11th Annual Practical Software and Systems Measurement Users' Group Conference

Measuring the Reliability and Value of a Checklist

Dan Houston, Ph.D.

July 26, 2007

Honeywell

Problem

Honeywell

- Software developers rely heavily on natural language instruments.
  - Task names and descriptions
  - Checklists and questionnaires
  - Development procedures
  - Defect categories

- Have you ever …
  - spent more time than necessary charging your time because task names didn't cover your work?
  - had to skip a checklist item because you couldn't figure out what it was asking for?
  - been unable to perform a procedure the way it was written?
  - argued over defect classification because the category descriptions weren't clear?

- Questions
  - Can we measure the reliability of natural language instruments such as a checklist?
  - How can the problematic items be identified accurately?

Outline

Honeywell

- Specific problem: defects leaking through test phase due to poor test specifications
- Context: a process improvement project
- The process analysis
- Process improvements
  - Producing a checklist
  - Measuring checklist reliability
  - Identifying checklist items to be improved
  - Improving checklist
- Checklist validation and project savings
- Other software development applications of subjective measurement system evaluation

DMAIC: Process Improvement

Honeywell

- **Analyze and measure process for variation**
  - **Uses qualitative and quantitative, especially statistical, tools.**
    - Subjective measurement system evaluation (MSE)
  - **Categorize inputs to process steps**
  - **Statistically characterize variation in process outputs**
- **Identify improvement opportunities**
- **Implement improvements and measure savings**

Phase	**Steps**
Define	Identify an opportunity and define a project to address it.
Measure	Analyze the current process and specify the desired outcome.
Analyze	Identify root causes and proposed solutions.
Improve	Prioritize solutions; select, plan, validate, and implement solution.
Control	Develop a plan for measuring progress and maintaining gains.

