PRACTICAL APPROACHES TO SOFTWARE METRICS

Corporate Presentation

March, 2013

Cem Kaner

Florida Institute of Technology

These notes are partially based on research supported by NSF Grant CCLI-0717613 "Adaptation & Implementation of an Activity-Based Online or Hybrid Course in Software Testing." Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Copyright (c) Cem Kaner 2013

4 Things to Remember

1. Yes, most software metrics are (to some degree) invalid. However, that doesn't reduce the need for the information we are trying to get from them.
2. I think it's part of the story of humanity that we've always worked with imperfect tools and always will. We succeed by learning the strengths, weaknesses and risks of our tools, improving them when we can, and mitigating their risks.
3. We need to look for the truths behind our numbers. This involves discovery and cross-validation of patterns across data, across analyses, and over time. The process is qualitative.
4. Qualitative analysis is more detailed and requires a greater diversity of skills than quantitative. Qualitative analysis is not a free (or even a cheap) lunch. We evaluate the quality of quantitative measures by critically considering their ***validity***. An equally demanding evaluation for qualitative measures considers their ***credibility***.

How I Teach Metrics at School

Core task in the course = a pair of research essays

- I create a core task because students learn what they do
- This course has the usual Kaner-style exams (essay questions drawn from a study guide with, currently, 121 essay questions)
- But the key task is research and write-up for the essays.
  - Pick one metric (your choice)
    - Choose a widely-used metric that seem-to-be-famous people say is a good one
  - Apply it to some data to see how it works
  - Find out how thoroughly the metric has been researched and evaluated
    - Dig through the scientific literature
    - Dig through the practitioner web

Copyright (c) Cem Kaner 2013

Essay: How Does the Literature Answer These Questions?

Describe the metric

- What attribute does it measure?
- What other constructs does the metric rely on?

Describe the operations involve in computing the metric

- Compute it. What happened?
- What instruments do you use?
- How do you take the readings?
- What are the scales of the attribute and the readings?
- Are results commensurable across projects? Companies?

Evaluate the metric's construct validity

- What evidence is there that this measures the attribute?
- What model allows us to map values of the measure to values of the attribute?
- What research has been done on key threats to construct validity? Evaluate it.

Evaluate the metric's operational validity

- Were problems / risks noted in the literature?
- What problems did you consider when applying the metric?
- What research has been done on key threats to operational validity? Evaluate it.

Evaluate the metric's generalization validity

- To what extent does research on this metric support or raise concerns about generalization
- What research has been done on key threats to generalization validity? Evaluate it.

What are the sources of error in the value of the metric and what causes them?

- What directional errors (biases)?
- Variance of the attribute? Of the measurements?
- Do intended use, potential consequences of use, or scope of use create bias?

Are there natural / foreseeable negative side effects?

- Distortion or dysfunction caused by intended use, potential consequences of use, scope of use, or by deflecting attention from something else?

Flow of the Course: First third

- Prepare students to write the essay.
- They need
  - Lectures on measurement theory
    - Representational theory of measurement
    - Properties of numbers
    - Constructs, attributes, and validity
  - Lectures on attributes of software quality
  - Lectures on issues (giving rise to metrics) in software project management
  - Library research skills (we work with a research librarian)
  - Quick overviews of measurement distortion, dysfunction, and of threats to validity of a metric

FLOW OF THE COURSE: TRANSITION 1ST TO 2ND

- They have completed Midterm Exam #1 and got feedback
- They have completed the essay
  - We did interactive grading. They each got personal coaching for 1-3 hours plus grading against a detailed rubric
  - They found it very hard to find information needed to answer most of the questions. Issues were often
    - Not mentioned by anyone, or
    - Mentioned reassuringly but not considered, or
    - Discussed briefly / reassuringly but not researched, or
    - Allegedly researched but little or not detail on the experiments or the data
- They know they have to do better on the next essay. How?

Copyright (c) Cem Kaner 2013

Flow of the Course: 2nd

We read Bossavit

Bossavit wrote the book to encourage better scientific practices (how to do research, how to report it)

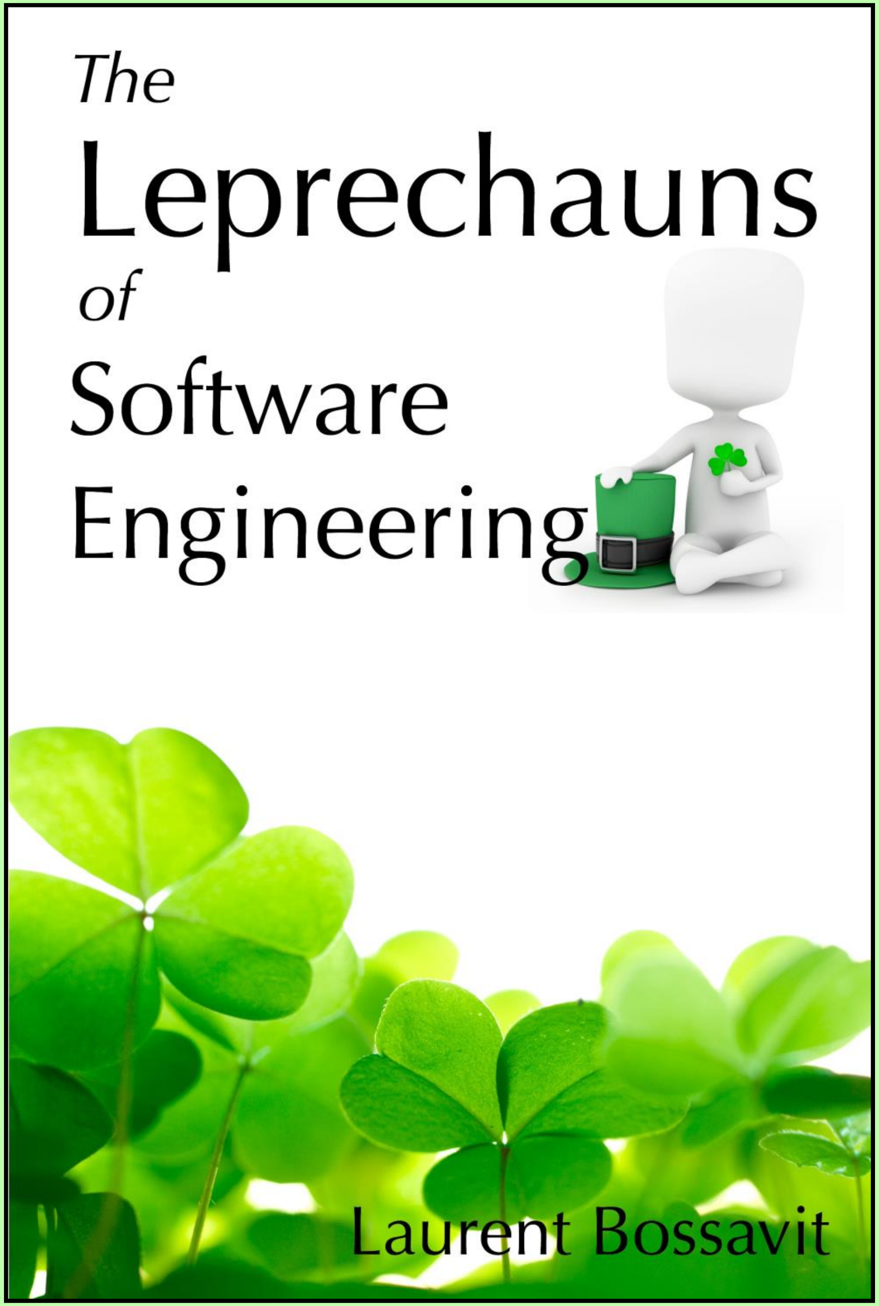
He exposes serious problems in quantitative claims about software engineering that are allegedly based on research

