Software Development Effort Estimation: Why it fails and how to improve it

Auckland, March 2010

Magne Jørgensen
Simula Research Laboratory & University of Oslo

UNIVERSITY OF OSLO

About me

- Scientific researcher at Simula Research Laboratory, Oslo, Norway
- prof. at Univ. of Oslo
- Industrial experience as programmer, project manager, process improvement manager and general manager.
- Responsible for estimation work and training in several companies.
- Conduct advisory work and seminars for software companies.
- Research reports can (free of charge) be downloaded from: *simula.no/research/engineering/projects/best*

BASIC EFFORT ESTIMATION KNOWLEDGE

Poor estimation work is an important cause of IT-project failure

- A recent (2007) survey of more than 1,000 IT-professionals reports that two out of the three most important causes of IT-project failure were related to poor resource estimation.
    - The third cause was related to poor communication.
- See: *certification.comptia.org/project*
    - *www.informationweek.com/news/management/showArticle.jhtml?articleID=198000251*

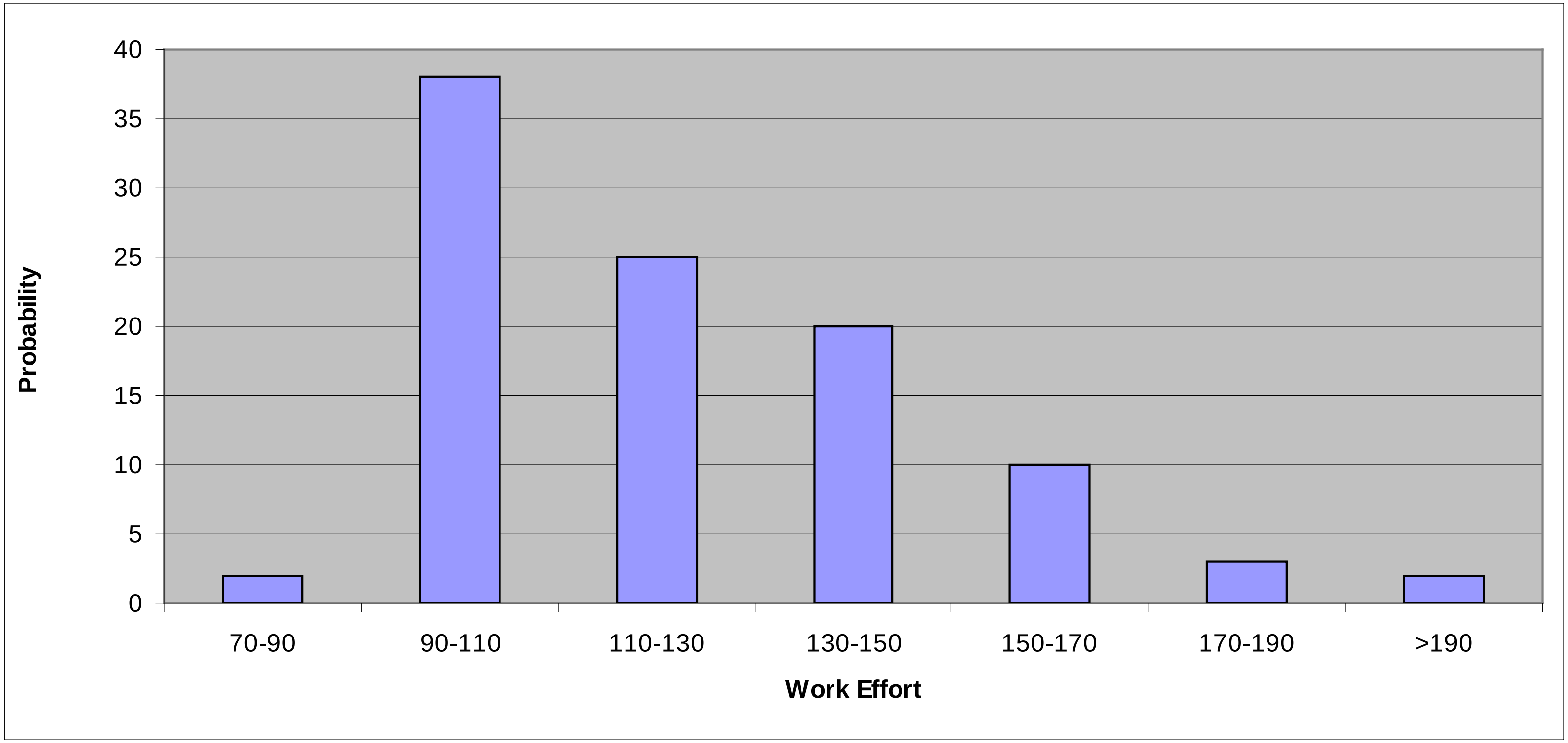
Estimation error

- Average estimation overrun in IT-projects is reported to be about 30%
  - Sometimes the estimation error is much higher.
  - Large estimation errors cause huge project management problems, low profitability, low client satisfaction and poor investment analysis!
  - No substantial changes in average estimation error from 1970 until today. Why are we unable to learn from previous experience?
- But first: What is the meaning of ”estimate”?

Do we know what we mean by ”estimate”?

An effort estimate is sometimes the: i) most likely effort (mode), ii) 50% estimate (median), iii) most optimistic effort, iv) ideal effort, v) 70% estimate, vi) planned effort, vii) budgeted effort, viii) priced effort, ix) effort used as input to the bid and sometimes not even defined.

