# Expecting Others to Be Good – Explaining the Polarity Effect

DR. PASCALE WILLEMSEN, UNIVERSITY OF ZURICH

*(Pascale.Willemsen@uzh.ch; www.PascaleWillemsen.com)*
*Link to related publications: https://onlinelibrary.wiley.com/doi/10.1002/tht3.488*

## WHAT ARE THICK CONCEPTS?

Philosophers distinguish between two types of evaluative concepts: *thin* and *thick*.

> *Thin concepts*: merely evaluate positively or negative ('good', 'bad', 'right', 'wrong'…)
>
> *Thick concepts*: evaluate positively or negative and also provide descriptive information ('honest', 'friendly', 'rude', 'cruel'…)

The attempt to define thick concepts has revolved around five central questions:

1. *Separability Question*
   Can the evaluative content of a thick concept be separated from the descriptive content – at least in principle?
2. *Location Question*
   By what means is the evaluative content communicated – by semantic or pragmatic means?
3. *Centrality Question*
   Are thick or thin concepts more basic – what can be reduced to what?
4. *Variability Question*
   Can a thick concept vary with respect to the evaluative content it carries?
5. *Action-Guidingness Question*
   How are thick concepts related to reasons for action?

## THE LOCATION QUESTION

The *Location Question* asks where exactly we can find the evaluative dimension of a thick term or concept. The *Location Question* presupposes the separability of the evaluative and the descriptive, at least for the sake of theoretical purposes.

In a nutshell, two options are discussed

1. *Semantic View*: The evaluation is part of the semantic content of a thick concept; it is semantically entailed.
2. *Pragmatic View*: The evaluation is communicated beyond what is literally said and part of the pragmatically conveyed speaker-meaning.

## CANCELLABILITY TEST FOR CONVERSATIONAL IMPLICATURES

Conversational implicatures are information that is conveyed beyond what a speaker literally says.

Example: "Sally regrets drinking instant coffee this morning".

From this statement, we learn that Sally had instant coffee this morning and that she has a negative feeling towards her choice of beverage. We might likely infer that Sally dislikes the taste of instant coffee or that she usually does not drink instant coffee. Let us call every piece of information that can be inferred from the target statement an implication of this statement.

Statement: Sally regrets drinking instant coffee this morning.

(1) Implication 1: Sally drank instant coffee this morning.
(2) Implication 2: Sally has a negative feeling towards drinking instant coffee this morning.
(3) Implication 3: Sally does not like the taste of instant coffee.
(4) Implication 4: Sally usually does not drink instant coffee.

The cancellability test can help us identify which implications are conversational implicatures. The idea is to take the original, implication-triggering statement

"Sally regrets drinking instant coffee this morning"

and to check if we can cancel the triggered implication explicitly without creating a contradictory statement. Here are the resulting statements, with '#' indicating that the statements sound contradictory and '!' indicating a felicitous statement.

(1*) **#** Sally regrets drinking instant coffee this morning. However, I am not saying that she drank instant coffee this morning.
(2*) **#** Sally regrets drinking instant coffee this morning. However, I am not saying that she has negative feelings about drinking instant coffee this morning.
(3*) **!** Sally regrets drinking instant coffee this morning. However, I am not saying that she does not like the taste of instant coffee. *(She just wanted to drink less coffee.)*
(4*) **!** Sally regrets drinking instant coffee this morning. However, I am not saying that she usually does not drink instant coffee. *(She drinks it all the time, but the fancy hotel in which she had breakfast had so many better options.)*

Sentences that sound contradictory (1* and 2*) are not conversational implicatures. Being cancellable is a necessary condition for being a conversational implicature. 3* and 4* might be conversational implicatures.

## CANCELLABILITY AND THICK CONCEPTS

Let's apply this rationale to thick concepts. We take a statement containing a thick concept (e.g. friendly or rude). The implication triggered is that this is something positive or negative, respectively. We then cancel this evaluation explicitly to determine whether the resulting statement sounds contradictory:

*Example:* "What Rachel did last week was friendly/rude."

*Implication triggered:* "This is something positive/negative about her behaviour."

*Cancellability Statement:* What Rachel did last week was friendly/rude, but by that I am not saying something positive/negative about her behaviour that day.

What prediction do the Semanticists and the Pragmatists make?

*Semantic View:* If the Semantic View is correct, then the evaluation of a thick concept should not be cancellable. There is no significant difference between Semantic Entailments and Thick Concepts.

