MEASUREMENTS, BIASES, JUDGMENTS: UNDERSTANDING VARIATIONS FOR RELIABLE ESTIMATES

SAND2011-7709C

11 DE NOVEMBRO DE 2011

2ª CONFERÊNCIA BRASILEIRA DE MEDIÇÃO E ANÁLISE DE SOFTWARE

SÃO PAULO, BRAZIL

Joe Schofield
joescho@joejr.com
SANDIA NATIONAL LABORATORIES

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

Recent Numbers from Capers Jones favorably ranks Brazil among 65 countries

APPROXIMATE INTERNATIONAL SOFTWARE QUALITY LEVELS EXPRESSED IN IFPUG FUNCTION POINTS

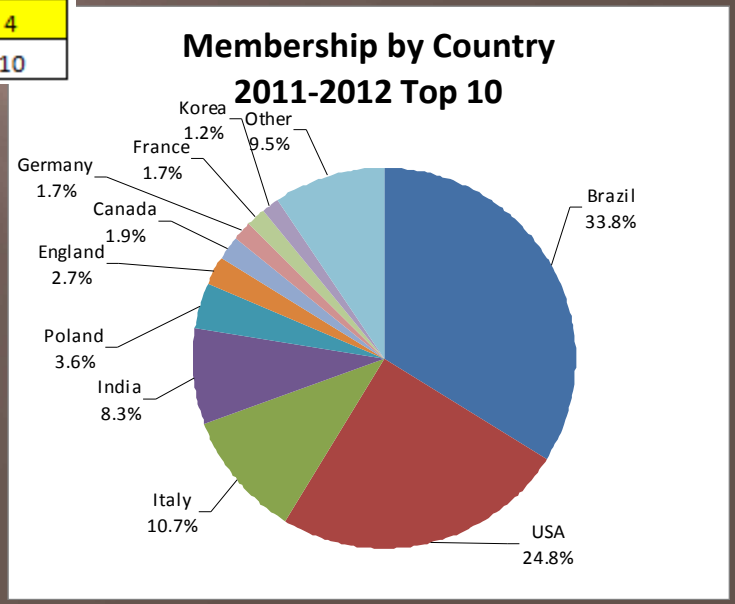
Version 2.0 6/12/2011

Note 1: Preliminary data from small samples.
Note 2: This table was created to illustrate the need for more extensive international studies.
Note 3: Defect potentials include requirements, design, code, document and bad fix defects.
Note 4: Defect removal efficiency includes results of inspections, static analysis, and testing.
Note 5: Delivered defects are based on 90 days of customer usage of software applications.
Note 6: To average > 80% in defect removal efficiency inspections and static analysis are needed
Note 7: Defect potential range is from < 2.5 defects per function point to > 7.0 defects per function
Note 8: Defect removal efficincy range is from < 75% to > 99.75%.
Note 9: Delivered defect range is from .00625 to 1.75 per function point
Note 10: IFPUG stands for International Function Point Users Group
Note 11: Software population includes analysts, software engineers, testers, quality assurance, m

		Average Defect Potentials in 2011	Average Defect Removal Efficiency	Average Delivered Defects in 2011	National Population from CIA Fact Book in 2011	Estimated Software Population in 2011
1	Japan	4.50	90.00%	0.45	126,475,664	986,510
2	India	4.90	89.00%	0.54	1,189,172,206	2,140,510
3	South Korea	4.90	87.00%	0.64	48,754,767	282,778
4	Canada	4.75	86.50%	0.64	34,030,589	231,408
5	Netherlands	4.80	86.50%	0.65	16,847,007	131,407
6	Sweden	4.75	86.00%	0.67	9,088,728	61,803
7	Norway	4.75	85.50%	0.69	4,491,849	35,036
8	Switzerland	5.00	86.00%	0.70	7,639,961	59,592
9	Australia	4.85	85.50%	0.70	21,766,771	148,014
10	Ireland	4.85	85.50%	0.70	4,670,996	22,421
11	Israel	5.10	85.50%	0.74	7,473,052	50,817
12	United Kingdom	4.95	85.00%	0.74	62,698,362	489,047
13	Denmark	4.80	84.50%	0.74	5,529,888	32,073
14	United States	5.00	85.00%	0.75	313,232,044	2,443,210
15	France	4.85	84.50%	0.75	65,312,249	509,436
16	Finland	4.70	84.00%	0.75	5,259,250	35,763
17	Austria	4.75	84.00%	0.76	8,217,280	64,095
18	Belgium	4.70	83.50%	0.78	10,431,477	81,366
19	Mexico	4.85	84.00%	0.78	113,724,226	432,152
20	New Zealand.	4.85	84.00%	0.78	4,290,347	24,884
21	Germany	4.95	84.25%	0.78	81,471,824	635,480
22	Spain	4.90	84.00%	0.78	46,754,784	224,423
23	Portugal	4.85	83.50%	0.80	10,760,305	51,649
24	Iran	5.05	84.00%	0.81	77,891,220	295,987
25	Hong Kong	4.85	83.00%	0.82	7,122,508	34,188
26	Iceland	4.75	82.50%	0.83	311,058	1,493
27	South Africa	4.90	83.00%	0.83	49,004,031	235,219
28	Taiwan	4.90	83.00%	0.83	23,071,779	87,673
29	Brazil	4.97	83.00%	0.84	203,429,773	976,463
30	Jordan	5.00	83.00%	0.85	6,508,271	24,731
31	Bahrain	4.75	82.00%	0.86	1,214,705	2,186
32	China	5.20	83.50%	0.86	1,336,718,015	2,406,092
AVERAGE		4.87	83.65%	0.80	5,041,253,245 Total	17,444,711 Total