For example

- Present "results" of research when no actual research was done
- Exaggeration
- Citations of sources that are impossible to find
- Appeal to authority

Let me add:

- Emotional charge (e.g. guilt)
- Selective presentation (present only confirmatory data or only opposing data/arguments that are weak)
- And, what we can't say in print: Fiction



Copyright (c) Cem Kaner 2013

FLOW OF THE COURSE: 2ND THIRD

- Detailed discussion of measurement validity
  - **A measurement is valid to the extent that it provides a trustworthy description of the attribute being measured.**
  - Validity might be "threatened" (be less trustworthy as a descriptor of the attribute) in several ways
    - Conceptually weak – very common to talk about a measure without carefully considering what it is that you think you are actually trying to measure
    - Operationally weak – how you *do* the measurement is weak or sloppy or in some other way misleading
    - Can't be generalized to other projects, other teams, other companies, other industries.

THE PROBLEM OF SURROGATE MEASUREMENT

Austin notes that in many cases of dysfunction, the employees were evaluated against a surrogate measure:

- The “true” measure of what the employer wanted would be too difficult or too expensive
- The employer measures something that is probably correlated with the desired attribute
- The employee optimizes performance against the measure, rather than against the attribute
- The result is a good number but poor underlying value

SURROGATE MEASUREMENT

I started these notes after discussions with

- Mark Johnson (see his M.Sc. Thesis) & with
- Shari Pfleeger (see tutorial notes *Evaluating Software Technology*, presented at ICSE 97 (19th International Conference on Software Engineering), http://dl.acm.org/citation.cfm?id=253778 (this discusses the need for models in measurements)

I think of surrogates as

- providing unambiguous assignments of numbers according to rules
- but they don't provide an underlying theory or model that relates the measure to the attribute allegedly being measured.

Measurement: The empirical, objective assignment

- of numbers
- to attributes of objects or events
- according to a rule
- derived from a model or theory
- with the intent of describing them.

(Kaner & Bond, 2004)

Surrogate measure: The empirical, objective assignment

- of numbers
- to attributes of objects or events
- according to a rule
- ~~derived from a model or theory~~
- with the intent of describing them.

Copyright (c) Cem Kaner 2013

SURROGATE MEASURE

I haven't seen many definitions of surrogate measure or proxy measure (search Bing or Google, there are surprisingly-to-me few), so here is mine

A surrogate (or proxy) measure is a type of measure whose values seems to be correlated with an underlying attribute, but we have no trusted theory or model to quantify the correlation or to specify the conditions under which the correlation occurs.

We normally use surrogates

- when they are much less expensive than better measures, or
- as converging measures to assess the trustworthiness of a proposed measure, model, or experimental result.

Copyright (c) Cem Kaner 2013

SURROGATE MEASURE

Shorter definition:

- A surrogate measure is easy or cheap to collect, seems self-evidently related to the attribute of interest, but cannot be tied to the attribute with a quantitative theory or model.

Typical strategies for validating surrogates:

- Correlate them with measurements that have more-easily-defended construct validity
  - Example: F.W. Lipert, R.E. Wyzga, J.D. Baty, & J.P. Miller (2006) "Traffic density as a surrogate measure of environmental exposures in studies of air pollution health effects: Long-term mortality in a cohort of US veterans", Atmospheric Environment 40(1) 154-169, http://www.sciencedirect.com/science/article/pii/S1352231005008459
  - Example: D. Gettman & L. Head (2003) "Surrogate Safety Measures from Traffic Simulation Models", United States Federal Highway Administration http://www.fhwa.dot.gov/publications/research/safety/03050/03050.pdf
  - See discussions of the Transportation Research Board's Safety Data, Analysis & Evaluation (ANB20) committee, https://sites.google.com/site/trbanb20/home, such as https://wiki.umn.edu/view/TRB_ANB203/