We have to think probabilistically about effort usage to enable good communication about what we mean by an effort estimate!



Recommendation: Use X% estimates

- Always communicate the type of estimate that you are providing (or receiving)
  - 50% estimate = just as likely to observe over- and under-run
  - 80% estimate = most likely effort + a risk buffer that makes it unlikely (only 20% likely) that there will be overruns. Could for example be the budget or the basis for the price to client.
  - 30% estimate = a close to best case estimate of the effort. Could for example be the bid in a situation where there are long term benefits of a client relationship.
- A method for the assessment of the likelihoods, (e.g., “80% likely not to exceed”) is presented later.

Recommendations

- Use a precise, probability-based terminology to communicate what you mean by an effort estimate.
- Use different terms and processes for different purposes:
  - Estimated effort (pX estimates). Purpose: **Realism**, and just that!
  - Planned use of effort (e.g., based on a 70%-estimate). Purpose: **Project control**.
  - Budget (e.g., based on an 80%-estimate). Purpose: **Financial control** of project portfolio.
  - Price (e.g., based on 40%-estimate). Purpose: **Profitability** on short or long term.
- Different purposes should lead to different processes. Mixing realism (e.g., when estimating effort) and market considerations (e.g., winning a bidding round) means that realism will suffer!

Reasons for Estimation Error
(and how to improve the processes)

UNIVERSITY
OF OSLO

The better-than-average effect….

Over-confidence …

Motivation

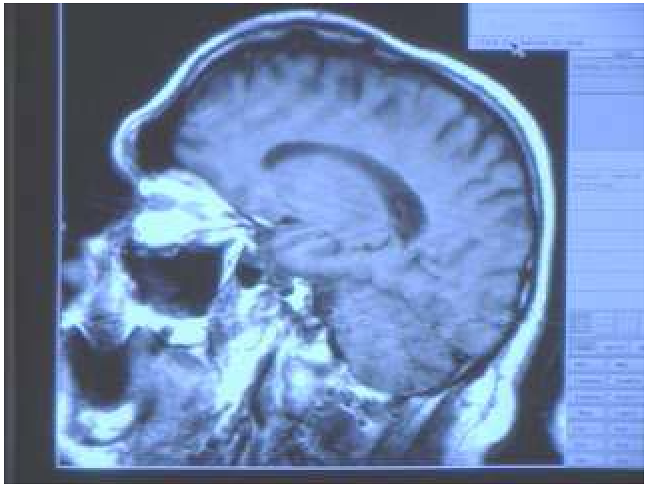
- Mix of “I hope this does not take more than …” and “This will not take more than …”
- Optimism can have a positive impact on performance, BUT
  - Only for a short period of time.
  - The effect is over-rated.

Motivation (cognitive dissonance)

- A over-confident self-evaluation may be beneficial
  - For yourself
  - Because it’s used an performance indicator by others
- Low effort estimates = high performance = better (but less realistic) self-evaluation.
  - Otherwise, we have a cognitive dissonance, i.e., a difference between what we estimate and who we want to be.

Cognitive processes

- Planning (scenarios of the future) makes us more optimistic than looking back (use of historical data).
- Illusion of control sometimes very strong
  - Perhaps the most important reason for over-optimism?
  - More risk analysis can increase the illusion of control!!!!



Bidding round format frequently leads to over-optimism

- The winner’s curse
  - You only win bidding round when being over-optimistic.
  - Many bidders should lead to lower bids to avoid the winner’s curse.
- Bidding anchors
  - Budget
  - Early price indications
  - Expectations

Recommendations to reduce over-optimism

1. Educate a ”cost engineer” that will be evaluated wrt realism of estimates and not him/herself be a part of the projects estimated.
2. Use separate processes (and people?) for estimation, planning and bidding.
3. Avoid irrelevant information (prepare information material before given to the estimators)
4. Use historical data
5. Ask for estimation justification based on historical data. Require very good arguments if the estimates are based on assumption of much less effort compared to similar projects.
6. Do not assume that you have learned very much from previous projects.
7. When there are no relevant historical data available, try to find experts with relevant experience and historical data outside the organizations.
8. Do not let the most skilled estimators estimate the effort of junior developers. Use instead medium skilled developers.
9. If a person benefits from low effort estimates (really wants to start the project etc.), find another person to estimate the effort.
10. Combine estimates from different sources. Use a Delphi-like process (e.g., Planning Poker) to combine these estimates.

When Should We Trust Expert Judgment in Software Development?

UNIVERSITY OF OSLO

Who are the Experts?

