

[**simula** . research laboratory]

RELATIVE ESTIMATION OF SOFTWARE DEVELOPMENT EFFORT:

IT MATTERS WITH WHAT AND HOW YOU COMPARE

Magne Jørgensen

scientia

NM in effort estimation (end of this session)

The winner gets a gift card worth 2000 NOK at a nice restaurant
(Restaurant Eik) and this trophy



[**simula** . research laboratory]

scientia



[**simula** . research laboratory]

scientia

Judgment is relative

Explicit/conscious relative judgment:

1. How much larger/smaller (in work-hours, cost or % effort) is this project compared to the project we did earlier?
2. How many “story points” is “user story” X? (comparison with a reference (baseline) user story)
3. Which category of task is this? (comparison with other categories of tasks, e.g., whether a new task is “small”, “medium” or “large”)

Implicit/unconscious relative judgment:

1. Estimation of one task just after another, frequently leads to the use of the first task as the reference for the second.
2. We unconsciously activate and compare with previous experience and get a first impression of project size. The activation can easily be manipulated, e.g., through activating information about smaller tasks by describing a new task as “minor extension”.

[**simula** . research laboratory]

scientia

Four comparison effects

- **Assimilation vs contrast:** Are tasks getting more or less similar through comparison?
- **Asymmetry vs symmetry:** Does it matter whether we compare X with Y or Y with X?
- **Framing:** Do we get impacted by how comparison tasks (estimation instructions) are formulated (framed)?
- **Central tendency:** Does it matter what is considered as the middle or default value or category of projects?

[**simula** . research laboratory]

scientia

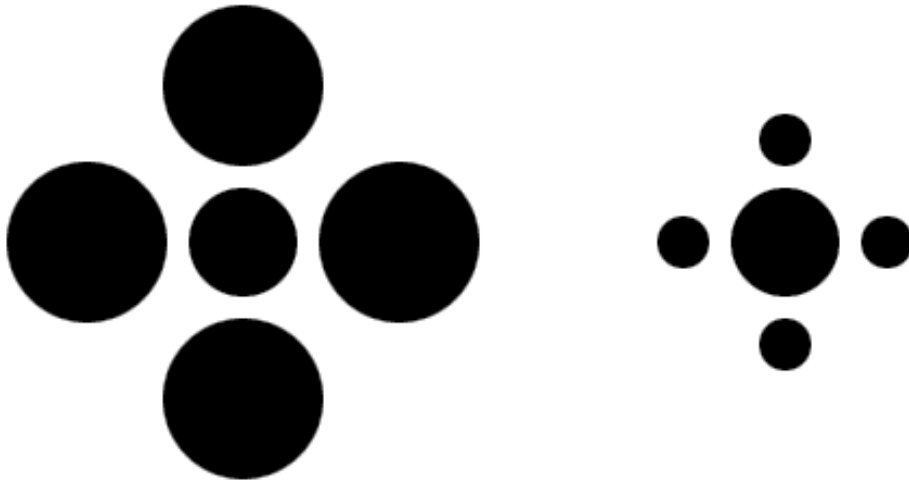
Assimilation vs contrast: Jumping after Wirkola or Eddy the Eagle?



[**simula** . research laboratory]

scientia

Sometimes there is a contrast focus ...



[**simula** . research laboratory]

scientia

And sometimes a similarity focus



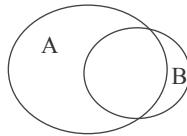
[**simula** . research laboratory]

scientia

Assymetry vs symmetry

Tversky would predict asymmetry in most comparisons, based on the observations that:

- Comparisons are “pattern matching” processes.
- There is a neglect of features present only in the reference.



- B (the target) is perceived similar to A (the reference), due to the neglect of the features in the large set (A-B).
- A (the target) is not similar to B (the reference) due to the neglect of less feature, i.e., those in the smaller set (B-A).

[**simula** . research laboratory]

scientia

Comments

Other consequences of the Tversky’s asymmetry:

- We will tend to think that “You are more similar to me, than I am similar to you”.
- USA is perceived more similar to Mexico (also in number of inhabitants) than Mexico is to USA.
- The less we know about a new project, the more similar it will be to the project with which we compare with.
- Your spouse will better off if you compare her/him with “the perfect person”, rather than comparing “the perfect person” with her/him.



[**simula** . research laboratory]

scientia

Study 1: Assimilation and asymmetry in relative estimation

- Two specifications:
 - A: A very simple web-registration system for a summer party
 - B: A system for registration of scientific studies
- Real difference in actual effort: B requires at least 100 work-hours more than A
- Participants: about 100 developers from an outsourcing company
- Randomly allocation of comparison direction in the relative estimation.
- All participants were asked to estimate the effort of developing the larger system (B), in work-hours, *after* completing the relative estimation.
- More about the study on:
 - http://simula.no/publications/Simula.simula.814/simula_pdf_file (Study 1)

[**simula** . research laboratory]

scientia

System A specification: Registration of Summer Party Participants

Participants should be able to register their participation at the summer party on the web by submitting their name and their email address. The layout should be as in the figure below. The system confirms (on web) that the data is registered. There is no data validation (duplicate check, etc). The data is stored in a database. Generation of reports, such as attendee lists, is done manually, i.e. by querying the database.

Layout of input screen

Summerparty 2011

Yes of course I will attend the 2011-edition of the Simula Summerparty!