Brazil is now the world leader in Function Point membership & adoption

Country	Total membership	WWC	RC	ARC	RI	AFF	UNIV	STU
Poland	15				15			
Canada	8		6		2			
France	7	1	2		3			1
England	11		3	1	7			
Italy	44		16		27			1
India	34	3	5		26			
Thailand	1							1
USA	102	6	17	1	77			1
Brazil	139	3	42	4	84		2	4
Total	411	13	107	6	272	0	3	10



These numbers extracted from the IFPUG monthly report on membership, September, 2011.

(Conduct) Survey Here!

A Survey Related to Decision-Making[1]

1. How many countries have at least one McDonald's?
2. What is the range of a Minuteman (III) Missile?
3. How long (minutes & seconds) was the song "Stop in the Name of Love" recorded by the Supremes?
4. If the air temperature (F) is 5 degrees below zero and the wind speed is 15 mph, what would be the wind chill?
5. How many sovereign rulers has England had in the last 1000 years?
6. What is the average cost of testing in software development relative to total cost?
7. How many meters high is the Sears Tower?
8. The Airbus A380 has 525 seats when configured for three classes. How many seats would it hold if all the seats were economy class?
9. How many inches does the hair on a human head grow in a year?
10. On average, a software development project projected to take 17 months actually takes how long?

[1]These questions were posed to ISMA Cinco! attendees by Dr. Ricardo Valerdi, MIT, as part of his keynote presentation. Used with permission 8-16-2011.

Response Range Considerations

Q#	General Knowledge	Bounds
1	Been in or at or seen on TV	Less than number of countries
2	Been in AF or a defense "expert"	Less than ½ the circumference of the earth
3	Radio listener in or since 1960s	Less than 4 minutes
4	Live in a cold climate	Likely not a positive number
5	Knowledgeable about English history (monarchs rule longer than presidents)	> 1, < 100
6	Software engineer or project manager	> 1, < 100
7	Been to Chicago; an architect	> 1, < ~1000
8	Aviation buff; a pilot	> 525, < ~1000
9	Have hair; cut hair	> 3, < 10
10	Software engineer or project manager	> 17, < 51

The Questions, Results, and Observations . . .

	Questions	R1R> R2R	R2R > R1R	Tied
1	How many countries have at least one McDonald's?	31	20	4
2	What is the range of a Minuteman III Missile?	20	28	7
3	How long (minutes & seconds) was the song "Stop in the Name of Love" recorded by the Supremes?	24	22	9
4	If the air temperature (F) is 5 degrees below zero and the wind speed is 15 mph, what would be the wind chill?	25	20	10
5	How many sovereign rulers has England had in the last 1000 years?	31	15	9
6	What is the average cost of testing in software development relative to total cost?	16	26	13
7	How many meters high is the Sears Tower?	32	18	5
8	The Airbus A380 has 525 seats when configured for three classes. How many seats would it hold if all the seats were economy class?	21	26	8
9	How many inches does the hair on a human head grow in a year?	26	16	13
10	On average, a software development project projected to take 17 months actually takes how long?	19	30	6

Thanks to Jacqueline Dominguez who contributed to the development of the formats and underlying formulae used above.

R1R = Round 1 responses
R2R = Round 2 responses

1. (R1R) Unsatisfied with allowing respondents to cavalierly answer questions with a wide, but *meaningless* range, I attempted to alter the "experiment" by offering a prize for *closeness*. One might think of this as the "horse shoe game" twist—closer matters.
2. (R2R) On four questions (2, 6, 8, 10) when respondents were placed under pressure to be "better than their peers" their answers actually got worse. (Daniel Pink talks about this in his book *Drive* as it relates to cognitive tasks.)
3. More alarming was that two of the four questions (6, 10) in which answers got worse, the questions were related to software project management.
4. The answers to those two questions ONLY are refutable; that is, there are similar studies in which answers would be distinguishably different.
5. Questions 2 & 8 are the other two questions in which response grew worse. Since these questions were posed (by me) to folks who work in an area related to #2, this is troubling.

The Questions, Results, and Observations (cont'd) . . .