*Pragmatic View:* If the Pragmatic View is correct, then the evaluation should be cancellable. There is no significant difference between Conversational Implicatures and Thick Concepts.

## EXPERIMENT 1: HOW IS THE EVALUATION CONVEYED?

7 × 1 between-subject design (Semantic Entailment; Generalised Conversational Implicature, Particularised Conversational Implicature, Thick Positive & Thick Negative in 2 different Embeddings (Behaviour & Character))

779 participants recruited on Prolific

62.3% female, 36.9% male, 0.8% non-binary; Mean Age: 34.5 years

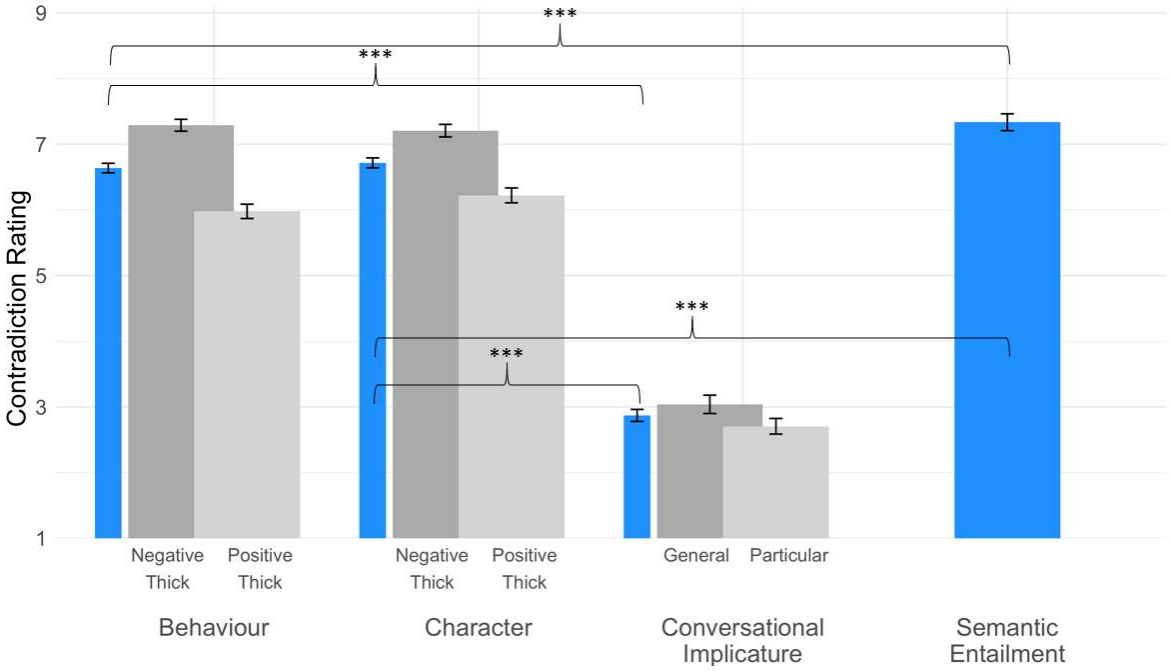
### STIMULI:

*Behaviour*: "What Tom did last week was X, but by that I am not saying something positive/negative about her behaviour that day."

*Character*: "Tom is X, but by that I am not saying something positive/negative about him".

| Domain | Positive | Negative |
|---|---|---|
| Meta-Concepts | Virtuous | Vicious |
| Care/Harm | Compassionate | Cruel |
| Loyalty/Betrayal | Honest | Manipulative |
| Authority/Subversion | Friendly | Rude |
| Fairness/Cheating | Generous | Selfish |
| Fariness/Cheating & Care/Harm | Courageous | Cowardly |