Test Specifications Project

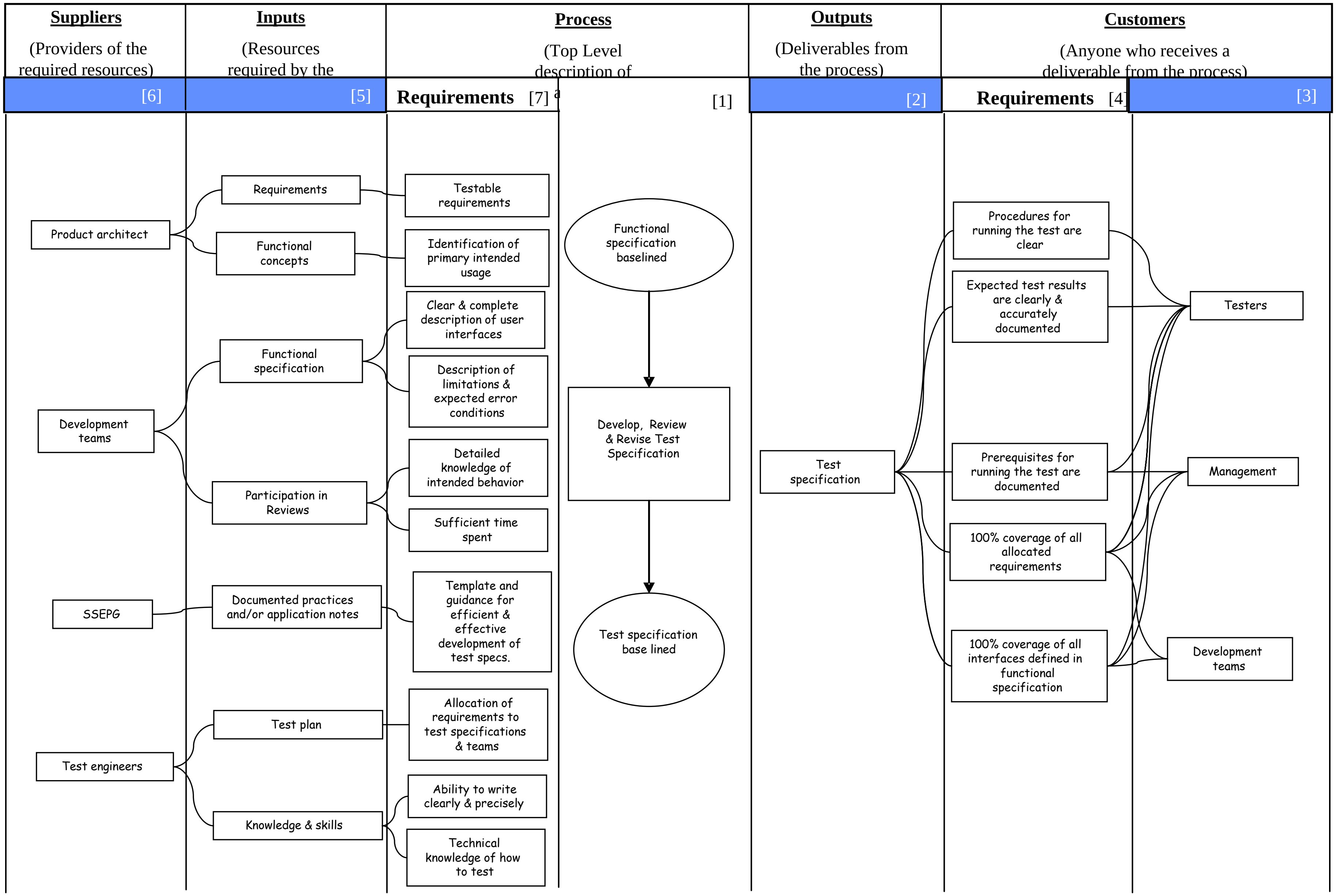
Honeywell

- Context: Fagan-style inspections of all work products
- System testers realized the need for guidance in reviewing test specifications.
  - Lack of content guidance caused concern about specification incompleteness.
  - Were defects passing through the system test phase?
- Project focus: test specification process
  - Emphasis on the quality of test specification content.
  - No savings were anticipated, but as the project progressed, the project team saw an opportunity to measure savings from use of the checklist.

Define	Measure	Analyze	Improve	Control
Problem statement SIPOC (supplier, inputs, process, outputs, customers)	As-is process map	Failure modes and effects analysis (FMEA)	To be process map Checklist drafted and reliability measured	Control plan Results

SIPOC Diagram

Honeywell



Test Specification Process Map

Honeywell

Write Test Specification → Inspect Test Specification → Revise based on Inspection → Baseline

Write Test Specification	Inspect Test Specification	Revise based on Inspection	Baseline
Output: Test Specification	Output: Log of Review Findings	Output: Revised Test Specification	Output: Baselined Test Specification
• n- Functional prototype • nx- DFS • nx- Author skill and allotted time • s- Framework Template for Test Specification • n- Test Plan (allocation of requirements & functionality to this Test Spec.) • nx- FWS part 2 • n- FWS part 3	• yx- Updated Test Specification • n- DFS • n- Test Plan • s- FrameWork review process • nx- Reviewers' skill and allotted time • c- Moderator's skill	• y- Test Specification • y- Log of review findings • n- Author's skill and allotted time • cx- Moderator's skill and alloted time	• Y- Revised Test Specification

INPUTS KEY
s - standard operating procedure
n - noise
c - controllable
y – output from previous step
* - Not implemented in current process

Failure Modes and Effects Analysis

Honeywell

- For each process step or step output, list potential failure modes
- For each failure mode,
  - list potential failure effects,
  - rate the severity of each failure effect, and
  - list the causes of each failure mode.
- Rate the likelihood of each failure mode, effect, and cause combination occurring.
- Assess current controls on each combination.
- Recommend actions for highest risks.
- Select improvements.
- Re-rate risks after improvements.