Copyright (c) Cem Kaner 2013

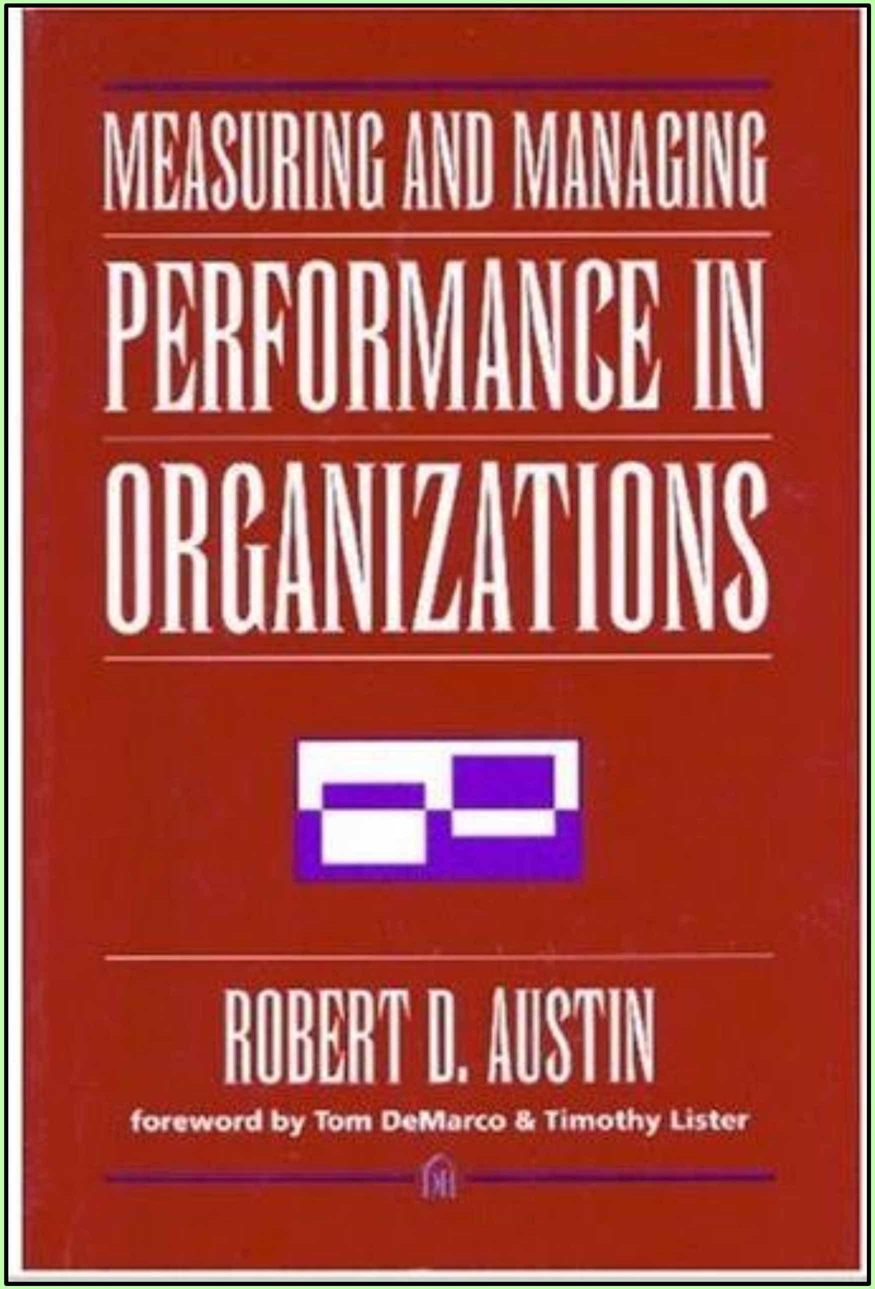
Flow of the Course. 2nd

Measurement distortion –introduction of measurement changes how people work in ways that improve the numbers but subtract value from what is not being measured

Measurement dysfunction – distortion, but to the extent that the overall value of the work has declined. The numbers look better but overall performance is worse

Examples

- Employment agency
  - Rewarded staff for # interviewed
  - Lots of interviews, no time spent recruiting employers, no one hired
- Reward testers for # bugs found
  - What happens to teamwork, bug troubleshooting, training of other testers, support of other groups, test documentation …



MEASURING AND MANAGING PERFORMANCE IN ORGANIZATIONS

ROBERT D. AUSTIN

foreword by Tom DeMarco & Timothy Lister

EXAMPLES OF DYSFUNCTION (SAMPLED FROM AUSTIN, 2004 AND HOFFMAN, 2000)

- In a federal law enforcement agency, agents with case quotas prioritized easy-to-solve cases over larger and more important cases
- Police departments evaluated by proportion of crimes solved chose not to record some citizen complaints or to wait to post them until the crime was solved
- Software development groups evaluated by proportion of bugs fixed chose to keep bugs out of the bug tracking system until they were fixed or to reclassify them as duplicates of other bugs. As a result, some bugs were lost.
- Programs were shipped with missing features because 100% code coverage was a requirement for shipment and had been achieved (all statements *in the program* had been tested).
- Teachers narrowed their course coverage to focus only on the areas tested in high-stakes tests.
- Commissioned staff at automobile service centers pushed customers into unneeded repairs or charged for repairs not done. Discovery of this created significant regulatory costs and negative market impact.

EXAMPLES OF DYSFUNCTION

- People might color or falsify the data you get
- People might stop doing important but unmeasured tasks
  - Managers reassign people who persist in important but unmeasured tasks or who persist in reporting undesired results
- Counts of undesirable things go down because activity is reduced, not improved. Examples:
  - Do less testing because you are writing more status reports: find and report fewer bugs.
  - Delay a critical kind of testing until after a milestone or until someone goes on vacation
- People delay or underreport "undesirable" things (causing consequences of not dealing with them)

Copyright (c) Cem Kaner 2013

EXAMPLES OF DYSFUNCTION

- People might create problems that they can then get credit for fixing
- Emphasis on individual performance and productivity can reduce
  - Coaching
  - Collaboration
  - Time invested in building tools, especially tools used by the broader group
- People might increase measured activities in ways that introduce unmeasured risks
  - Examples: drive to greater apparent productivity can wear out equipment from overuse, burn out people, yield errors, cause accidents

DYSFUNCTION NEED NOT INVOLVE "GAMING" THE SYSTEM

- In my experience (and Hoffman's), people sometimes did dysfunctional things because they believed they were doing what they were supposed to do
  - They weren't taking shortcuts
  - They were adapting to what they believed were management priorities or management-preferred practices
- Examples
  - Presentation from a manager at Microsoft on relationships between testers and programmers
  - Pressure on a programmer to check in code sooner
    - Is technical debt a bug or a feature?
  - Rating of testers by number of bugs reported
  - Competitive peer reviews
    - Is poor collaboration an objective or a dysfunction?