### RESULTS



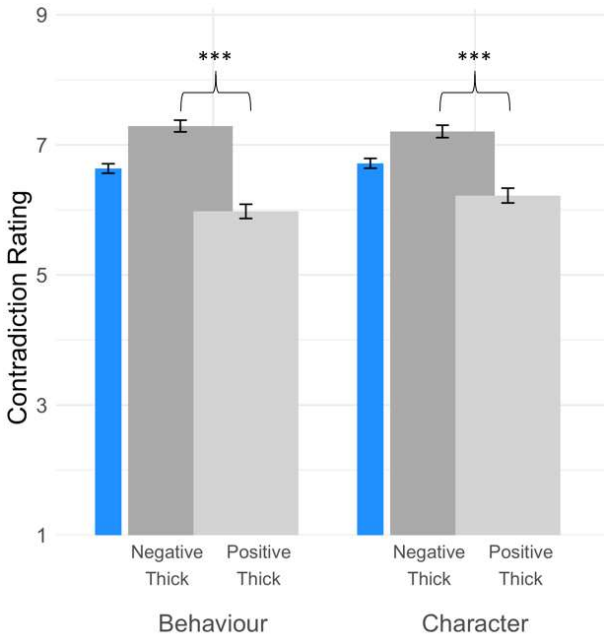
| Nested group | Mean | Median | SE | SD |
|---|---|---|---|---|
| Behaviour: Negative thick | 7.29 | 8 | 0.09 | 2.35 |
| Behaviour: Positive thick | 5.98 | 7 | 0.11 | 2.83 |
| Character: Negative thick | 7.21 | 8 | 0.10 | 2.48 |
| Character: Positive thick | 6.22 | 7 | 0.11 | 2.93 |
| Conversational implicature: General | 3.04 | 1 | 0.14 | 2.92 |
| Conversational implicature: Particular | 2.70 | 1 | 0.12 | 2.51 |
| Semantic entailment | 7.33 | 9 | 0.13 | 2.72 |

## INTERPRETATION OF STUDY 1: WHAT DOES THAT MEAN?

The results do not support the predictions of the Semantic View; they neither support the predictions of the Pragmatic View. One might think that sounds like bad news. But for whom? And does it actually?

Beyond philosophers' predictions, we found an asymmetry between positive and negative thick concepts.

Polarity Effect: The evaluation of a negative thick concept is harder to cancel than the evaluation of a positive thick concept.




## EXPERIMENT 2: CONFIRMATION OF THE POLARITY EFFECT
(OMITTED IN PRESENTATION)

So far, we only have limited data on the polarity effect, namely in a very specific embedding. Thick concepts are not only used to refer to an individual person or behaviour, but also to groups of people. Does the polarity effect show up again?

3 × 2 between-subject design with the independent factors Valence (Positive vs. Negative) and Scope (Proper Name, Limited Scope, Generic Statement)

387 participants

44% female, 56% male; Mean Age: 34.5 years

## STIMULI

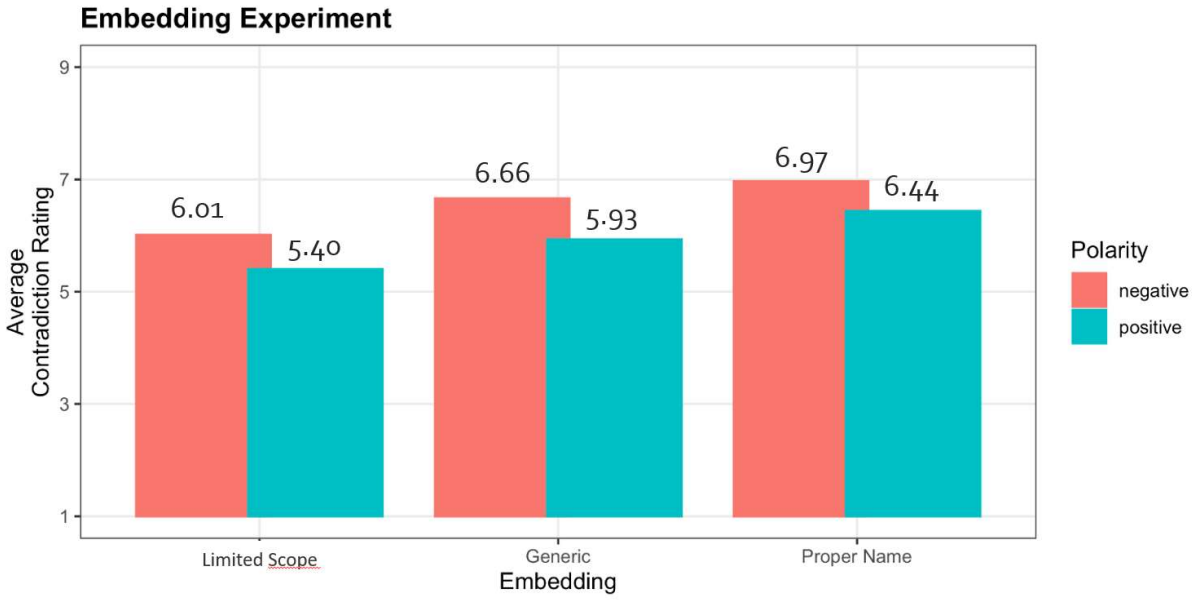
*Proper Name*: "Tom is X, but by that I am not saying something positive/negative about him".

