[simula . research laboratory]

When Should We Trust Expert Judgment in Software Development?

Magne Jørgensen
magnej@simula.no
München, May 20, 2008

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 / inconsistency*
- Those recognized as experts by at least one other person?
- 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."*

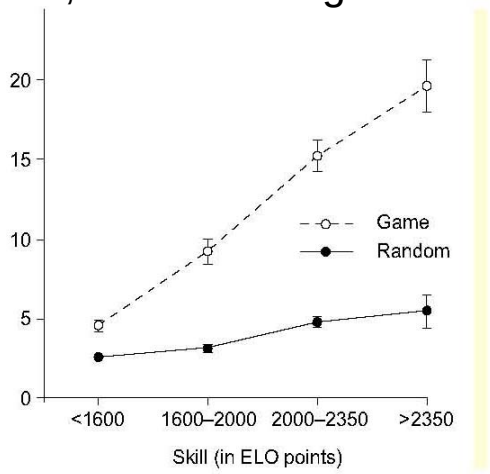
What is the Difference Between Experts and Non-Experts in Chess?

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

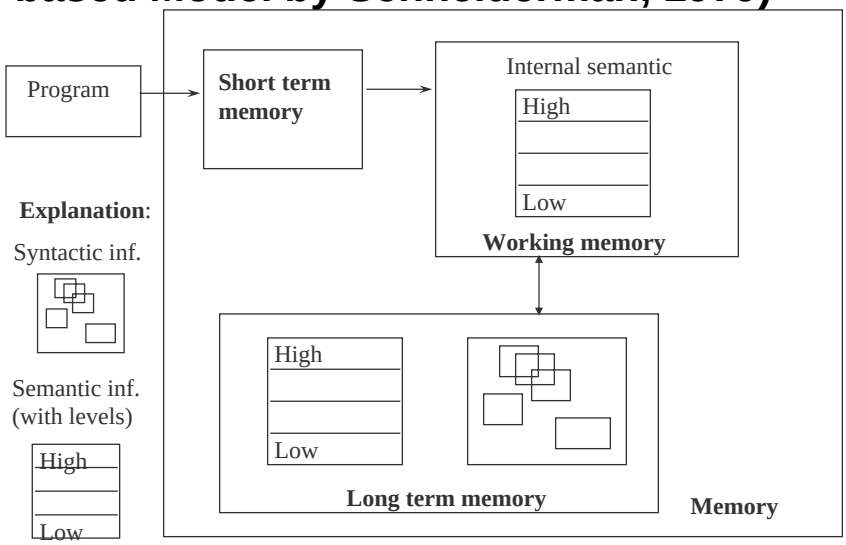
- number of moves analyzed per minute?
- depth of IF-THEN analysis?
- short term memory?
- search heuristic?
- filtering of bad moves?
- recall of randomly positioned chess pieces?
- better working memory capacity?
- ability to analyze larger units, e.g., analyze patterns rather than single pieces?

[simula . research laboratory]

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)



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?

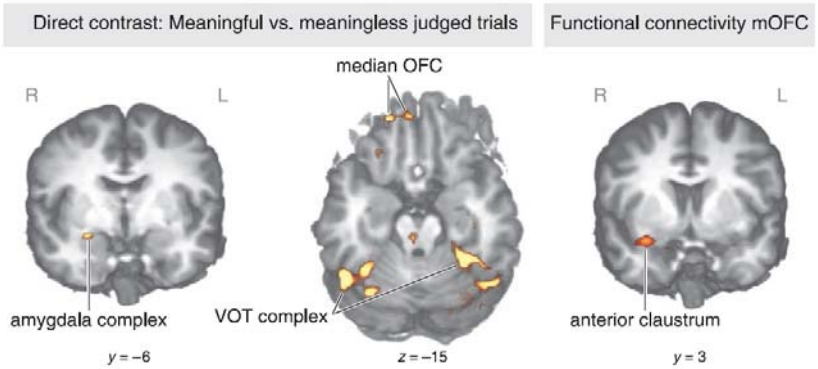
- There is an essential difference between expertise and ability to know how the expert judgment are derived.
  - Lagnado et al. 2006: "*Studies suggest that quite different regions of the brain are involved in learning and insight about learning."*
- Essential parts of the expert judgment are unconscious/intuition-based. 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.