Copyright (c) Cem Kaner 2013

Effects versus side-effects

When you set up a measurement system for managing human performance:

- ***The system will have some effects***.
  - People will change what they do in order to improve their measured results.
  - This is what we do expect, and should expect, in any work environment.
- ***The system will have some side-effects.***
  - Changing behavior has consequences
  - People reprioritize their tasks to achieve the goals you set for them.
  - People might take shortcuts or risks or simply not do things in order to achieve the goals you set for them
  - The effects that you did not intend when you set up the system, but which happen anyway, are the side-effects.
- ***The side-effect is unintended, probably unexpected, and probably undesired.***

SURROGATE MEASURES

- NO WAY TO GAUGE THE DEGREE OF CONSTRUCT VALIDITY
- HARD TO GAUGE INTERNAL OR EXTERNAL VALIDITY.
  - This was the theme of Johnson's thesis
  - For medical literature discussion of similar issues, see http://www.ganfyd.org/index.php?title=Surrogate_outcome_measure
- The performance structure (reward for improving the surrogate numbers) ***encourages*** dysfunction:
  - The employee is rewarded for improving the proxy
  - To the extent that the employee improves the underlying attribute without improving the proxy, the employee is penalized
  - It seems unreasonable to blame the employee for focusing on the proxy at the expense of the attribute. This is exactly what the employer's reward structure tells him to do.

Copyright (c) Cem Kaner 2013

RECAP ON DISTORTION AND DYSFUNCTION

People change what they do in response to how they are measured. This is normal. We expect this. It is the basis of measurement-based management.

But if people have finite time, and they give you more of what you measure, where do they cut back? What do you lose in order to obtain these perceived gains?

- **Measurement distortion**: An effect of taking these measurement is to change the system in ways that are undesirable. Example: reallocate resources in ways that starve an unmeasured task
- **Measurement dysfunction**: The measurement distortion is so bad that the system-under-measurement looks better than before measurement but is actually worse than it would have been without measurement.

Surrogate measures increase the risk of dysfunction because they explicitly incent the employee to prioritize work on the wrong things.

Copyright (c) Cem Kaner 2013

At the Root of the Problem

- When you measure the length of a table,
  - It can’t change its length to please you
  - It can’t even want to change its length, or to please you
- When you measure human performance
  - The humans can want to please you (or not)
  - The humans can change what you measure

Production and process measurement
Of things
Is fundamentally different
From production and process measurement
Of people

Copyright (c) Cem Kaner 2013

MOST SOFTWARE ENGINEERING METRICS ARE HUMAN PERFORMANCE METRICS

- Bug counts
- Programmer or team productivity
- Customer satisfaction
- Software complexity
- Software usability, maintainability, and all other measures of software design quality and software implementation quality
- Project cost (in time or money)
- Project cost compared to budget

All of these reflect the work of people who designed, wrote, tested and managed development of the code. All of them can be used, and have been used, to compare, reward or punish individuals or teams.

Management By Objectives

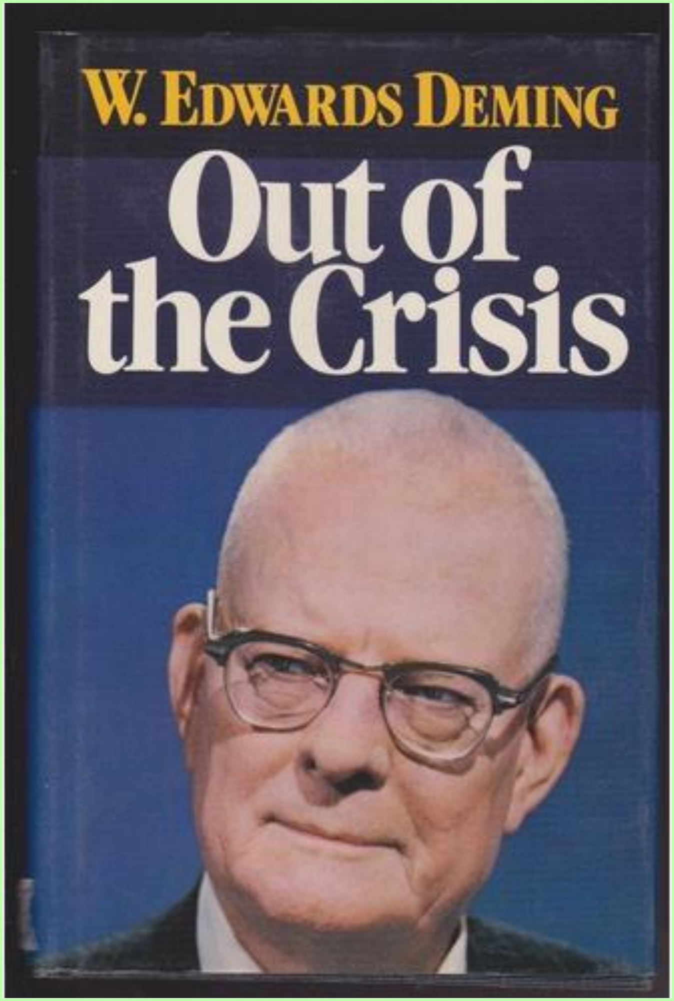
Peter Drucker, one of the most influential management theorists of the 20th century, advocated Management By Objectives – heavy emphasis on employee performance measurement.