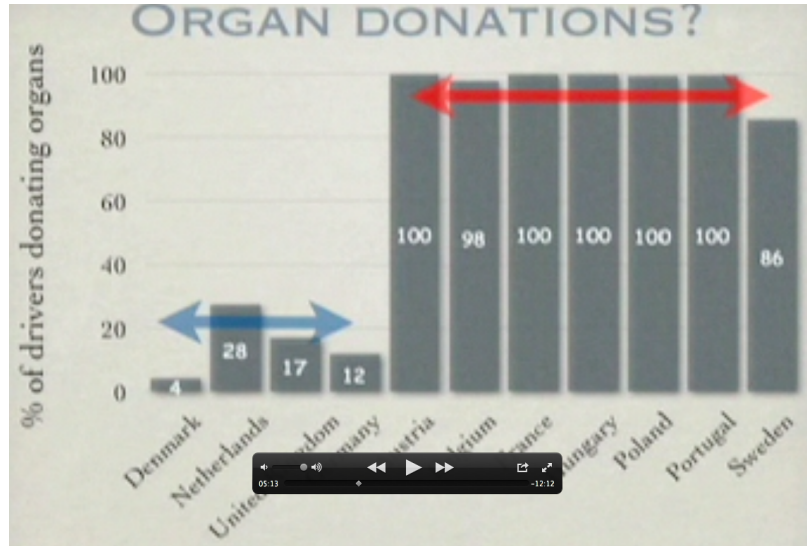
Name *

Your E-Mail Address *

[**simula** . research laboratory]

scientia

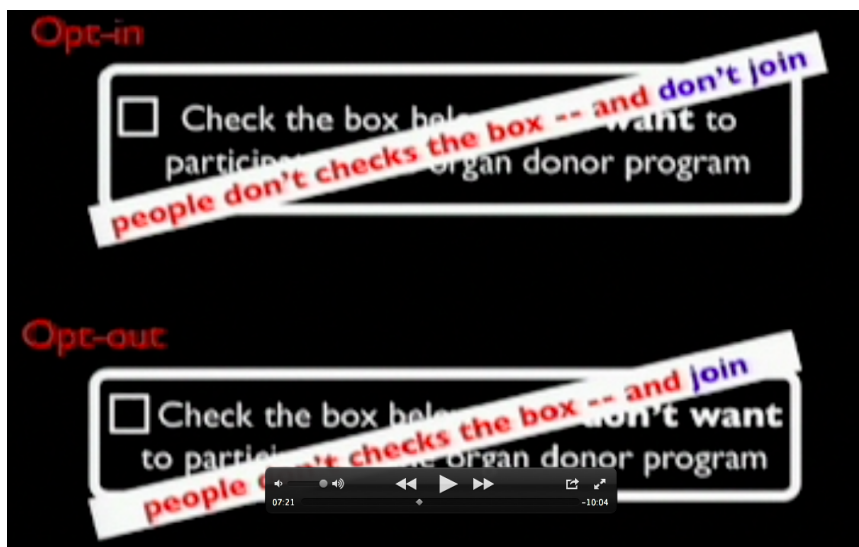
Why is there this huge difference in organ donations? (from a ted.com video with Ariely)



[[simula](#) . research laboratory]

scientia

Framing of the question



[[simula](#) . research laboratory]

scientia

Framing in effort estimation ...

- **Hypothesis:** If we ask “How many % is X of Y?” we would tend to believe that X is smaller of Y. (At least we will tend to believe that the person asking the question believes this, and get affect by that.)
- **Study:** Two systems (C and D) of about the same size (about 300 work-hours each). 35 software developers were asked to either state how many % C is of D (in terms of work-effort) or how many % D is of C.
- **Results** (median values):
 - “I think that C is 70% of D”
 - “I think that D is 78% of C”
- **Comments:**
 - We get similar results on other types of tasks (including results from a study on story points, which is implicitly a X % of Y type of question) and when tasks are very different in size, e.g., A is 400% of B.
 - The effect seems to go away when explicitly telling that the person asking has no knowledge!

[**simula** . research laboratory]

scientia

When in doubt, we tend to select the?



[**simula** . research laboratory]

scientia

... middle value

- The middle category, when in doubt, may have an immense power on our choices. Is this the case when we estimate effort, as well?
- Two studies:
 - Students estimating the same programming task using either the “Fibonacci” scale (1, 2, 3, 5, 8, 13, 20, 30, 40, more than 40 work-hours) or a linear scale with numbers 1-40 + the category more than 40 work-hours.
 - Replication with four professional teams using Planning Poker and the Fibonacci scale on some user stories and the linear scale on other user stories.
- Results:
 - The median estimates of the students were close to the middle value of the scale, i.e., 8 work-hours for the Fibonacci and 20 work-hours for the linear scale. The effect was highest on students with the lowest skill.
 - Similar effects for the professional teams, e.g., average task effort of 5 (Fibonacci scale) vs 7 (linear scale) work-hours for a user story.
- More results on: http://simula.no/publications/Simula.simula.1282/simula_pdf_file

[**simula** . research laboratory]

scientia

Summary and recommendations

- When we compare tasks, they usually get more similar.
 - Avoid dissimilar reference tasks. Use a medium large user story as baseline.
- The less we know about a new project, the more similar it gets to previously completed ones.
 - Be aware of this effect and put extra effort in searching for features only present in the reference when comparing with other projects.
- Comparing X with Y is typically NOT the same as comparing Y with X.
 - There is power in the control of the direction of comparisons!
- We get (unconsciously) affected by what (we think) the person asking believes him/herself through his/her framing of the question.
 - “Percentage of” seems to be especially strong and should be avoided.
- The less we know, the stronger we get affected by the middle or default value, e.g., of a scale.
 - There is power in the control of the middle or default value, especially when high estimation uncertainty.

[**simula** . research laboratory]

scientia