Empirical Software Engineering
CSE 8340 — Fall 2002

Prof. Jeff Tian, tian@engr.smu.edu
CSE, SMU, Dallas, TX 75275
(214) 768-2861; Fax: (214) 768-3085
www.engr.smu.edu/~tian/class/8340.02f

Module IVa: Metrics Evaluation

• Metrics and Measurement in ESE

• Self and Empirical Evaluation

• Formal Model Based Evaluation
Overview: Measurement

To achieve the goal of controlled software development, we need to:

- Develop an *engineering* discipline;

- Measure and evaluate the working product;

- Construct a *scientific* model for program measurement:
  - Techniques from other disciplines;
  - Develop new techniques if necessary;
  - Basic questions:
    - What to measure: goal & environ.
    - How to measure it: metrics & tools
    - Selection and validation
Complexity Measurement

**Basic assumption:** The lower the complexity, the more desirable:

- cheaper to build;
- easier to maintain;
- more reliable;
- ...

**Usage of Complexity Measurement:**

<table>
<thead>
<tr>
<th>activity</th>
<th>time</th>
<th>nature</th>
</tr>
</thead>
<tbody>
<tr>
<td>assessment</td>
<td>a posteriori</td>
<td>passive</td>
</tr>
<tr>
<td>prediction</td>
<td>a priori</td>
<td>passive</td>
</tr>
<tr>
<td>control</td>
<td>persistent</td>
<td>active</td>
</tr>
</tbody>
</table>
Internal/External Measures

**Internal Measures:** depend on programs only. Complexity measures \( \subset \) internal measures;

**External Measures:** depend also on other external factors — so called -ilities.

**Relations:** correlated but not uniquely determined. To use internal measures to predict external measure, we need:

- Discover *appropriate* internal measures;
- Establish *predictive* relations;
- Use and validate predictions.
Measures and Dimensions

Complexity measures are multi-dimensional because of:

1. Multi-facet *internal* organization:
   - Presentation;
   - Control;
   - Data.

2. Multi-purpose *external* usage under different activities.
   - Assessment: Basili’s GQM;
   - Prediction: Boehm’s COCOMO;
   - Control: Boehm’s spiral.
Measurement in ESE

• Measurement: central activity in ESE
  ▶ context of measurement/expr/study
  ▶ measurements associated with different experimental designs
  ▶ measurement and data collection
  ▶ measurement result analysis
  ▶ measurement/analysis result presentation, interpretation, and drawing conclusions

• Interpreted as measurement activities:
  ▶ definition: context, design
  ▶ gathering: data collection
  ▶ analysis/followup:
    analysis, presentation, interpretation
Measurement Evaluation

- Measurement typically used to evaluate SE artifacts/activities.

- Also need to evaluate measurements/metrics themselves:
  - properly defined?
  - properly used?
  - lead to useful results?

- Use of evaluation results:
  - selecting existing measures/metrics
  - proposing new ones
  - under what context?
Measurement Evaluation

- Types of metrics evaluation:
  - self evaluation
  - empirical evaluation
  - formal model based evaluation

- Self evaluation of new metrics:
  - when proposed/defined
  - demonstrate the use & usefulness
  - possible subjective bias
  - limited scope & validity
Metrics Evaluation

- Empirical evaluation of metrics:
  - a set of given metrics
  - empirical study set up
  - focus: how these metrics work
  - other performance measures not subjected to evaluation
  - typical evaluation objects:
    - internal (complexity) metrics

- Evaluation based on formal models:
  - based on empirical studies/evidences
  - generalized theory/models
  - development: after many empirical evaluation studies
Formal Evaluation Models: Prather

- Prather’s axioms:
  - \[ m(S_1; S_2; \ldots; S_n) \geq \sum_i m(S_i) \]
  - \[ 2(m(S_1) + m(S_2)) \]
  - \[ \geq m(\text{if } P \text{ then } S_1 \text{ else } S_2) \]
  - \[ > m(S_1) + m(S_2) \]
  - \[ 2m(S) \geq m(\text{while } P \text{ do } S) > m(S) \]

- Observations/discussions:
  - earliest attempt on axiomatic model
  - some intuition captured:
    - e.g., interactions
  - limited scope
  - justification for some axioms?
Formal Evaluation Models: Fenton

- Fenton’s hierarchical complexity:

  ▶ $m(seq(F_1, \ldots, F_n))$
  \hspace{1cm} = g_n(m(F_1), \ldots, m(F_n))$
  ▶ $m(F(F_1 \text{ on } x_1, \ldots, F_n \text{ on } x_n))$
  \hspace{1cm} = h_F(m(F), m(F_1), \ldots, m(F_n))$

- Observations/discussions:

  ▶ general framework
    - too general?
  ▶ contrast with Prather’s work
  ▶ relation to later work
    - add specifics
    - measurement theory based work
Formal Evaluation Models: Weyuker

- Weyuker’s Desirable Properties:

1. \((\exists P, Q) \ (\forall(P) \neq \forall(Q))\)
2. \(\{P, \forall(P) = c\}\) is finite
3. \((\exists P, Q) \ (P \neq Q) \land (\forall(P) = \forall(Q))\)
4. \((\exists P, Q) \ (P = Q) \land (\forall(P) \neq \forall(Q))\)
5. \((\forall P, Q)\)
   \((\forall(P) \leq \forall(P;Q) \land \forall(Q) \leq \forall(P;Q))\)
6. \((\exists P, Q, R)\)
   \((\forall(P) = \forall(Q) \land \forall(P;R) \neq \forall(Q;R))\)
7. \((\exists P, Q) \ (P = \text{perm}(Q) \land \forall(P) \neq \forall(Q))\)
8. \((\forall P) \ (\forall x, y) \ \forall(P) = \forall(P^x_y)\)
9. \((\exists P, Q) \ (\forall(P) + \forall(Q) < \forall(P;Q))\)
Formal Evaluation Models: Weyuker

- About Weyuker’s properties:
  - more systematic treatment
  - inspired/lead to many followup work
    - positive: refinement
    - negative: counter examples
    - other: development & alternatives

- Tian/Zelkowitz as followup:
  - merit of Weyuker’s properties
  - some universally satisfied
    - basis for universal axioms
  - some for certain types of metrics
    - classification?
  - a theory based on the above
  - an evaluation/selection process
Formal Eval. Models: Tian/Zelkowitz

- Tian/Zelkowith Theory/Framework

- **Axioms:** Define program complexity and state common properties.

- **Dimensionality Analysis:**
  provide the basis for metrics classification
  - Aspects or dimensions:
    - presentation, control, data;
  - Levels: lexical, syntactic, semantic.

- **Classification Scheme:** Define mutually exclusive and collectively exhaustive classes.

- More details later.