Test Specification FMEA

Honeywell

- Identified 39 failure modes
- Recommended actions for 28 failure modes
- Majority of the risks controlled by applying prior experience to ensure specification completeness.
  - Distill experience in a checklist.
  - Use different types of experts to review specific parts of a test specification.
- Identified five desirable attributes for test specification authors:
  - Analytical skills (identifying completeness of coverage with minimal redundancy)
  - Communications skills (clarity of instructions)
  - Customer usage knowledge
  - Technical systems knowledge (the architecture and interaction of components)
  - Testing experience

Improvements

Honeywell

- Process revisions
  - Specifications could be written incrementally and a draft could be inspected prior to baselining.
  - Test specifications can be revised and reviewed after execution.
- Checklist with type of expertise required for each item.
  - Needed to ensure the reliability of checklist
    - Is each item interpreted consistently?
  - Measure consistency of checklist usage
    - Have different raters use the checklist on the same specification: independently indicate whether the specification conformed to each item in the checklist.

Sample of the Test Specification Checklist Items

Honeywell

General

Reviewer	Checklist Item
Architect	1.1. Does the scope clearly specify the boundaries of the testing covered by this document?
Test expert	1.3. Does this spec tell the tester where to find all the files necessary to run each test?

Overall coverage

Reviewer	Checklist item
any	2.1 Are all requirements allocated by the test plan to this test team covered by this set of test cases?
(technical) Domain expert	2.4 Are there test cases with loads to stress the functionality to at least the level of the maximum realistic customer usage?

Individual test cases

Reviewer	Checklist Item
Test expert	3.1 Are the required files/databases and their location identified?

Nominal classification reliability*

Honeywell

κ (kappa) is defined as the proportion of agreement between raters after agreement by chance has been removed. The formula for κ, with two raters, is:

$$\kappa = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

Where

$P_{observed}$ is the proportion of units in which the raters agreed.

P_{chance} is the proportion of units in which agreement by chance is expected.

* Reliability estimates the interchangeability of judges by removing random measurement error variance.

Nominal classification reliability

Honeywell

For more than two raters,

$$\kappa_{overall} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} \overline{p}_j \overline{q}_j}$$

Where

x_{ij} is the number of ratings of the i^{th} unit in the j^{th} category

n is the number of units

m is the number of raters

k is the number of categories

$\overline{p}$ = ratings within a category / (n x m)

$\overline{q}$ = 1 - $\overline{p}$

κ calculation, first checklist

	Raters										
Item	PF	ES	KM	DS			1	0	$\sum_{i=1}^{2} x_{ij}^{2}$		
1.1	1	1	1	1			4	0	16		
1.2	0	0	1	0			1	3	10		
1.3	0	1	1	0			2	2	8		
1.4	0	1	0	0			1	3	10		
1.5	1	1	1	1			4	0	16		
1.6	0	0	1	1			2	2	8		
1.7	0	1	1	0			2	2	8		
2.1	0	1	1	0			2	2	8		
2.2	1	1	1	1			4	0	16		
2.3	1	1	1	1			4	0	16		
2.4	0	0	0	0			0	4	16		
2.5	0	0	1	1			2	2	8		
2.6	1	1	1	1			4	0	16		
2.7	1	1	1	0			3	1	10		
2.8	1	1	1	0			3	1	10		
3.1	0	0	1	0			1	3	10		
3.2	1	1	1	1			4	0	16		
3.3	0	1	1	1			3	1	10		
3.4	1	1	1	1			4	0	16		
3.5	1	1	1	1			4	0	16		
						sum	54	26	244		
n	20					$\bar{p}$	0.68	0.33			
m	4					$\bar{q}$	0.33	0.68		sum	κ
k	2					$\bar{p}\bar{q}$	0.22	0.22		0.44	0.28