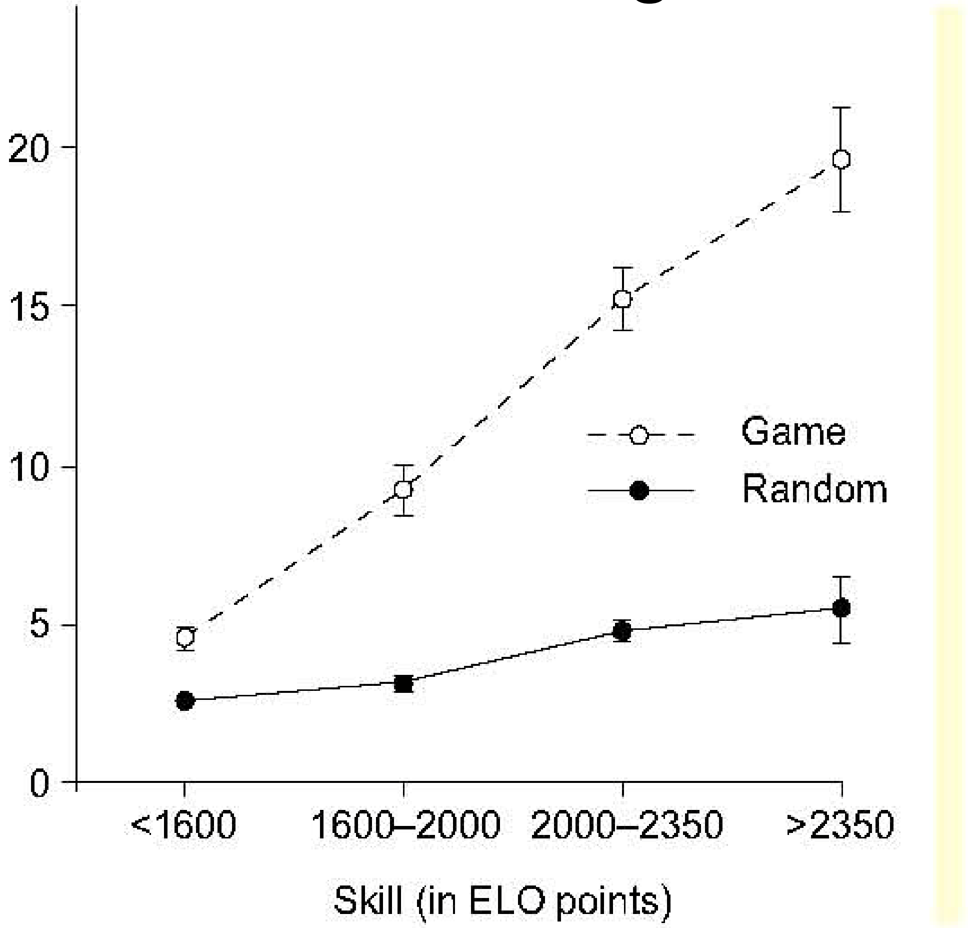
- Those with long experience?
- Those with accurate judgments?
- Those with high confidence in their judgment?
- Those with the best skill, knowledge and/or process?
- This with highest CWS-index? (CWS Cochran-Weiss-Shanteau)
  - *CWS -index = discrimination / consistency*
- Those recognized as experts by at least one other person? (or people away from home, such as me ….?)
- U.S. Supreme Court classifies legal experts in Federal Rule of Evidence 702 as:
  - *”individuals with scientific, technical, skill, experience, training, or education that will assist the trier of fact* [judgment of facts] *to understand the evidence or to determine a fact at issue.”*

What is the Difference Between Experts and Non-Experts in Chess?

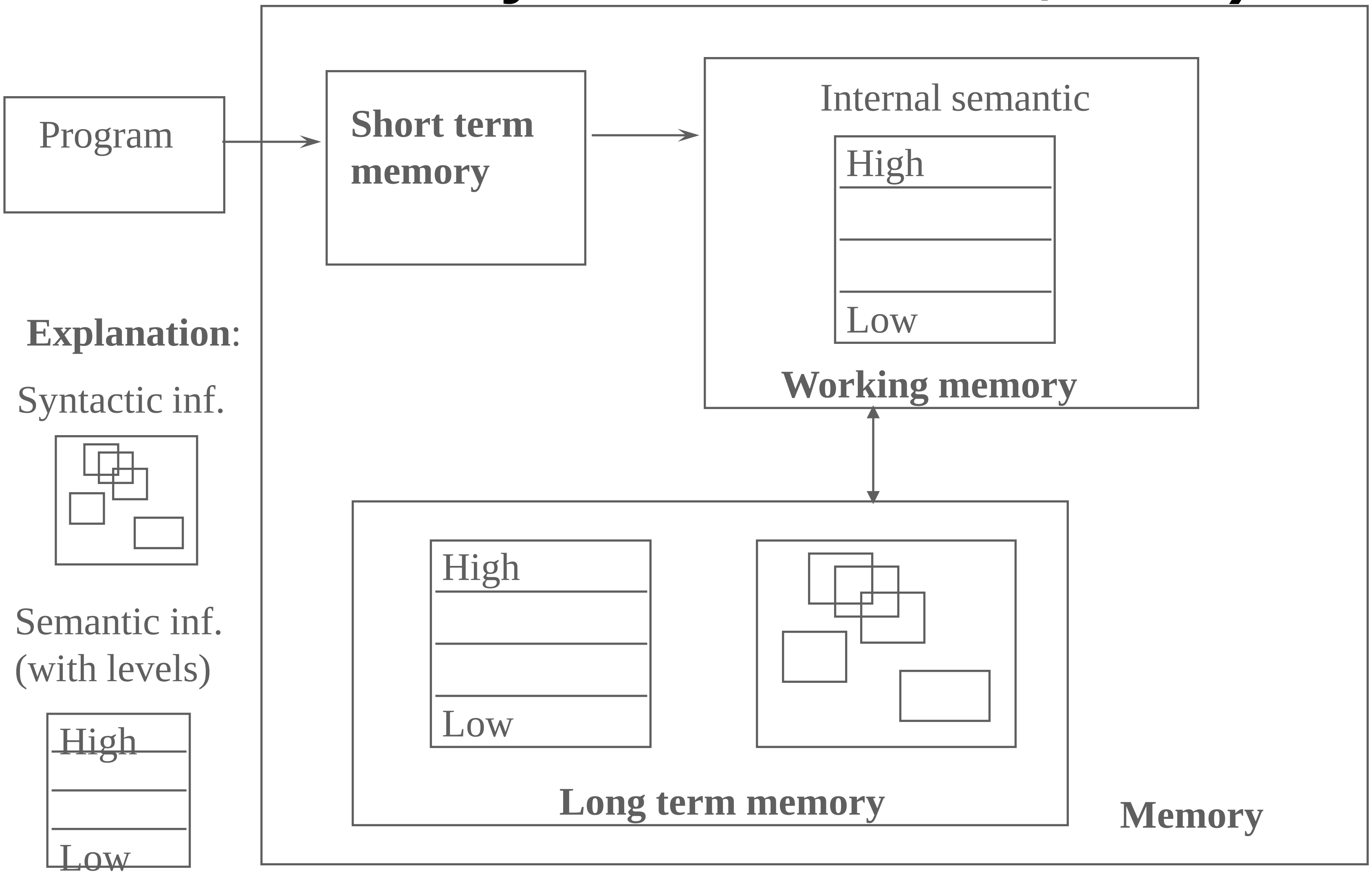
Is an expert better than a non-expert (advanced player) with respect to:

