr>
<tr>
<td>Simpson</td>
<td>0.93258</td>
<td>0.192450</td>
<td>4.79636</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Gold Standard: Accuracy and Precision

(σ)

Accurate but not precise

(μ)

Precise but not accurate

Both accurate and precise
## Analysis Evaluation: Appropriate Modeling

### Quantifying Relationships of X Factors with Y Outcomes

#### ANOVA & MANOVA

<table>
<thead>
<tr>
<th>X (Discrete)</th>
<th>Y (Continuous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Variable, 2 levels</td>
<td>1 Variable, 2 levels</td>
</tr>
<tr>
<td>1 Variable, 2 levels</td>
<td>1 Variable, 2 levels</td>
</tr>
<tr>
<td>Mixture of discrete &amp; continuous</td>
<td>Mixture of discrete &amp; continuous</td>
</tr>
</tbody>
</table>

- ANOVA
- 1 Way MANOVA
- 2 Way MANOVA
- 3 Way MANOVA

#### ANOVA & MANOVA in Minitab

- One-way ANOVA
- Two-way ANOVA
- MANOVA

<table>
<thead>
<tr>
<th>Y (Continuous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Variable</td>
</tr>
</tbody>
</table>

- Discriminant Analysis
- Canonical Analysis
- MANOVA

#### Y (Continuous)

<table>
<thead>
<tr>
<th>X (Discrete)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Variable</td>
</tr>
</tbody>
</table>

- Correlation & Regression
- Logistic Regression

#### Y (Continuous) in Minitab

- Stepwise Regression
- Regression Analysis

---

**You Begin Here**

© 2007 Carnegie Mellon University
Modeling Errors: Some Look Fors

Ordinal variables treated as continuous

- Regression model predicting effort deviation based on maturity level
- Regression model predicting repair effort based on defect severity

Use of correlated independent variables in a regression model
### Appropriate Analysis: Types of Hypothesis Tests

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Interval or Ratio (Parametric Tests)</th>
<th>Ordinal (Non-Parametric Tests)</th>
<th>Nominal Similarity</th>
<th>Proportion Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td># Samples (Data groups)</td>
<td>Mean</td>
<td>Variance</td>
<td>Median</td>
<td>Variance / Fit</td>
</tr>
<tr>
<td>1 Sample</td>
<td>1-sample t test</td>
<td>Chi-Square test</td>
<td>1 sample Wilcoxon Signed Ranks test</td>
<td>Kolmogorov-Smirnov Goodness of Fit test</td>
</tr>
<tr>
<td>2 Samples</td>
<td>Independent 2-sample t test</td>
<td>Normal F test</td>
<td>Independent Mann Whitney U test</td>
<td>= Medians</td>
</tr>
<tr>
<td></td>
<td>Paired t test</td>
<td>Levene test</td>
<td>Wilcoxon matched Pairs test</td>
<td>Moses test ≠ Medians</td>
</tr>
<tr>
<td>3+ Samples</td>
<td>Normal ANOVA (1 &amp; 2 way ANOVA; Balanced ANOVA; GLM) MANOVA (General &amp; Balanced)</td>
<td>Bartlett test</td>
<td>Independent Kruskal-Wallis 1-way ANOVA</td>
<td>Van der Waerden Normal scores test</td>
</tr>
<tr>
<td></td>
<td>Not Normal</td>
<td>Levene test</td>
<td>Friedman 2-way ANOVA</td>
<td>ANOM (Analysis of Means)</td>
</tr>
</tbody>
</table>
Hypothesis Test Errors: Some Look Fors

No formal statement of a hypothesis
- No specification of null and alternative (e.g., 1 or 2 sided test)
- Failure to specify rejection level of null

Confusing failure to reject the null as proof that means are equal
- Improved maturity reduces fielded defects
  - Null: Fielded defects in products from low maturity organizations are equal to those in products from high maturity organizations
  - Alternative: They are not equal
- Improved maturity does not increase development time
  - Null: Development time in high maturity organizations is greater than it is in low maturity organizations
  - Alternative: Development time in high maturity organizations is equal to or less than it is in low maturity organizations
How does M&A infrastructure Impact Stakeholders?

Customer satisfaction perspective
  • What are their views, their experiences?

Interviews, focus groups, and survey techniques
  • Is our sampling representative of the stakeholder groups?

What are the costs associated with M&A?
  • What are the costs (time, tools) associated with the M&A infrastructure?

What are the benefits?
  • What value do the stakeholders receive? Is it commensurate with the costs?

How can it be improved?
Outline

The Need for a Measurement and Analysis Infrastructure Diagnostic (MAID)
  • Why measure?
  • Measurement errors and their impact

The MAID Framework
  • Reference Model: CMMI and ISO 15939
  • Measure and Analysis Infrastructure Elements

MAID Methods
  • Process Diagnosis
  • Data and Information Product Quality Evaluation
  • Stakeholder Evaluation

Summary and Conclusion
Summary

Like production processes, measurement processes contain multiple sources of variation:

- Not all variation due to process performance
- Some variation due to choice of measurement infrastructure elements, procedures and instrumentation

Measurement Infrastructure Diagnostic:

- Characterizes performance of measurement system
- Identifies improvement opportunities for:
  - Measurement processes
  - Data quality
  - Stakeholder satisfaction/utility
MID Process Findings and Corrective Actions

- Missing or Inadequate
  - Processes and procedures
  - Measurement definition and indicator specification

- Incomplete stakeholder participation
- Failure to address important measurement goals

- Develop needed processes procedures and definitions
- Involve additional stakeholders
- Address additional measurement goals
MID Data Quality Findings and Corrective Actions

Frequently encountered problems include the following:

- invalid data
- missing data
- inaccurate (skewed or biased) data

Map the data collection process.

- Know the assumptions associated with the data.

Review base measures as well as indicators.

- Ratios and summaries of bad data are still bad data!

Data systems you should focus on include:

- manually collected or transferred data
- categorical data
- startup of automated systems
MID Stakeholder Findings and Corrective Actions

Information not used
Data too hard to collect
Mistrust of how data will be used

Check content, format, and timing of indicators and reports
Automate and simplify data collection
  - Tools and templates
  - Training
Visible and appropriate use of data
References