	Questions	R1R > R2R	R2R > R1R	Tied
1	How many countries have at least one McDonald's?	31	20	4
2	What is the range of a Minuteman III Missile?	20	28	7
3	How long (minutes & seconds) was the song "Stop in the Name of Love" recorded by the Supremes?	24	22	9
4	If the air temperature (F) is 5 degrees below zero and the wind speed is 15 mph, what would be the wind chill?	25	20	10
5	How many sovereign rulers has England had in the last 1000 years?	31	15	9
6	What is the average cost of testing in software development relative to total cost?	16	26	13
7	How many meters high is the Sears Tower?	32	18	5
8	The Airbus A380 has 525 seats when configured for three classes. How many seats would it hold if all the seats were economy class?	21	26	8
9	How many inches does the hair on a human head grow in a year?	26	16	13
10	On average, a software development project projected to take 17 months actually takes how long?	19	30	6

6. The sum of all variances for R1R was 51,032,348 and for R2R 870,167 – a decrease of 58.6 times the variation of R1R.
7. Excluding one set of (outlier) answers from both question sets, the sum of the variances for R1R was 1,009,619 and for R2R 868,643 or just 1.16 times the variance of R1R.
8. Excluding the next two largest outliers from both question sets, the sum of the variances for R1R was 453,277 and for R2R 864,146. This exclusion may mislead the casual observer into believing that the sum of the variances — when respondents are asked to "compete" — actually increases the variances significantly. This appears to be a distortion of the data and what the data is telling us.
9. Excluding the next largest "outlier" from both questions sets, the sum of the variances for R1R was 288,836 and R2R 251,859.

New questions were developed because:

1. Questions needed answers that were repeatable when searched and researched.
2. Questions needed answers that were current in the literature; that is, less subject to change over time.
3. Increased liberty was desired to analyze the data based on unforeseen inquiry.
4. Questions needed to have less global sensitivity.
5. Questions needed to have less industry specific sensitivity.
6. Questions could be added to trigger desired specific interest in respondents (latent defect estimation as an example below)

Three samples of fish are taken from a lake. 70 fish were found in the largest sample, a total of 90 fish in the other two samples. 50 fish were common to both the sample of 70 and the sample of 90 fish. What's the predicted number of fish in the lake that were not captured in the samples? (((70 * 90) / 50) – (70 + 90 – 50) = 130 – 110 = 20)

Ref: *Beyond Defect Removal: Latent Defect Estimation with Capture Recapture Method;* CrossTalk, August 2007

New questions include:

1. On average, how far is the sun from Neptune when compared to the distance of the sun to Earth? (or distance from Earth to sun)
2. How many tenths of an inch (centimeters) do fingernails grow within a year?
3. What is the flight distance in kilometers from New York City to Mumbai India?
4. How many feet (meters) above sea level is Mt. Kilimanjaro?
5. How long is the song “Hey Jude”, originally recorded by the Beatles in 1968?
6. The Oasis of the Seas is listed as the world’s largest cruise ship (circa 2011). What is the maximum passenger capacity listed for this vessel?
7. If it’s 80 degrees Fahrenheit, what’s the temperature in Celsius?
8. What is the estimated maximum number of military deaths that resulted from WWII expressed in millions?
9. What is the number of gallons (liters) in a US barrel of oil?
10. The gestation period of an elephant is how many months?

Response Range Considerations

Q#	General Knowledge	"Logical" Bounds
1	Astronomy interest	Uncertain
2	Have or cut "nails"	< six inches
3	Traveler, global geography	< ½ the global circumference times about .6?
4	African, African traveler, mountain climber	> 10,000, < ~30,000 ft.
5	Beatles fan, music enthusiast from 60s	> 3, < 10
6	Cruise traveler, trivia expert	> 3000, < 10,000
7	Celsius familiar, meteorologist	> 1, < 80
8	War, history buff	> 5,000,000, < 100,000,000
9	Oil person, savvy consumer	> 30, < 60
10	Elephantologist, veterinarian	> 3, < 20

New questions & results . . .

Q#	Question	R1R > R2R	R2R > R1R	Tied
1	On average, how far is the sun from Neptune when compared to the distance of the sun to Earth?	46%	34%	20%
2	How many tenths of an inch do fingernails grow within a year?	51%	31%	17%
3	What is the flight distance in kilometers from New York City to Mumbai India?	49%	49%	3%
4	How many feet above sea level is Mt. Kilimanjaro?	49%	31%	20%
5	How long in seconds is the song "Hey Jude" originally recorded by the Beatles in 1968?	26%	57%	17%
6	The Oasis of the Seas is listed as the world's largest cruise ship. What is the maximum passenger capacity listed for this vessel?	54%	34%	11%
7	If it's 80 degrees Fahrenheit, what's the temperature in Celsius?	43%	34%	23%
8	What is the estimated maximum number of military deaths that resulted from WWII in millions?	40%	46%	14%
9	What is the number of gallons in a US barrel of oil?	57%	34%	9%
10	The gestation period of an elephant is how many months?	40%	37%	23%

Note that ties are typically from respondents who enter the same answer for both sets of responses.
Virtually at least 1/3 of all "incented" responses were worse than initial responses to questions.