The ideas underlying the MBO review are:

- Multidimensional review: weighted score on several (perhaps 5) dimensions
- Clear communication of priorities
- Accountability and rewards for success
- If objectives are set in collaboration with the employee, the process may be positively motivating
- Protection against dysfunctional effects from focus on a single dimension. For example, an executive who focuses too much on cost-cutting will probably score poorly on client satisfaction.

Many Argue that the use of Human Performance Measurement in Business is Inevitably Toxic

- W. Edwards Deming
  - Key 20th century advocate of statistical process control
- considered this one of the Seven Deadly Diseases of the American management style
  - Created management by fear
  - Created emphasis on those aspects of the job that could be easily quantified
  - Created emphasis on short-term results
  - Undermines teamwork and diverts effort into competition



Copyright (c) Cem Kaner 2013

Battle of the Titans

- We have a total opposition between two giants (arguably the most significant thinkers in Management (Drucker) and Quality (Deming).
- The disagreement is broad, with many respected thinkers on each side, with deeply held views on all sides.

Flow of the Course: Final third

- So where are we?
- They're completing Essay #2
  - The literature hasn't improved
  - Their research skills are better
  - They find a bit more relevant data
  - And when they don't find data, they know it's not there.
- The research literature is very, very weak
- The claims are overblown

- And the metrics might be doing more harm than good

- It is very tempting to give up

> “METRICS THAT ARE NOT VALID ARE DANGEROUS.”

CEM KANER
JAMES BACH
BRET PETTICHORD
LESSONS LEARNED IN SOFTWARE TESTING

Hmmm, are these guys telling us that we SHOULD give up?

Copyright (c) Cem Kaner 2013

On the Other Hand, Managers Have Legitimate Needs

- They need metrics in order to (for example…)
  - Compare staff
  - Compare project teams
  - Calculate actual costs
  - Compare costs across projects or teams
  - Estimate future costs
  - Assess and compare quality across projects and teams
  - Compare processes
  - Identify patterns across projects and trends over time

- Executives need these, whether we know how to provide them or not.
  - Hung Quoc Nguyen

Copyright (c) Cem Kaner 2013

Flow of the Course: Final third

What do people in other fields do?

Hmmm,

Let’s study some Financial Analysis

Metrics in Finance

Bank of America (BAC)

- Summer 2012

• Assets (book value) per share	$19.83
• Price per share	$ 7.82
• “Price to Book” Ratio	0.39

- According to these statistics, if you closed BAC and sold its assets, you could get nearly 3x as much as it is worth as a running company.

http://finance.yahoo.com/q/ks?s=BAC+Key+Statistics

Copyright (c) Cem Kaner 2013

CONTRAST

	Bank of America	Wells Fargo
• Assets per share	$19.83	$25.70
• Price per share	$ 7.82	$33.91
• Price to Book Ratio =	0.39	1.32

What's going on?

http://finance.yahoo.com

Copyright (c) Cem Kaner 2013

PERHAPS BAC'S "BOOK VALUE" IS INCREDIBLE

- Maybe no one will pay book value for BAC because they don't believe its estimates?
  - Foreclosed houses – what are they worth?
  - How many loans are bad?
  - How does this compare to its loan loss reserves?
- I don't have knowledge of BAC's assets, but a large discrepancy between book value and market value can suggest that someone believes that someone is gaming the numbers

Copyright (c) Cem Kaner 2013

FINANCIAL RATIOS

- Price to earnings ratio – how much you pay for each dollar of earnings.
- Price to book ratio – how much you pay for each dollar of assets
- Price to sales ratio – how much you pay for each dollar of gross revenue

Copyright (c) Cem Kaner 2013

FINANCIAL RATIOS

P/E, P/S, and P/B are all
widely used by investors,
including well-informed professionals

Copyright (c) Cem Kaner 2013

FINANCIAL RATIOS

Almost no one thinks they are valid

Copyright (c) Cem Kaner 2013

FINANCIAL RATIOS

I don’t think they are valid

Copyright (c) Cem Kaner 2013

Financial ratios

I use them every day

Copyright (c) Cem Kaner 2013

Triangulation: Hunting patterns in weak data

For me, the key to working with a financial ratio is understanding **what that's supposed to tell me about**.

For Price / Book, the underlying concept is how much asset I get for my money. If the company is at risk, this is important.

But if I am actually concerned about that, I look at other indicators of the company's assets and who else has claims against them:

- What potential losses are on the horizon?
- How much do they owe?
- When are those debts payable?
- What challenges have been made to the valuations?
- What history does the company have of surprising revaluations?

Taken together, the collected data might tell a useful story.

Copyright (c) Cem Kaner 2013

LET ME PUT THIS A DIFFERENT WAY

- The numbers aren't very good
  - People can game them (and they do)
  - People can distort their business in order to make the numbers look better (and they do)
  - Companies with "great" valuation ratios can be terrible investments for years and years (we call them "value traps")
- When an experienced investor analyzes a company's metrics
  - She reviews data over several years
  - She reviews many types of data
  - **She looks for the truth behind the data**
- But even though the individual numbers are weak, they can be useful on their own (e.g. "too much debt!") or as flags that signal that it's time for a deeper look

When I say that these metrics can be useful, even though they are not very good, some people tell me I am contradicting myself.

> “Metrics that are not valid are dangerous.”

Cem Kaner
James Bach
Bret Pettichord
Lessons Learned in Software Testing

Copyright (c) Cem Kaner 2013

DANGEROUS = UNUSABLE?

- It's snowing today.
- I'm from Canada. When it snowed, we used to think about doing some cross-country skiing.

Copyright (c) Cem Kaner 2013

DANGEROUS = UNUSABLE?

- It's snowing today.
- I'm from Canada. When it snowed, we used to think about doing some cross-country skiing.
- Did you know?

Skiing is dangerous!!!

Skiing is dangerous!!!

Maybe we should never ski

Fire is dangerous!!!

Maybe we should never use matches

Cars are dangerous!!!

Maybe we should never drive

Maybe we should never drive !! ? ! ? !!

How would Flo feel about that?