*Limited Scope*: "Some people are X, but by that I am not saying something positive/negative about them".

*Generic Statement:* "People are X, but by that I am not saying something positive/negative about him".

We used the same thick concepts as in Experiment 1.

## RESULTS




## INTERPRETATION OF EXPERIMENT 2

The Polarity Effect shows up in all three embeddings. Negative evaluations are always harder to cancel than positive ones.

## EXPERIMENT 3: WHAT ABOUT THIN CONCEPTS?
(OMITTED IN PRESENTATION)

So far we have only investigated thick concepts, and we have implicitly assumed that the Polarity Effect is a Thick Concept Effect. But this assumption might be wrong. If it is wrong, we need to search for a very different kind of explanation.

2 × 1 between-subject design

101 participants recruited on Prolific

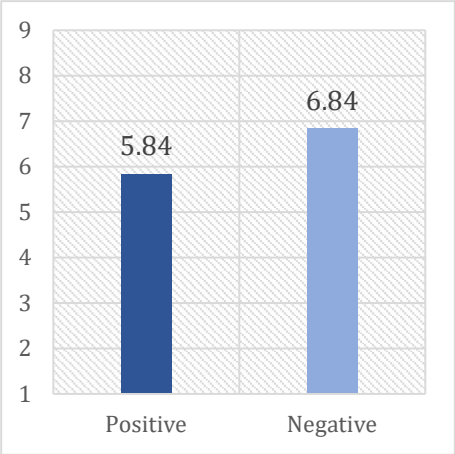
58% female, 41% male, 1% non-binary; Mean Age: 34.3 years

## STIMULI

Instead of thick concepts, we now used *good* and *bad*; *right* and *wrong*. This is the statement that we used:

"What Amy did last week was good/bad/right/wrong, but by that I am not saying something positive/negative about her behaviour that day."

## RESULTS





## INTERPRETATION OF EXPERIMENT 3

The Polarity Effect shows up for thin concepts as well. Negative evaluations are always harder to cancel than positive ones.

The Polarity Effect is not a Thick Concept Effect, it is an Evaluative Language Effect. This calls for a more systematic explanation.

## EXPERIMENT 4: REVERSAL OF THE EVALUATION
(OMITTED IN PRESENTATION)

Perhaps the Polarity Effect is explained by an asymmetric flip in the anticipated speaker-meaning. There are cases in which being friendly, generous, honest, etc. are not good, bad rather bad things (e.g. telling someone something that is true but really hurtful or being friendly or generous to a person who really does not deserve it). Positive cancellability sentences are interpreted as a reversal of the evaluation, such as:

"What Amy did last week was friendly, but by that I am not saying something positive about her behaviour that day. *In fact, I'm saying something negative about it. She was friendly to a racist, and this was really bad*."

Coming up with similar cases is much harder for negative cases. As a consequence, people might interpret the Cancellation Statements differently, depending on whether they are in the Positive or Negative Condition. That might explain the effect – and that would be a problem.