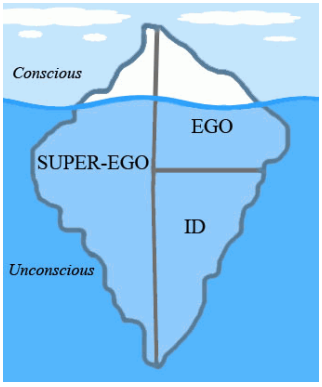


[simula . research laboratory]

9

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.



[simula . research laboratory]

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

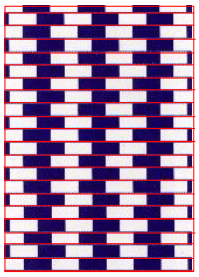
Intuition:

- 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
  - E.g., in many types of clinical judgment and effort estimation uncertainty judgments
- can be extremely inconsistent
  - E.g., our studies on expert estimation of software development effort
- may be unable to transfer extensive knowledge into accurate judgment
  - E.g., mutual funds
- are impacted by many irrelevant factors
  - E.g., the weather may impact how people's abilities are judged (see next page)

“Clouds Make Nerds Look Better”

- Sunshine increases tipping, impacts stock-market, and, increases happiness.
- Study of university applicants:
  - 12% higher chance when sunshine compared to worst cloudcover.
  - Nerds had significantly higher chance compared to non-nerds on cloudy days.
    - Nerd-factor measured as academic rating divided by social rating (e.g., leadership).

The dilution effect - Design

- 44 industry participants
- INSTRUCTIONS: Please, give each of the following estimation model evaluation factors a weighting in %. (Weight should corresponding to the weight they would put on the score in an real decision process)
- 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 highest 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.**
  - Comment 1: More information is not only good, especially when it's only slightly important.
  - Comment 2: Questionnaires that ask people to assess importance of factors on Likert scales (e.g., 1-7) are also vulnerable to the dilution effect, but seemingly less than the relative weighting measure.

Priming I - design

- 111 industry participants
- Phase 1
  - **Group Average**: Complete 3 average estimation related tasks (average height of Norwegian 18-year old men, etc.)
  - **Group Analogy**: Complete 3 analogy identification tasks (town most similar to Lillehammer wrt inhabitants, etc.)
- Phase 2
  - Estimate the productivity of a project based on historical data
  - We could derive from the estimate whether an analogy or average-based use of the historical data had been used.
- Do you think the enforced strategy in Phase 1 had an impact on the strategies used on the unrelated task in Phase 2?

Priming I - results

- Results:
  - Group Average: Approx. 90% selected an average-based strategy
  - Group Analogy: Approx. 60% select an analogy-based strategy
- And, they did not notice the impact from the previous use of strategy!
- **Expert judgment can be very inconsistent, partly due to the priming effect.**
  - A previous study gave that the median effort estimation inconsistency when estimating the same task with one month in-between was about 50%! (Grimstad & Jørgensen, 2007)

Priming II - 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 II - results

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

External validity?

- Previous studies were mainly in contexts with small tasks and/or high time pressure.
  - Which is relevant, but not the only (or even the typical) situation.
- This may lead to increased use of surface indicators in comparison to estimation processes where several hours are spent and more information collected.
- Would we be able to replicate the findings (i.e., that it's very easy to impact the estimates) in field settings?

A field experiment (analysis in progress) ...