Copyright (c) Cem Kaner 2013

OK, SO METRICS ARE DANGEROUS BUT WE NEED THEM ANYWAY

So how can we do that?

Anyone got a spare magic wand?

- Maybe we could measure more directly. Like, get rid of all surrogates!
  - "Not everything that counts can be counted" (Einstein)
  - If we don't include the things that are hard to measure, those are what will get prioritized out when we measure the other stuff
- Maybe we could use a magic dashboard
  - Use colors instead of numbers, call it "qualitative measurement" (that's a shiny new buzzword)
  - What are those colors based on? Why should a skeptical executive trust them? Why should anyone trust them?

What Can We Do?

- Multidimensional measurement, like balanced scorecard?
  - Example: If a pharma company measures only profits, they can increase profits today by cutting R&D. But a few years later, their drugs come off patent and they've got nothing to sell. Who would want that?
    - Some companies did that to themselves and are in deep doodoo today? Ooops
    - Maybe if they tracked R&D successes as another metric, people would notice if someone cut R&D out to improve the profit metrics
  - Maybe if we track a few key areas, we can counterbalance the risks of focusing on just one
- Austin (and others) attack this too
  - We still miss key areas and the dimensions don't work for everyone

CURRENT EXAMPLE: CITIBANK, MARCH 2013

Wall Street Journal (Kapner, 2013), article opens:

> "Four months after taking over as chief executive of Citigroup Inc., Michael Corbat is putting his stamp on the company with a simple formula: You can't manage what you can't measure. … 'You are what you measure,' [Corbat]"

According to the article, Corbat will grade executives using a composite measure based on scores on five dimensions:

- Capital
- Clients
- Costs
- Culture and
- Controls.

Copyright (c) Cem Kaner 2013

BUT MAYBE THERE'S A PROBLEM

"Mr. Corbat's plan to roll the score cards out to all divisions has raised objections from some executives who view the strategy as too much of a "one-size-fits-all" approach, said people familiar with their thinking....

They have argued that technology, legal, risk-management and other divisions don't lend themselves to this type of measurement, these people said."

Capital, clients, costs, culture and control might be great for those sales guys (or someone) but how will a structure like this change the performance of IT? For the better?

Copyright (c) Cem Kaner 2013

I DON'T THINK THERE'S AN EASY SOLUTION

Some principles:

- **Pay attention to your tools and their quality**
  - For example, be aware (and skeptical) of the relationship between the measurement and the attribute you're trying to measure
    1. Why should we think that THIS measures THAT?
    2. What would the best arguments be that THIS does not measure THAT or that THIS is a very bad measure of THAT?

Copyright (c) Cem Kaner 2013

I DON'T THINK THERE'S AN EASY SOLUTION

Some principles:

- **Present sets of points, showing relationships**
  - – Show values of the same variable(s) over time
  - – Show alternative variables or alternative analyses that "should" show the same patterns or lead to the same conclusions
    - ▪ Do the alternatives present consistent pictures?
    - ▪ If they disagree, your theory that they "should be" equivalent is mistaken. Look deeper.
  - – Show relationships of variables that should affect each other

I DON'T THINK THERE'S AN EASY SOLUTION

Some principles:

- **Tie your analyses to value**
  - If you're writing sales-support software,
    - How do your measures tie in with the success of the sales reps? Have your improvements
      - » Yielded sales improvements over time?
      - » Made the sales reps happier over time?
      - » Made sales transactions complete more quickly over time?
      - » Reduced sales-rep errors over time?
      - » Reduced customer dissatisfaction with erroneous sales claims over time?

I DON'T THINK THERE'S AN EASY SOLUTION

Some principles:

- **Measure different things for different people / groups**
  - Black box tester who primarily finds and reports bugs
  - Technology advocate who writes test tools that others use to find bugs, and trains them to use the tools
  - Test designer for tests to be done by inexperienced staff (heavy emphasis on scripting, need to anticipate and script for maintenance errors because the script followers won't know how to usefully improvise)
- **Ask how well people fulfill the job description of THEIR job, not of some generic**