5 × 1 between-subject design

544 participants recruited on Prolific

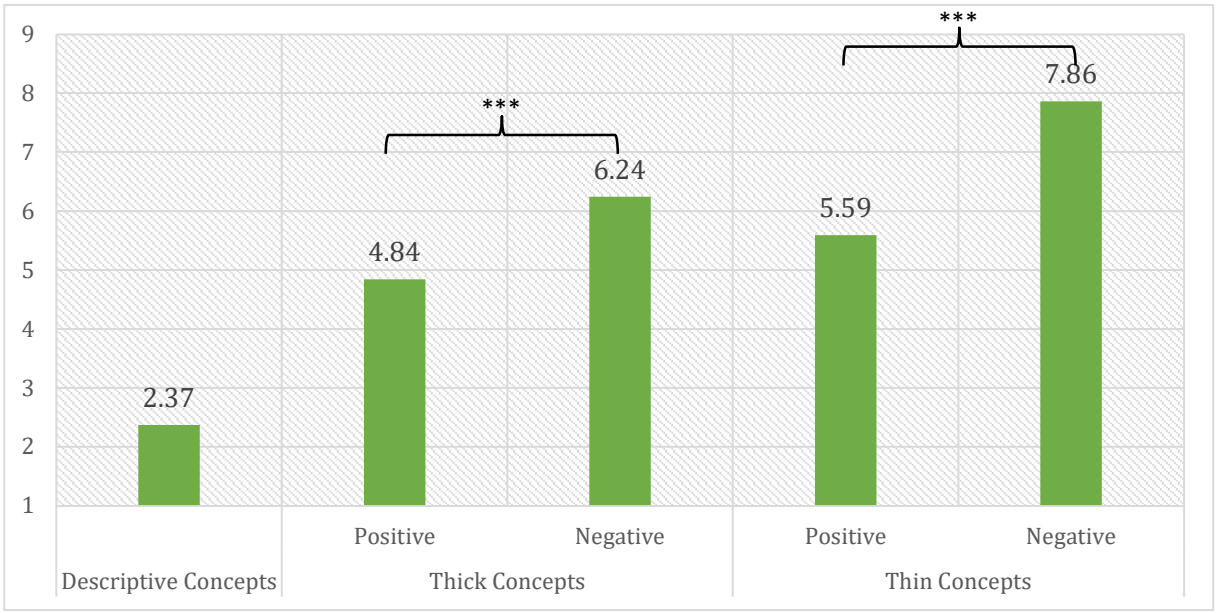
62% female, 37% male, 1% non-binary; Mean Age: 34.5 years

## STIMULI

We used the same thick concepts as in Experiments 1 and 2. We also used thin concepts (Experiment 3) and some descriptive concepts as a control. We made the following modifications:

"What Tom did last week was X, but by that I am not saying something positive or negative about her behaviour that day. I mean this in a fully neutral way."

## RESULTS




## INTERPRETATION OF EXPERIMENT 4

The Polarity Effect cannot be explained by a flip in the anticipated speaker-meaning. If the speaker makes it explicit that she uses the term in a fully neutral way, the Polarity Effect still occurs.

# EXPECTING OTHERS TO BE GOOD

How can we explain this effect? We believe that data is reliable, and it shows something deep about the psychology underlying evaluative language use. We suggest that people believe others to be good and to do good things. And this shapes their use of evaluative terms.

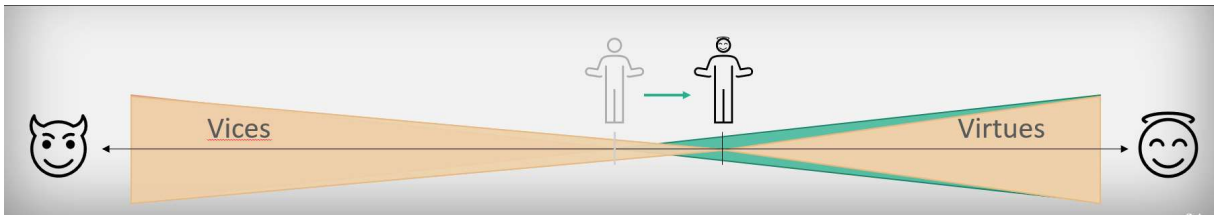
For a society to not fall apart, we need people to be cooperative and respects the rights of others. They cannot harm others, they cannot cheat or betray others, and they cannot destroy cooperation by being offensive and disrespectful. In addition, for a society to work properly, we need people to do positive, good things for others, such as giving a bit of what you don't need to others, forming stable relationships by saying nice things, or being sensitive to the other people's needs.

**Suggestion 1:** We generally expect others to be good and to do good things. Being slightly virtuous is the standard of an average, normal, morally acceptable person who we accept as part of our society.

How can this explain the effect?

Because we expect others to be a bit generous, a bit honest, a bit friendly…, there is a way of being these things that is nothing extraordinary or praiseworthy. Saying "Tom is friendly" can mean two things, depending on the conversational context:

1. *Non-Evaluative Use:* Tom meets the general expectation of decent behaviour. He is just as friendly as he should be.
2. *Evaluative Use:* Tom is an extraordinarily friendly person and deserves praise for being more friendly than other people.




**Suggestion 2:** Positive terms (both thing and thick) can be used in two ways. They can be used in a non-evaluative and an evaluative way. Negative terms can only be used in an evaluative way.