Sources of variation from our:

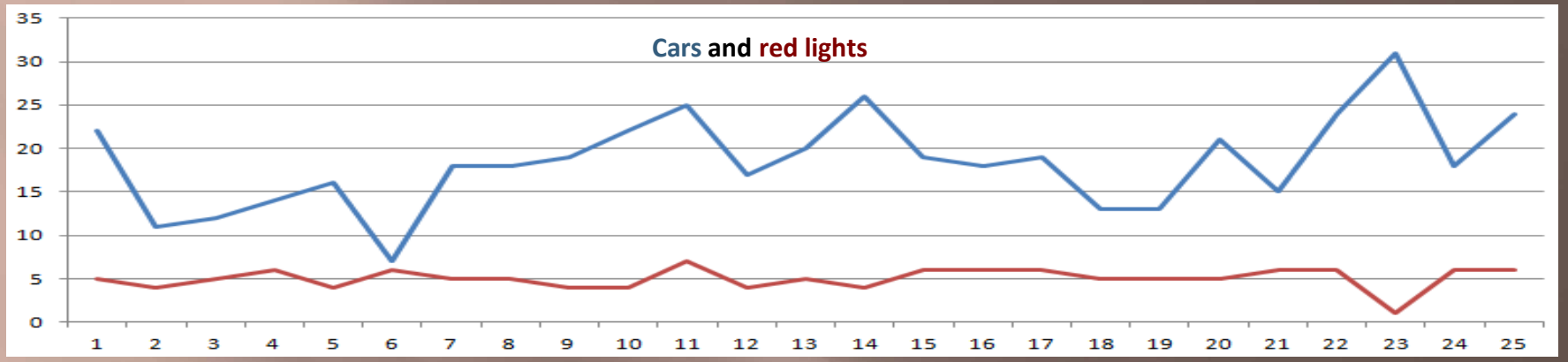
- Measurements
- Measurement presentation
- Thinking
- Judgment

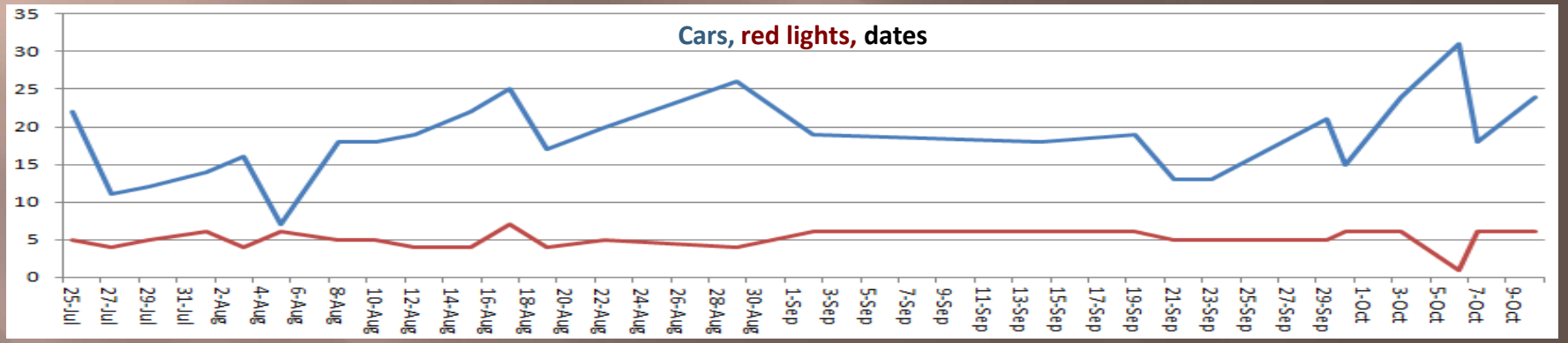
"You may conclude that any proximity associated with an estimate and an actual value is solely to coincidental and serendipitous"

OR

"it's better to be lucky than to be good!"