- number of moves analyzed per minute?
- depth of IF-THEN analysis?
- short term memory?
- search heuristic?
- filtering of bad moves?
- recall of randomly positioned chess pieces?
- better working memory capacity?
- ability to analyze larger units, e.g., analyze patterns rather than single pieces?

Chunking mechanisms in human learning, Gobet et al., Trends in cognitive science, 2001



What Separates an Expert and a Novice in Program Comprehension? (Chunking-based Model by Schneiderman, 1979)



Program

Short term memory

Internal semantic

High

Low

Working memory

High

Low

Long term memory

Memory

Explanation:

Syntactic inf.

Semantic inf. (with levels)

High

Low

Some Expert Characteristics ...

- Experts excel mainly in their own domain (expertise is narrow)
- Experts has a large knowledge base, e.g., consisting of chunks (more than 10,000?), rules and schemata.
- The experts perceive large meaningful patterns in their domain (e.g. identify chunks stored in their knowledge base)
- Experts see and represent a problem in their own domain at a deeper (more principled) level than novices; novices tend to represent a problem at a superficial level.
- It takes at least 10 years with “deliberate practice” to achieve top performance.
- Experts do not differ from non-expert in basic information-processing power, but mainly in amount of “deliberate practice”.

For an overview, see, for example: *Expertise, models of learning and computer-based tutoring*, by F. Gobet and D. Wood, 1999.

We don’t know much about expert judgment in software development. Why not?

- Lagnado et al. 2006: “*Studies suggest that quite different regions of the brain are involved in learning and insight about learning.”*

- Essential parts of the expert judgment are unconscious/intuition-based. We don’t have easy access to such processes.

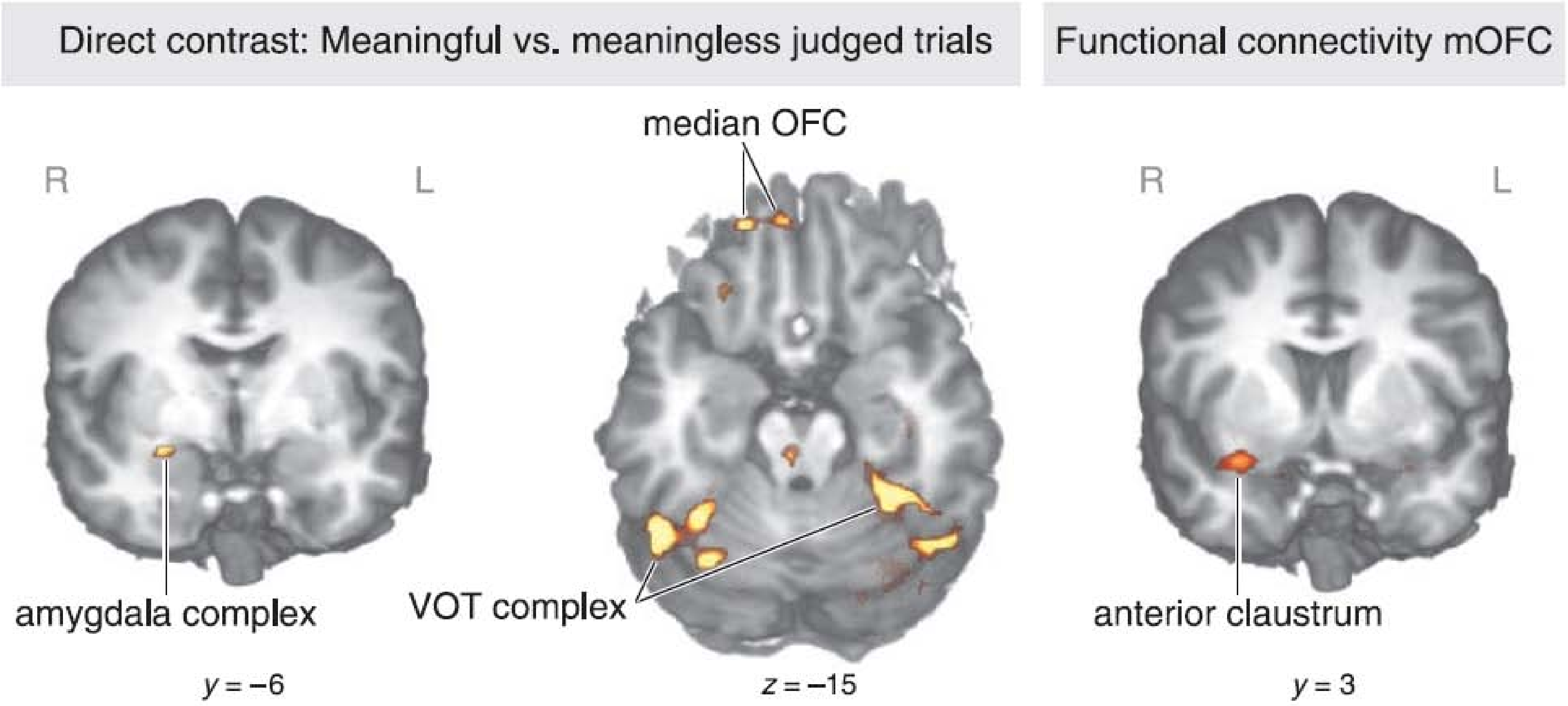
- Lack of knowledge/awareness about the underlying process means that it’s difficult to assess when it is likely to work well and when it will fail.