I DON’T THINK THERE’S AN EASY SOLUTION

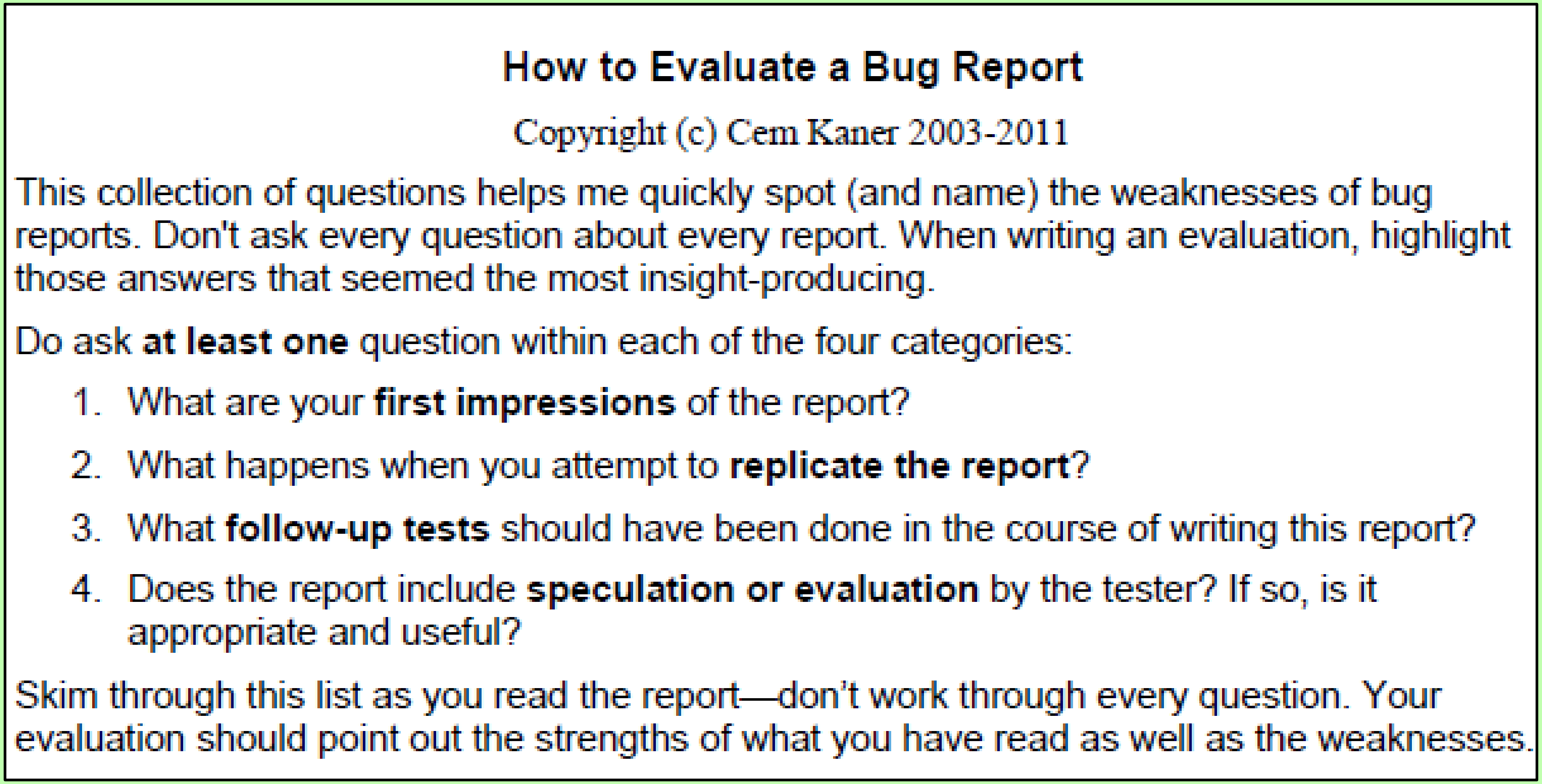
Some principles:

- **Pay attention to side effects**
  - Actively look for indicators that something unexpected / undesirable is happening
  - Some side effects are foreseeable
  - Some side effects show up as anomalies – unexpected changes to things that shouldn’t be changing – so ask what’s going on
  - You might do this with numbers. You might do this with “management by walking around”. However you do it, make it part of your process.

I DON'T THINK THERE'S AN EASY SOLUTION

Some principles:

- **For complex tasks, use a multidimensional rubric that allows a diversity of approaches to be recognized as successful**



How to Evaluate a Bug Report

Copyright (c) Cem Kaner 2003-2011

This collection of questions helps me quickly spot (and name) the weaknesses of bug reports. Don't ask every question about every report. When writing an evaluation, highlight those answers that seemed the most insight-producing.

Do ask **at least one** question within each of the four categories:

1. What are your **first impressions** of the report?
2. What happens when you attempt to **replicate the report**?
3. What **follow-up tests** should have been done in the course of writing this report?
4. Does the report include **speculation or evaluation** by the tester? If so, is it appropriate and useful?

Skim through this list as you read the report—don't work through every question. Your evaluation should point out the strengths of what you have read as well as the weaknesses.

YOU ARE CREATING MULTIDIMENSIONAL STORIES

Welcome to qualitative measurement

“**All quantitative data is based on qualitative judgment.**”

Trochim
http://www.socialresearchmethods.net/kb/qualdeb.php

Copyright (c) Cem Kaner 2013

YOU ARE CREATING MULTIDIMENSIONAL STORIES

"My belief is that the heart of the quantitative-qualitative debate is philosophical, not methodological. Many qualitative researchers operate under different **epistemological assumptions** from quantitative researchers. For instance, many qualitative researchers believe that the best way to understand any phenomenon is to view it in its context. They see all quantification as limited in nature, looking only at one small portion of a reality that cannot be split or unitized without losing the importance of the whole phenomenon. For some qualitative researchers, the best way to understand what's going on is to become immersed in it. Move into the culture or organization you are studying and experience what it is like to be a part of it. Be flexible in your inquiry of people in context. Rather than approaching measurement with the idea of constructing a fixed instrument or set of questions, allow the questions to emerge and change as you become familiar with what you are studying."

Trochim
http://www.socialresearchmethods.net/kb/qualdeb.php

Copyright (c) Cem Kaner 2013

Quality Criteria for Qualitative Reports

Based on Guba & Lincoln (1989)

- When you describe qualitatively, you are describing your perceptions, your conclusions, your analysis. You back it up with examples that you choose, quotes that you choose, data that you choose.
- Why should someone else trust your work?
  - Do you know what you’re talking about?
  - Did you collect the data in a reasonable way?
  - Are you summarizing the data fairly?
  - How are you managing your biases (people are often not conscious of the effects of their biases) as you select and organize your observations?
  - Are you prone to wishful thinking or to trying to please (or displease) people in power?

Copyright (c) Cem Kaner 2013

Quality Criteria for Qualitative Work

Just as we can question the quality of a traditional measurement or set of measurements by considering its **validity**

> **A measurement is valid to the extent that it provides a trustworthy description of the attribute being measured.**

We can question the quality of a qualitative analysis / report in terms of its **credibility**

Copyright (c) Cem Kaner 2013

Quality Criteria for Qualitative Work

Credibility:

- What makes you an expert in this? Why should I be interested in your observations or opinions?
- Prolonged engagement
- Persistent observation
- Peer debriefing
- Negative case analysis: for example, how well and how often and how willing are you to revise working hypotheses in the light of hindsight
- Progressive subjectivity: as you observe situations or create and look for data to assess models, how much are you paying attention to your own expectations versus how much are you paying attention to the expectations and observations of others

Quality Criteria for Qualitative Work

Credibility:

- Member checks: If you are observing / measuring / evaluating others, how much do you involve them in the process? For example, who creates your rubrics, and how much influence do reviewers have?