Selected potential sources of variation in measuring - 1

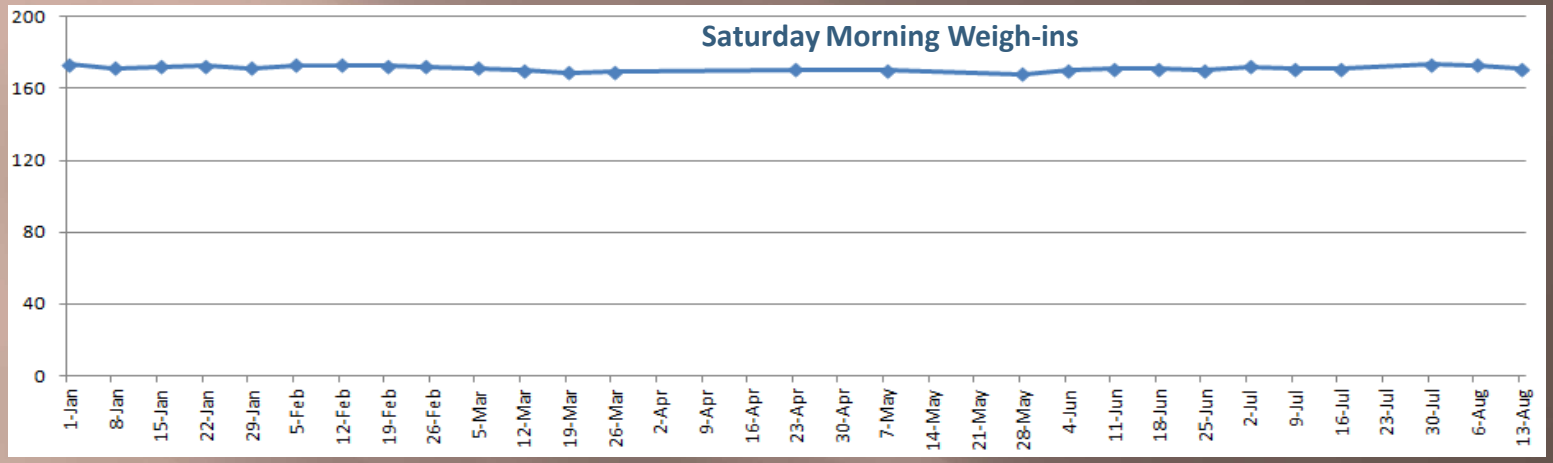


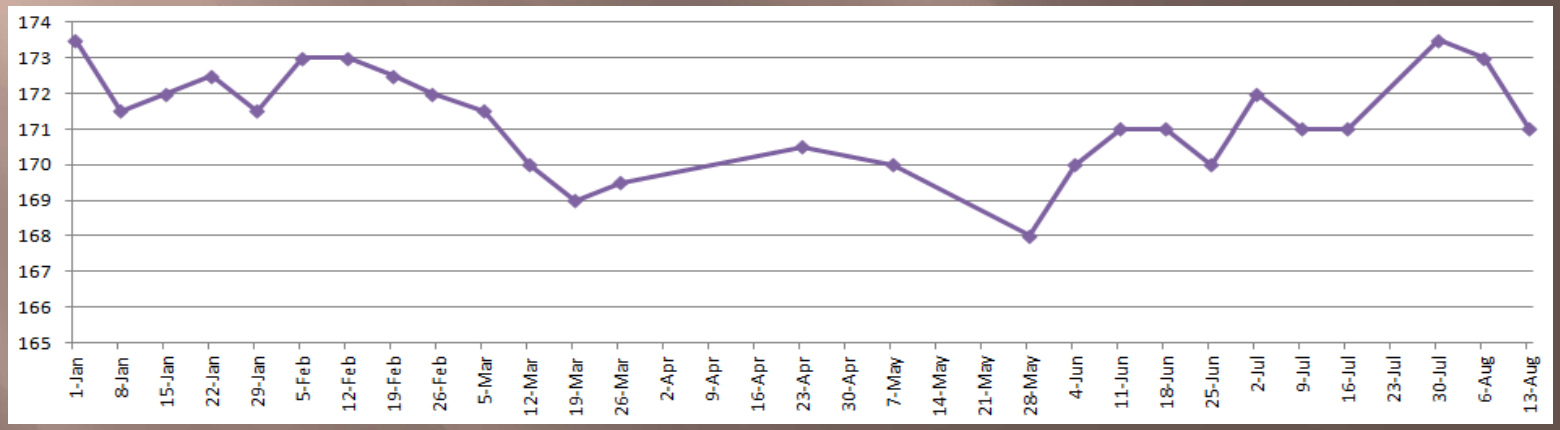


Notes:

- **Attempting to assign a relationship where none exists**
- **Missing a relationship where one exists**

Selected potential sources of variation in measuring - 2





Notes:

- **Hiding data with inappropriate scaling**
- **Overemphasizing expected variation with inappropriate scaling**
- **Unable to *monitor and control* and take *corrective action***