Example: Judgment-based effort estimation.

- Ask a software professional about his judgment-based estimation process or use a think-aloud protocol to collect this information, and you will NOT get much valuable information.
  - They typically respond with “don’t know”, “it felt right” or present vague statements about their use of experience.
  - The may also feel that they should know how they did the estimation work, and start to rationalize, e.g., by describing how they believe they should have done this as rational beings.
- The same goes, I guess, for expert-judgment based assessment of properties like “maintainability”, “user friendliness” and “quality”.
- It is consequently not possible to gain much insight into these expert judgment-based processes by asking people (think-aloud protocols, interviews, experience reports) or observing their actions. (We have tried and failed several times ...)

The feeling that a judgment is “right” seems to involve brain regions different from those involved in conscious, analytic processes ...

- ”the median OFC, the lateral portion of the amygdala, anterior insula, and ventral occipito-temporal regions ...”
  - *What Neuroscience Can Tell about Intuitive Processes in the Context of Perceptual Discovery*, by Kirsten G. Volz and D. Yves von Cramon, 2006.


Direct contrast: Meaningful vs. meaningless judged trials
Functional connectivity mOFC
median OFC
R
L
R
L
amygdala complex
VOT complex
anterior claustrum
y = −6
z = −15
y = 3

The dual theory of cognition ...

- ”*Both theory and a substantial body of evidence, some of it derived from neuro-imagining studies of the brain employing fMRI technology, support the view that humans employ at least two distinct systems to process information, a rational system and an intuitively-oriented experiential system*” (Goel & Dolan, 2003)
- The “gut feeling” (intuitive) based system is probably the oldest and the one that feels most natural to follow.
- When our “gut feeling” (e.g., judgment-based estimation) says one thing, while your “head” (e.g., an analytic quantification step) says something else, we have a conflict between the two thinking systems.

More on differences between these two systems (Hammond et al, 1987)

Analysis:

- High insight into judgment process, and, hence publicly retraceable
- Low confidence in outcome, high confidence in method
- Slow rate of processing
- High cognitive consistency

Intuition:

- Low insight into judgment process, and, hence difficult to retrace and defend
- High confidence in outcome, low confidence in method
- Fast rate of processing
- Low cognitive consistency

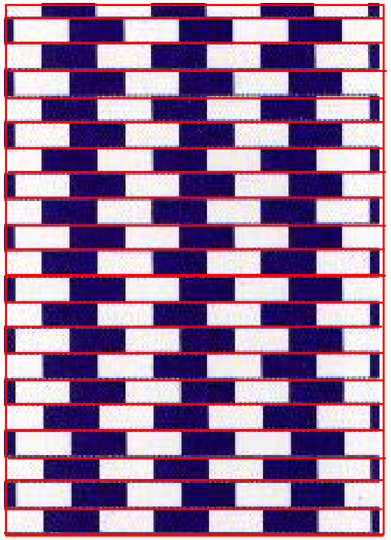
A minor distraction: Do women base their judgments more on intuition than men?

- NO. Only small differences in use of intuition (unconscious processes) in judgment and decision processes.

- Men, however, seem to have a larger need to explain judgments analytically!
  - Individual Differences in Intuitive-Experiential and Analytical-Rational Thinking Styles, Seymour Epstein and Rosemary Pacini, Journal of Personality and Socia Psychology, 1996, Vol. 71, No. 2, 390-405

- All of us, independent of gender and profession, are strongly dependent on intuition!

Example of conflict: Are the lines parallel?



Experiment: (Denesraj, V, Epstein, S: Conflict between intuitive and rational processing – when people behave against their better judgment)

- From the paper abstract:
  - “When offered an opportunity to win $1 on every "win" trial in which they drew a red jelly bean, subjects frequently elected to draw from a bowl that contained a greater absolute number, but a smaller proportion, of red beans (e.g., 7 in 100) than from a bowl with fewer red beans but better odds (e.g., 1 in 10). **Subjects reported that although they knew [analytically] the probabilities were against them, they felt [intuitively] they had a better chance when there were more red beans**.”
- Even some of those selecting the “right” bowl described that they had to fight against the desire of selecting the non-optimal bowl.