Copyright (c) Cem Kaner 2013

QUALITY CRITERIA FOR QUALITATIVE WORK

Transferability

- Somewhat like generalization validity
- How well would your conclusions apply in a different setting?
- How likely would it be that people would make similar observations in another setting?

- Thorough description is the key element for some researchers. You might not be able to predict your generalizability (whether people in other settings will see the same things) but you might be able to describe what you see well enough to help them recognize that they are seeing things very similar to what you were seeing.
- Over time, a sense of how general something is can build as multiple similar observations are recorded in different settings

QUALITY CRITERIA FOR QUALITATIVE WORK

Dependability

- Is your work methodologically sound?
- Qualitative work is more exploratory than quantitative (at least as quantitative is described). You change what you do as you learn more or as you develop new questions.
- Therefore consistency of methodology is not an ultimate criterion in qualitative work, as it is for some quantitative work
- But we can still ask how well (methodologically) you do your work. For example
  - Do you have the necessary skills and are you applying them?
  - If you lack skills, are you getting help?
  - Do you keep track of what you're doing and make your methodological changes deliberately and thoughtfully?

Copyright (c) Cem Kaner 2013

QUALITY CRITERIA FOR QUALITATIVE WORK

Confirmability

- If someone else worked through your data
  - Would they see the same things as you?
  - Would they generally agree that things you see are representative are representative and things that are idiosyncratic are idiosyncratic?
  - Would they be able to follow your analysis, find your records, understand your ways of classifying things and agree that you applied what you said you applied?

Copyright (c) Cem Kaner 2013

QUALITATIVE WORK

Qualitative measurements tell a story (or a bunch of stories)

- Anyone can tell a story
- Some people can tell persuasive stories even though they don't have much basis, but if people rely on those stories to make decisions and the decisions go wrong, that storyteller loses credibility
- Telling stories that can stand up to scrutiny over time takes enormous work
- Sometimes your stories will be told with numbers, and people who read the numbers will treat them as quantitative and make predictable decisions based on them. Your challenge is to find a way to evaluate whether the stories those numbers would tell would be consistent with the stories a storyteller would tell.

End of Slide Deck

This isn't the end of the story for software metrics

But it's as far as I've gotten in my thinking. There's a lot of road left to travel.

But I think this takes us pretty far.

Copyright (c) Cem Kaner 2013

4 Things to Remember

1. Yes, most software metrics are (to some degree) invalid. However, that doesn't reduce the need for the information we are trying to get from them.
2. I think it's part of the story of humanity that we've always worked with imperfect tools and always will. We succeed by learning the strengths, weaknesses and risks of our tools, improving them when we can, and mitigating their risks.
3. We need to look for the truths behind our numbers. This involves discovery and cross-validation of patterns across data, across analyses, and over time. The process is qualitative.
4. Qualitative analysis is more detailed and requires a greater diversity of skills than quantitative. Qualitative analysis is not a free (or even a cheap) lunch. We evaluate the quality of quantitative measures by critically considering their ***validity***. An equally demanding evaluation for qualitative measures considers their ***credibility***.

Readings

- Austin, R.D. (1996). Measurement and Management of Performance in Organizations. Dorset House. *Alternatively, see* Austin, R.D.(1994). Theories of Measurement and Dysfunction in Organizations. Ph.D. Dissertation, Department of Social and Decision Sciences, Carnegie Mellon University. *The dissertation is the original source material for the book.*
- Campbell, D.T. (1976). Assessing the Impact of Planned Social Change. Public Affairs Center, Dartmouth College. http://portals.wi.wur.nl/files/docs/ppme/AssessingImpact_Campbell.pdf
- Chambers, R.J. (1960). Measurement and misrepresentation. Management Science, Volume 6(2), 141-148.
- Deming, W.E. (1982). Out of the Crisis. MIT Press.
- Hoffman, D. (2000). The darker side of metrics. Pacific Northwest Software Quality Conference. http://www.testingeducation.org/BBST/foundations/Hoffman_DarkerSideMetrics.pdf
- Johnson, M. (1996) Effective and Appropriate Use of Controlled Experimentation in Software Development Research, Master's Thesis (Computer Science), Portland State University. http://www.worldcat.org/title/effective-and-appropriate-use-of-controlled-experimentation-in-software-development-research/oclc/36599919
- Kapner, S. (March 4, 2013). Citi's CEO is keeping score. Wall Street Journal, http://online.wsj.com/article/SB10001424127887324539404578340413764265022.html
- Lincoln, Y.S. & Guba, E.G. (1985) Naturalistic Inquiry. Sage.
- Guba, E.G. & Lincoln, Y.S. (1989) Fourth Generation Evaluation. Sage.
- Kaner , C. & Bond, W.P. (2004), “Software engineering metrics: What do they measure and how do we know?” http://www.kaner.com/pdfs/metrics2004.pdf
- Kaner, C. & Kabbani, N. (2012) Software Metrics: Threats to Validity. Video at http://testingeducation.org/BBST/metrics/CAST2012Metrics.mp4. Slides at http://testingeducation.org/BBST/metrics/MetricsValidityLecture2012.pdf
- Nichols, S.L. & Berliner, D.C. (2005). The Inevitable Corruption of Indicators and Educators Through High-States Testing. Education Policy Research Unit, Arizona State University. http://portals.wi.wur.nl/files/docs/ppme/AssessingImpact_Campbell.pdf
- Patton, M.Q. (2002, 3rd Ed.). Qualitative Research & Evaluation Methods.
- Ridgway, V.F. (1956). Dysfunctional consequences of performance measurements. Administrative Science Quarterly 1(2): 240-247. http://www.jstor.org/stable/2390989
- Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002) Experimental and Quasi-Experimental Designs for Generalized Causal Inference.
- Suddaby, R. (2010) Construct clarity in theories of management and organization, Academy of Management Review. 35(3) 346-357. http://www.aom.pace.edu/amr/editorials/Suddaby.Editor%20Comments.Construct.pdf
- Trochim, W. & Donnelly, J.P. (2006, 3rd Ed.) The Research Methods Knowledge Base