Selected potential sources of variation in measuring - 3

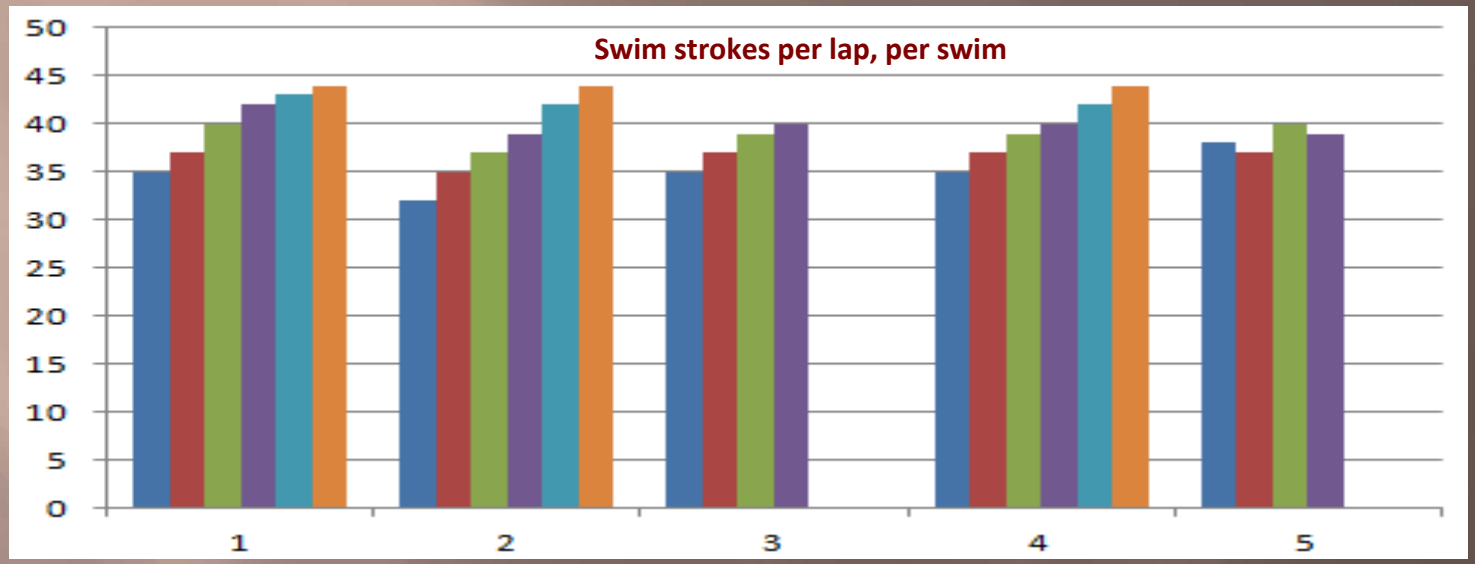
Book List

Beyond the Summit
Change Your Questions Change Your Life
Competing on Analytics
Drive
Emotional Intelligence
How the Mighty Fall
Outliers
Predictably Irrational
Super Freakonomics
The Fourth Turning
The Future of the Internet and how to stop it
The Oz Principle
The Real Business of IT
Tipping Point
Vaults, Mirrors, Masks
What the Dog Saw
Who Moved My Cheese (17)
Change the Culture Change the Game

Notes:

- Don't read novels
- Recognize outliers
- All pages are not created equal – measure the right / same thing

Selected potential sources of variation in measuring - 4



Notes:

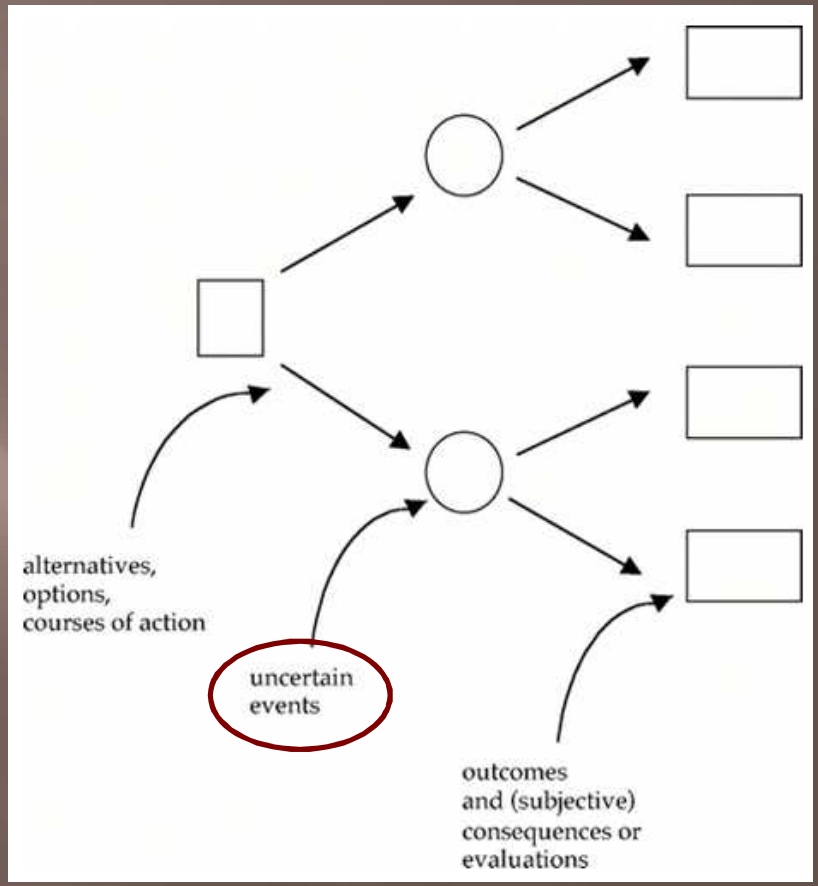
- Clear trend that swimmer uses more strokes per lap as the number of laps increases
- Or is this an example of measuring for the sake of measuring?
- Or, as one works (exerts), does it take increasing levels of energy to produce a similar results?

Sources of variation from our thinking (and estimates) - 1

The role of uncertainty in events

OR,

predicting the future is hard



PROBLEMS FOR JUDGMENT AND DECISION MAKING
Annual Review of Psychology
Vol. 52: 653-683 (Volume publication date February 2001)
DOI: 10.1146/annurev.psych.52.1.653
R. Hastie

Sources of variation from our thinking (and estimates) - 2

Decisions about financial investments, litigation, environmental disasters, and insurance are usually **based on a lack of knowledge** about relevant probabilities.

. . .

Lack of calibration based on comparisons of confidence judgments against percentages of correct items have led researchers to argue that **people are often overconfident**.

JUDGMENT AND DECISION MAKING
Annual Review of Psychology
Vol. 49: 447-477 (Volume publication date February 1998)
DOI: 10.1146/annurev.psych.49.1.447
B. A. Mellers[1], A. Schwartz[2], and A. D. J. Cooke[3]

Sources of variation from our thinking (and estimates) - 3

When people make interventions to a system they expect the effects to be nearly instantaneous. Unfortunately, in most of the cases the intervention intended to improve the process actually causes outcomes to get worse before they get better, if they get better at all.