The same conflict (analysis vs. intuition) is present when, for example, estimating effort

- Suppose that we have a simple model, e.g., the rule that a medium complex “user story” takes 8 work-hours.
- Use of that model implies that a task with five medium complex user stories should take about 40 work-hours.
- The estimator, however, feels that 40 work-hours is too high, and, that 30 work-hours should be sufficient. We now have a conflict between analysis and intuition.
- As reported earlier, we tend to have more confidence in the analytical **process**, but at the same time more confidence in the intuition-based **output** (our expert judgment). How is this conflict solved?
  - A strongly analytical person: Trust the model
  - A strongly intuitive person: Trust the intuition
  - Conflict-averse person: Adjust the model input so that it gives the desired output. In the example, this may be achieved through categorization of some of the medium complex user stories as “simple”. This conflict-avoiding adjustment may happen both consciously and unconsciously.

Experts can be very good, BUT ...

- are frequently outperformed by simple models
- can be extremely inconsistent
- may be unable to transfer extensive knowledge into accurate judgment
- are impacted by many irrelevant factors

The dilution effect - Design

- 44 industry participants
- INSTRUCTIONS: Please, give each of the following estimation model evaluation factors a weighting in %.
- FACTORS:
  1. More accurate effort estimates than expert judgment.
  2. Ease of understanding the model.
  3. Ease of using the model.
  4. The model uses only data typically available in the specification work.
  5. The model is flexible.
  6. The model enables minimum-maximum intervals.
  7. Other factors.
- Group A: Presented the factors 1-3 + 7 (other factors),
  Group B: Presented all factors.
- Who do think had the heighest weighting of Factor 1 (accuracy)?

The dilution effect - results

- Results, Factor 1 (accuracy of model):
  - Group A’s assessment of importance: median 38%
  - Group B’s assessment of importance: median 23%
- **When there is much information of low relevance, experts tend to weight the most relevant information too little.**
  - “Holistic thinking” (complex, highly inter-connected view of the world) may be the most correct thinking mode, but may lead to even higher degree of dilution than simple, less correct causal models.

Priming study - design

- We divided 65 software professionals randomly into three groups: Low (22 participants), Control (23 participants), and High (20 participants).
- We gave all participants the same programming task specification but varied the words describing some of the requirements slightly.
- The most notable difference in wording is that we asked the:
  - Low group to complete a “minor extension”
  - Control group to complete an “extension”
  - High group to develop “new functionality.”
- We told all the estimators:
  - “You shouldn’t assess how much the client will spend on this project, but what’s required by development work with normal delivery quality.”

Priming study - results

- The resulting median effort estimates were
  - Low (minor extension): 40 work-hours
  - Control: 50 work-hours
  - High (new functionality): 80 work-hours

So, when should we trust experts?

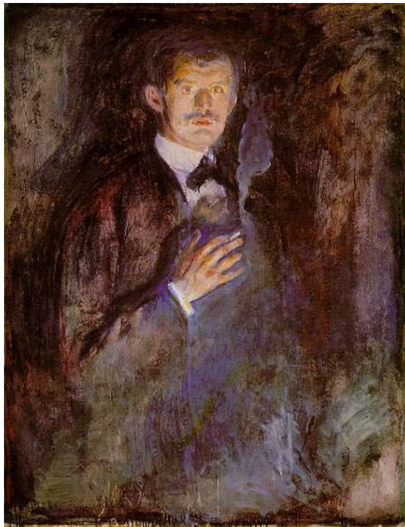
- When they have extensive “deliberate practice” in the particular problem to be solved.
  - See studies by Ericsson et al. and by Shanteau.
- When the context includes little irrelevant and/or misleading information leading to well-known effects (dilution, anchoring, priming, wishful thinking).
  - See the “human biases” studies, e.g., by Kahneman & Tversky