κ calculation, revised checklist

	Raters								$\sum_{i=1}^{2} x_{ij}^2$		
Item	PF	ES	KM	DS			1	0			
1.1	1	1	1	1			4	0	16		
1.2	0	0	0	0			0	4	16		
1.3	1	1	1	1			4	0	16		
1.4	1	1	1	1			4	0	16		
1.5	1	1	1	1			4	0	16		
1.6	0	0	0	0			0	4	16		
1.7	1	1	1	1			4	0	16		
2.1	0	1	1	1			3	1	10		
2.2	1	1	1	1			4	0	16		
2.3	1	1	1	1			4	0	16		
2.4	1	1	1	1			4	0	16		
2.5	1	1	1	0			3	1	10		
2.6	1	1	1	1			4	0	16		
2.7	1	1	1	1			4	0	16		
2.8	1	1	1	1			4	0	16		
3.1	1	1	1	1			4	0	16		
3.2	1	1	1	1			4	0	16		
3.3	1	1	1	1			4	0	16		
3.4	1	1	1	1			4	0	16		
3.5	1	1	1	1			4	0	16		
						sum	70	10	308		
n	20					$\bar{p}$	0.88	0.13			
m	4					$\bar{q}$	0.13	0.88		sum	κ
k	2					$\overline{pq}$	0.11	0.11		0.22	0.77

Validation and Savings

Honeywell

- Validation
  - Used revised checklist to inspect a test specification that had already been used for testing.
  - Found and corrected specification deficiencies.
  - Used the revised test specification to run additional tests and found three high-priority defects.
- Savings
  - Estimated additional costs to fix defects found in the field.
  - For three defects, additional cost of leaked defects was estimated at $10,100.
- Further validation
  - Used the Test Specification Checklist to re-inspect another test specification.
  - Additional testing with the second revised test specification discovered two more defects at the same time they were being discovered by customers in beta testing.

Components of Savings Calculation

Honeywell

- **Defect management costs**
  - **# defects found * (total defect effort / # defects) * burdened rate**
- **Rework costs**
  - **Effort to fix defects found * burdened rate**
- **Release costs**
  - **Cost of release * Probability of release due to a high priority defect**
    - Cost of release
      - Management at project and program levels
      - Release management
      - Software configuration management
      - Product and system tests (planning, testing, analysis, and reporting)
      - Media verification and documentation
      - Installation documentation
    - Probability of release is calculated from problem database and release records
- **Less improvement project cost**
- **Additional unmeasured costs avoided**
  - **Schedule impact**
  - **Customer dissatisfaction**
  - **Contracted customer support**

Other uses of κ in software development

Honeywell

- Process improvement: classification of process inputs
  - Mapped process across sites, then independently classified inputs to each step. Low κ value for critical inputs. Discovered differing perspectives between sites on criticality of inputs (product knowledge and resolution options) to a software rework process.
- Process capability assessment instruments*
- Project planning: test a project risk classification scheme
- Project tracking: test activity/task labels for time charging
- SQA: test a project's application of documented software processes
- Process reliability: test multiple projects' interpretation and use of a procedure (projects' usages are the raters)
- Usability: test usability questionnaire
- Defect management: test defect classification schemes

* Khaled El Emam. 1998. Benchmarking Kappa for Software Process Assessment Reliability Studies. International Software Engineering Research Network Technical Report ISERN-98-02. Available at http://www.ehealthinformation.ca/documents/isern-98-02.pdf (June 2007)

Conclusions

Honeywell

- Time and energy can be wasted using unreliable instruments due to:
  - Missing or incomplete items
  - Ambiguous items
  - Unclear or meaningless items

- Measuring the reliability of assessment instruments, questionnaires, and nominal categories prior to widespread usage …
  - can identify problems items in the instrument,
  - provides a basis for improving the instrument,
  - engenders confidence in and encourages use of the instrument, and
  - avoids rework, frustration, and wasted time.

Resources

Honeywell

- David Futrell. 1995. When quality is a matter of taste, use reliability indexes. *Quality Progress* 28: 5 (May), 81-86.
  - This article is a practical guide for applying both the kappa and the intraclass correlation techniques.

The following articles are recommended for further study of **κ** and other interrater agreement measures.

- Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 37-46.
  - Presents the kappa coefficient and discusses its statistical characteristics.
- Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, Debajyoti Sinha. 1999. Beyond Kappa: A Review of Interrater Agreement Measures. *The Canadian Journal of Statistics* 27:1 (Mar) 3-23.
  - Reviews and critiques various approaches to the study of interrater agreement, for which the relevant data comprise either nominal or ordinal categorical ratings from multiple raters.