Underestimation in the "When It Gets Worse Before it Gets Better" Phenomenon in Process Improvement
Advanced Concurrent Engineering, 2011, Part 1, 3-10,
DOI: 10.1007/978-0-85729-799-0_1
Ricardo Valerdi and Braulio Fernandes

Sources of variation from our thinking (and estimates) - 4

Gambling is a widespread form of entertainment that may afford unique insights into the interaction between cognition and emotion in human decision-making. . . . The cognitive approach has identified a number of erroneous beliefs held by gamblers, which cause them to over-estimate their chances of winning.

Decision-making during gambling: an integration of cognitive and psychobiological approaches
Luke Clark

Sources of variation from our thinking (and estimates) - 5

Research has shown that the confidence individuals express in their judgments generally exceeds the accuracy of those judgments on difficult tasks (Fischhoff, Slavic, & Lichtenstein, 1977; Lichtenstein & Fischhoff, 1977, 1980; Lichtenstein, Fischhoff, & Phillips, 1982). **(see also, next slide)**

ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
48, 100-130 (1991)
Influences on the Appropriateness of Confidence in Judgment: Practice, Effort, Information, and Decision-Making
PAUL W.PAESE
University of Missouri-St. Louis
JANET A. SNIEZEK

Sources of variation from our judgment (and estimates)

An optimist is a person who is unrealistic about the short term but realistic about the long term, while a pessimist is a person that is realistic on both their short term and long term thinking. **(Seligman, 2006)**

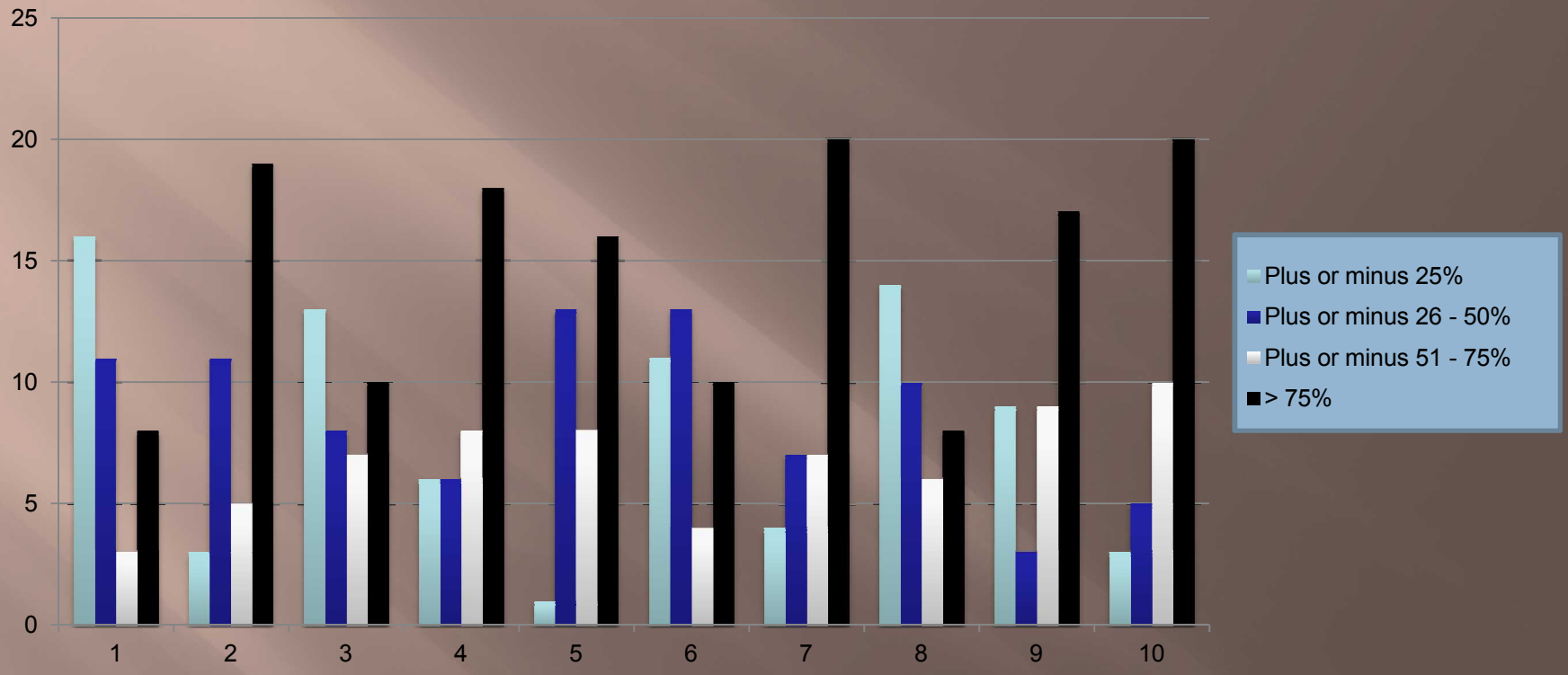
Optimism and pessimism can be quantified using a Brier score

(tested system engineers should) have been in the 90 percent confidence category; most were actually in the 30 percent (mode) category

"how can system engineers be more like bookies and meteorologists and less like software project bidders in their judgments of probabilities?"