Indicators of estimation expertise

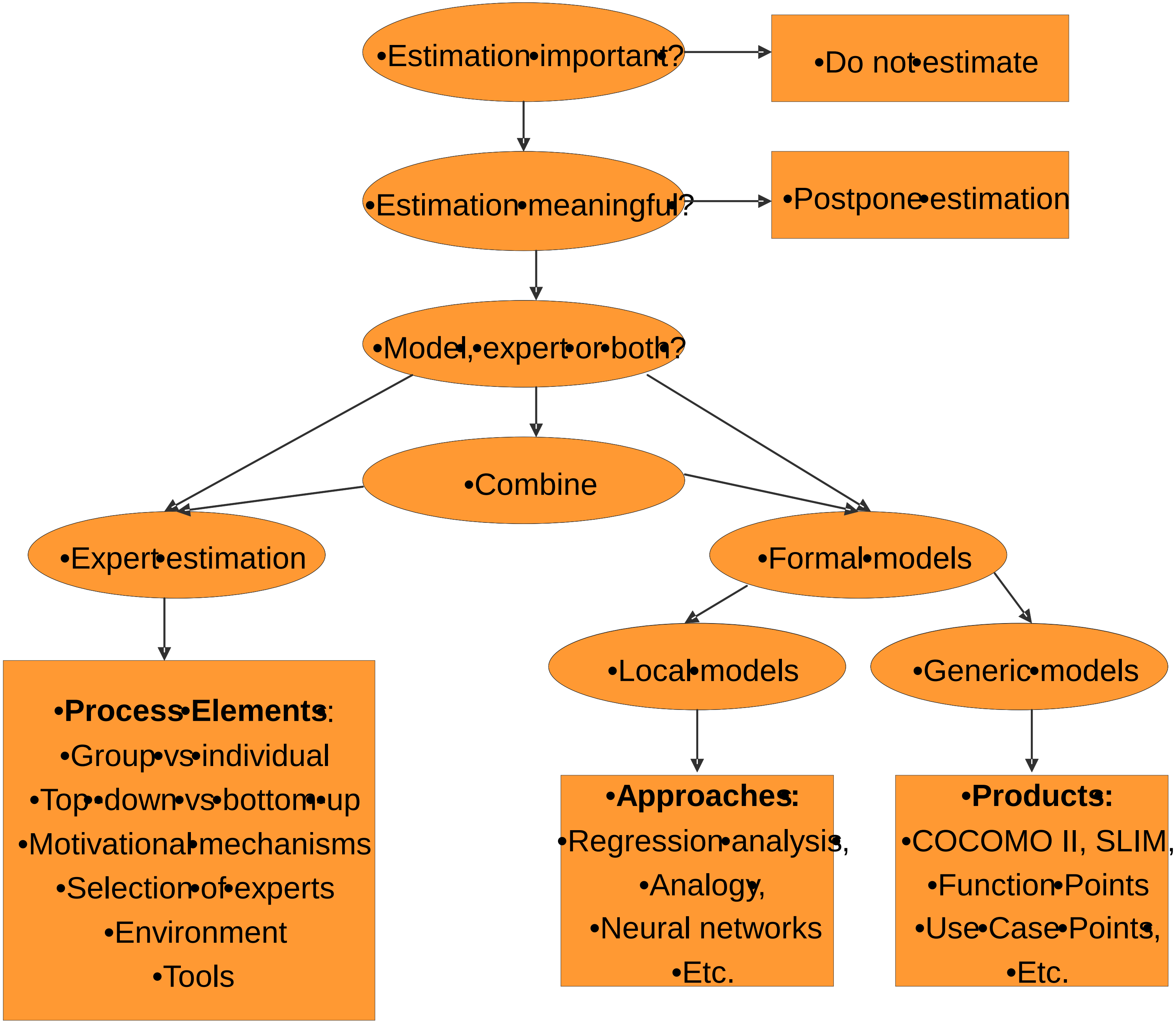
- Length of experience?
  - Not a good indicator.
- Experience from similar projects?
  - Definitively yes, but remember that expertise is "narrower" than typically assumed.
- The best developer?
  - Not always. The best developer may not be suited for the estimation of work effort for novices.
  - "Outside view" (less know-how) sometimes a better strategy.

Indicators of estimation expertise

- The one with highest confidence in his/her estimate?
  - No. We observed the opposite. The most confident are typically the most over-optimistic.
- Those historically most accurate?
  - Yes, but only a medium good indicator. We observed that the software professional (out of two) most over-optimistic on previous estimate had a 70% probability of being the most over-optimistic on the next estimate.
- Personality? (optimism tests, suggestibility test, Big five test, IQ-test, ...)
  - Probably not of much help.
- Slightly depressive people?
  - Yes ☺. They are on average most realistic regarding own abilities.



Selection of Estimation Method



Estimation important?

Do not estimate

Estimation meaningful?

Postpone estimation

Model, expert or both?

Combine

Expert estimation

Formal models

Local models

Generic models

Process Elements:

Group vs individual

Top-down vs bottom-up

Motivational mechanisms

Selection of experts

Environment

Tools

Approaches:

Regression analysis,

Analogy,

Neural networks

Etc.

Products:

COCOMO II, SLIM,

Function Points

Use Case Points,

Etc.

Estimate or not estimate?

- **Essential question**: Do you really need a cost estimate?
- **Rationale**: An estimate, if it is too low or too high, frequently have unwanted impacts on the project behavior, e.g., poor design (too low estimate) and “gold plating” (too high estimate).
- There are several alternatives to estimation that should be considered, such as:
  - Incremental development with the philosophy of do as much as possible within budget, starting with “need to have”-functionality.
  - The client has selected you on the basis of the belief (and previous history in support of this) that the company will work efficiently and with proper quality and says “Just do it!”

Estimate now or later?

- **Essential question**: Is there sufficient knowledge about the requirement and solution to enable meaningful estimation?
- **Rationale**: Early estimates based on insufficient knowledge may easily become over-optimistic and reduce the organization’s ability to derive realistic estimates when more information gets available due to so-called anchoring.
- Alternatively,
  - estimate only the well-understood parts of the project (or the next sprint/release/increment/time-box/…), or,
  - describe the uncertainty through wide minimum-maximum cost intervals, or
  - collect more information

Formal estimation model, expert judgment or both?

- **Essential questions:**
  - Do people in your organization have the necessary statistical and analytical skill to properly use formal estimation models?
  - Is the organization willing to spend effort on implementing, monitoring and, if needed, tailoring the models?
  - Are there important domain knowledge not included in the formal models?
  - Are essential relationships likely to be stable?
  - Is it likely that both formal models and expert estimation provide meaningful estimates?
  - Are there software professionals with experience from similar projects available for estimation?
  - Would you believe in and use the model-based estimate if it diverges substantially from your expert judgment of required effort?

Use of local or generic models

- **Essential questions:**
  - Are there evidence of accurate estimates of the relevant generic model wrt your type of projects and organizational context?
  - Are there necessary statistical and analytical skill to tailor the local model to the relevant types of projects, based on historical data?
- **Rationale:** Most previous studies show that tailoring (local models) is required.