- Forty-six companies from various low cost countries estimated the same five projects: Russia (15 companies), Ukraine (5), India (7), Bulgaria (4), Romania (3), Pakistan (5), Belarus (2), Moldovia (1), Poland (1), Serbia (1), Slovakia (1), and Vietnam (1). T
- We accepted only estimators with professional experience from projects similar to those to be estimated, i.e., we allowed only reasonably experienced estimators.
- The companies were hired and paid for their estimation work, i.e., they did not (seen from their point of view) participate in an experiment.
  - The companies were on average paid about 1500 USD for the estimation work, ranging from 400 to 4000 USD.
  - The effort a company estimated to spend on the estimation of the five projects varied from about 40 work-hours to about 200 work-hours.
  - They were told that they would not be invited to develop the systems, but that their job was to provide realistic effort estimates.
- Random allocation to “manipulations” of requirement specification.

High variance in estimates ...

Effort Estimation Distributions

Project	Minimum	Q1	Median	Q3	Maximum
RDinner	45	119	190	339	1320
DocAssist	61	186	330	438	1200
AA	160	316	509	715	2280
DES	17	134	192	347	1160
IMWOS	240	649	895	1316	3371

Length of specification

- **H1:** A reduction in number of pages of the requirement specification leads to lower effort estimates, even when the written content is exactly the same.
  - Manipulation: Text identical. One version 3 pages, the other 12 pages.
  - Length of specification is clearly not relevant for the development effort, but will it be used as an indicator?

Results: Length of specification (H1) [System: DocAssist]

The Effect of the Reduced Length of Specification

Group	Median
Manipulated (3 pages spec.)	295 work-hours (n=24)
Ordinary (12 pages spec.)	330 work-hours (n=22)

A small effect (perhaps).

Effect seems to be reduced with more time and expertise

Numerical anchor

- **H2:** Presenting the actual effort of the system to be replaced (a low numerical value in our case) early in the requirement specification leads to lower effort estimates.
  - The following text was included early in the manipulated requirement specifications: "*The preliminary budget of the new system is $10 000 [corresponding to about 100 work-hours with typical pricing in the country in which it will be built]. The preliminary budget is not built on any knowledge about the actual cost of developing the new system, and will, if needed, be extended to cover the expenses necessary to build a quality system with the desired functionality.*"
  - 100 work-hours is a very low value for this project and the companies were instructed to not use this as input to their effort estimate, but they may use it unconsciously.

Results: Client expectation (H2) [System: IMWOS]

Numerical Anchor

Group	Median estimate
Manipulated (client's expectation)	724 work-hours (n=23)
Ordinary	956 work-hours (n=23)

A significant, large effect.
However, lower effect than in our previous laboratory experiments.

Time schedule pressure

- **H3:** Information about that the client requires a short development period leads to lower effort estimates.
  - The following text was included early in the manipulated requirement specifications: "[the client] *expects that the system development starts February 3, 2008 and can be launched February 23, 2008. This three week period should include all development and testing.*"
  - A short development period should lead to more rather than less use of effort, but may also induce "wishful thinking" or the belief that the system is small.

Results: Time schedule pressure (H3) [System: DES]

The effect of time schedule pressure

Group	Median
Manipulated (Informed that the client expected the system to be developed during 3 weeks period.)	142 work-hours (n=24)
Ordinary	214 work-hours (n=21)

Very large effect!

So, when should we trust experts?

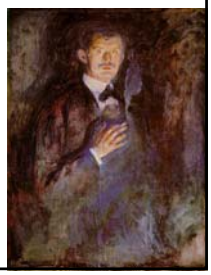
- When they have extensive “deliberate practice” in the particular problem to be solved.
  - See studies by Ericsson and by Shanteau.
- When the context includes little irrelevant and/or misleading information leading to well-known effects (dilution, anchoring, priming, wishful thinking).
  - See the “human biases” studies, e.g., by Kahneman & Tversky
- When the learning environment is not “wicked” (feedback is timely and enable learning).

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 not a very 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, Big five test, IQ-test, ...)
  - Probably not of much help.
- Slightly depressive people?
  - Yes ☺. They are on average most realistic regarding own abilities.



[simula . research laboratory]