The Human Element of Decision Making in Systems Engineers: A Focus on Optimism; Ricardo Valerdi & Craig Blackburn
Massachusetts Institute of Technology – Lean Advancement Initiative

Survey responses within ranges . . .



Note:

- **The values represented as greater than 75 percent can exceed 100 percent; a number of them exceed 10,000 percent, a small number exceed 60,000 percent**
- **Before concluding that the excessive percentages outliers are irrelevant, review the values for the Taurus project**

Project and product performance values:
> 75 percent industry variation

	% Variation
UK's National Health Care Service, budgeted at $12B, closer to $24B in 2007[1]	100
London Stock Exchange's Taurus project: estimated 6 million pounds, actual 800 million[1]	13,200

Note: The *mode* of the responses to 6 (of the 10 questions) fell in the range > 75 percent; questions 2, 4, 5, 7, 9, and 10 (fingernails, Kilimanjaro, "Hey Jude", Fahrenheit to Celsius, gallons to barrel, gestation)

[1]MIS Quarterly Executive Vol. 6 No. 2 / June 2007; University of Minnesota

Project and product performance values:
50 – 75 percent industry variation

	% Variation
70% of large IT programs don't reach their goals in the allotted time and budget[1]	70
Rework consumes almost 50 percent of resources for lower quality large software projects[2]	50

Note:

- **While about 10 percent of responses fell in this range, the range was not the statistical *mode* for any question**
- **the largest 500 U.S. companies lose in excess of $14 billion a year because of failed technology projects[1]**

[1]Taming Information Technology Risk A New Framework for Boards of Directors, National Association of Corporate Directors, 2011

[2]*Software Project Failure Costs Billions. Better Estimation & Planning Can Help*; Dan Galorath on Estimating:, June 7, 2008

Project and product performance values:
0 – 50 percent industry variation

	% Variation
Formal inspections find twice as many defects at 1/5 the cost of testing. [1]	20
Defect detection and removal account for at least 40 percent of total software costs, exceeding the cost of its development. [1]	40
Thirty percent of project effort can be traced to rework[2]	30
Rework consumes almost 50 percent of resources for lower quality large software projects[3]	50
Reworking defective requirements, design, and code typically consumes 40 to 50 percent or more of the total cost of most software projects and is the single largest cost driver.[4]	40 – 50

Note:

- **3 questions' responses *mode* fell in the range of 0 – 25 percent (astronomical units, flight kilometers, WWII deaths)**

[1]MIS Quarterly Executive Vol. 6 No. 2 / June 2007; University of Minnesota

[2]*Dr. Dobb's Report*; informationweek; July 12, 2010; study by Dean Lefingwell, 1997

[3]*Software Project Failure Costs Billions. Better Estimation & Planning Can Help*; Dan Galorath on Estimating:, June 7, 2008

[4]Jones, Capers. *Estimating Software Costs*, New York: McGraw-Hill, 1998

Are these comparisons reasonable?

- Both survey and project results were measured, not confabulated
- Both sets of results are drawn from populations of software engineers and project managers
- The comparison of results suggests that the differences aren't that significant; that is, they are well within the range of credibility
- Comparisons can trigger introspection
  - What are our numbers?
  - How do we compare to the broader population?
  - How do we use these numbers to improve in targeted areas?
- This is my data from which I've drawn insights and value. Where's the data to support your assertions?

Takeaways -1

- Surveys evidence that our confidence often exceeds our ability to predict results
- Respondents under pressure are likely to produce worse results
- How we define a measure, what we measure, and how we represent a measure all introduce biases and variation in our judgments
- The way that we think clearly adds a dimension of unpredictability to estimates
- Our interpretation of measures via judgments likely increases variation
- Less than desirable results in performance are not much different (by percentage) than answers to questions with which we have very little confidence

The elusiveness of reliable measures *increases* the significance of refining our measurement processes; it does not excuse it . . .

Takeaways - 2

"*The quality of our decision-making is deteriorating across the board. Not because the people in charge are stupid. But because they're all running too fast, making too many decisions, too fast, about too many things they know too little about.*" - Alvin Toffler, futurist

Preliminary conclusion: The *responses* to a series of general knowledge questions by IT professionals displays wide ranges of variation that parallel *results* for software projects, perhaps those with which they are familiar.

- **Eagerness to do well may result in worse estimation results (not including the influence of optimism)**
- **Cost & schedule are easier to track, so we do (PMPs, Earned Value, actual vs. estimated, Scorecards, budget reports, budget analysts, forecasts, schedulers)**
- **Limited use of or availability of objective, relevant, and quantitative "actual" data during estimating**
- **Changing and evolving technology becomes an excuse for under performance because we don't manage and understand the changes well**
- **The best data for estimating is your (you, your organization's) most recent *relevant* (proximity of technology, methods, people, and business) data**
- **The best *test* of measurement effectiveness is its usage in decision-making**

Thank you!

(easy) Questions please . . .

Thanks to . . .
Conference planners for inviting me to spend time here with you
RT for her review and constructive comments