Selection of tailoring approach (local model)

- **Main model-building approaches (over-lapping)**: Regression, analogy (including case-based reasoning), function points (including story points, use case points, expert system, feature points, etc), classification and regression trees, artificial neural network, Bayesian Belief Networks (which models expert knowledge).
- **Main size variables**: Estimated lines of code, function points, use case points, user story point, number of screens, etc.
- **Principles for model building/selection of model-building approach:**
  - Select a small set of variables you believe are the most meaningful in your context.
  - Develop estimation models with few variables and apply a simple model development approach (e.g., regression analysis or Bayesian Belief Networks).
  - Use the record on previous projects of relevant type to guide selection of model-building approach.

Selection of generic model

- **Examples of generic models**: COCOMO, SLIM, PRICE-S, etc.
  - These models may have tailoring possibilities, but are typically fixed regarding choice of variables and basic formulas.
- Look at the track record on projects similar to yours.
  - Non-calibrated (generic model use) is, at its best, highly discussable, i.e., hardly any study supports the use of such generic models.
- Do you understand the model?
  - Do not use “black-box” models, i.e., models where the tool vendor does not reveal the “inside” of the model.

Tailoring of expert estimation processes

Tailor the process with elements from all the categories below:

- **Selection of expert(s) relative to:** Skill, “motivation”, experience, accuracy record
- **Problem solving approach:** Top-down (outside view), Bottom-up (inside view), “Inside-out” (inside view + outside view on activity proportions)
- **Group process:** Mechanical combination (experts in “isolation”), Unstructured, Structured (e.g., Delphi-method, Planning Poker or Role-playing)
- **Variance in experience/background/role**
- **Remove irrelevant information**
- **Avoid conflicting goals, etc.**
- **Use of historical data**
- **Require an explicit process (no “gut feeling”)**
- **Checklists**
- **Work-breakdown structure**
- **Combine with rules-of-thumb (one Use Case-point equals about X work-hours)**

Effort estimation uncertainty analysis

Probabilities: A late invention (and we are not good at assessing it)

Task: What is the number of inhabitants in Norway

Minimum

Maximum

Be 99% confident to include the correct number in the min-max interval!

How sure is “almost sure”?

- Our field studies of software companies:
  - Some projects use a minimum-maximum interval method (e.g., PERT)
  - Some did not state how likely they thought it would be to include the actual effort, other assumed a 90% or 98% likelihood.
  - In reality, as much as 40% of the projects was outside the min-max interval!
- In experiments we find that when project managers claim:
  - Almost certainty, this mean about 60% certain
  - “60% certain” = “75% certain” = “90% certain = “99% certain”

Realism-conflicting goals

- Informative assessments excludes wide (realistic) intervals
- Rewards for over-confidence
  - Realism used as indicator for lack of skill!
- The clients don’t like high uncertainty ….
- If the uncertainty is too high we will not be allowed to start this project ….

In the middle of this one is asked to be realistic regarding the uncertainty!

Two views on the development effort uncertainty: Inside view

- Inside view, i.e., break-down of uncertainty:
  - min-max per activity
  - analysis of known risk (High/medium/low)
- **Strength**: Identification of risk elements and the need for risk management
- **Weakness:** Under-estimation of uncertainty through poor methods of combining individual risk elements and lack of focus on “unknown risk”.

Two views on the development effort uncertainty: Outside view

- Outside view, i.e., look at the project and it’s uncertainty as a whole
  - Compare with uncertainty of previously completed, similar projects.
- **Strength:** Increased realism in uncertainty assessment.
- **Weakness:** Does not contribute much to how to reduce the risk. Dependent on that similar projects are available and that learning effects are properly adjusted for.

They need to be combined!

- Inside view necessary for planning.
- Outside view necessary for proper budgeting.
- When the total uncertainty derived from the two viewpoints differ, this indicates that more analysis is needed.

It matters how you ask ...

- The realism of the uncertainty assessment depends strongly on how you ask:
  - Don’t ask like this:
    - What is the maximum/minimum effort?
  - Ask rather like this:
    - How large proportion of similar project have been overrun with more then X (where X for example is 50%)
    - Require documentation, if realism is essential.
  - The improvement in realism may be surprising large.

Example from a Norwegian organization

Table 2. Distribution of Estimation Error of Similar Projects

	Teams (Group B only)									
Estimation Error Category	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**Mean value**
>100% overrun	45	18	10	10	10	5	10	0	18	14
50-100% overrun	20	40	35	20	10	5	20	5	25	20
25-49% overrun	15	22	25	30	30	35	40	20	30	27
10-24% overrun	10	15	25	20	30	45	20	40	15	24
+/- 10% of error	7	4	0	5	10	10	10	20	12	10
10-25% too high estimates	3	1	0	10	5	0	0	10	0	3
24-50% too high estimates	0	0	0	0	5	0	0	5	0	1
>50% too high estimates	0	0	0	0	0	0	0	0	0	0

What would be the p70% estimate of Team 17?

Recommendations

- Assume over-confidence, particularly in large and complex projects if the judgment is based on an inside view.
- Reward realism and create situations that enable realism (e.g., no pricing, bidding elements).
- Require documentation of uncertainty assessment, do not rely on the experts' feeling-of-risk.
  - Simple models outperform expert judgment in uncertainty assessment (but not in effort estimation!).
- Use the proposed method (and not the traditional min-max method) when asking for uncertainty assessments.