Various philosophers have also suggested that actions are only praiseworthy if they are supererogatory, that is good beyond an expectable, minimal degree of decency. See, e.g., Nathan Stout (2020. On the Significance of Blame), Andrew Eshleman (2014. Worthy of Praise), and McKenna (2012. Conversation and Responsibility).

## EXPERIMENT 5: DO WE EXPECT OTHERS TO BE GOOD?

To test this hypothesis, we designed the fifth experiment.

2 × 2 x 2 mixed design, with the independent factors Concept (Thin; Thick) and Valence (Positive; Negative) which were tested between subjects, and Question (Expectation; Approval) being tested within-subject.

352 participants recruited on Prolific, 12 had to be excluded

### STIMULI

Participants received the following prompt:

> **(EXPECTATION)**
> Please think about what kind of behaviour you expect of people in general.
>
> Please consider the following statement:
>
> *What Tom did was [thick or thin term].*
>
> To what degree is Tom's behaviour below your expectations, is exactly what you would expect of him, or exceeds your expectations?

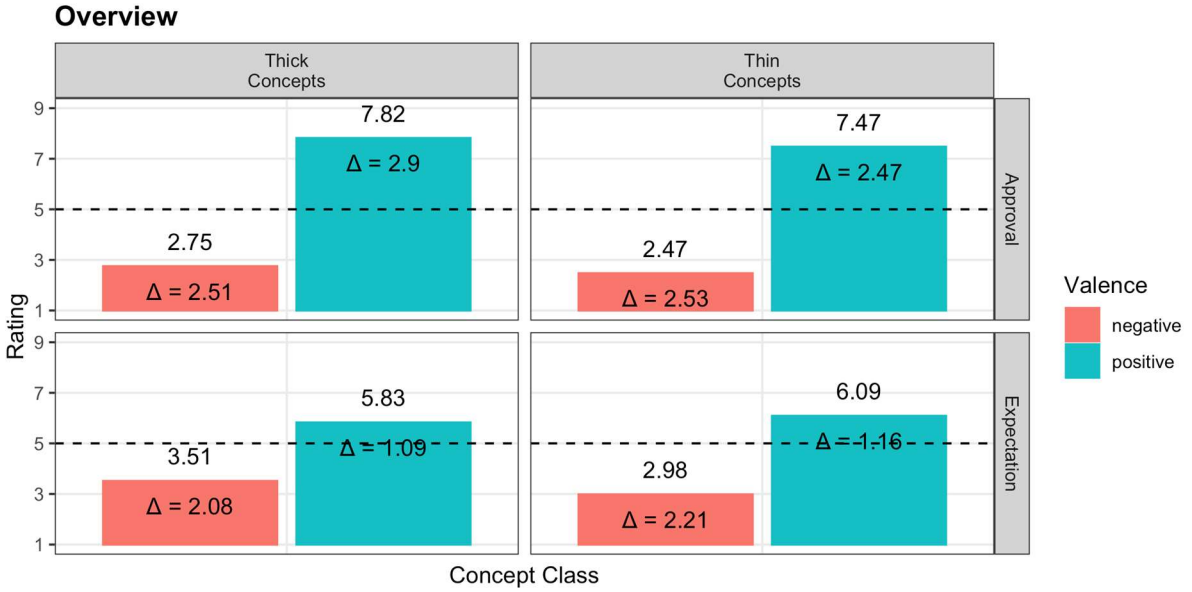
Ratings were given on a scale from "1 = strongly below my expectations", "5 = exactly what I expect of him", and "9 = strongly exceeds my expectations".

On the next page, all participants received the following question:

> **(APPROVAL)**
> Next, we would like to know:
>
> *How strongly to you disapprove or approve when people do something [thin or thick term]?*

Approval ratings are measured on a 9-point Likert scale anchored at 1 = "strongly disapprove", 5 = "neither approve not disapprove", and 9 = "strongly approve".

## RESULTS




In this Experiment, we are not interested in the absolute values, but in how much they differ from the neutral midpoint of "5".

For (Expectation), positive thick concepts get expectation ratings closer to the midpoint (Δ=1.09) than negative thick concepts (Δ=2.08). We find the same effect for thin concepts as well (Positive: Δ=1.16; Negative: Δ=2.21). This effect also found at the level of individual pairs of items (e.g. generous vs selfish).

For (Approval), we find no such difference between positive and negative thick concepts (Positive: Δ=2.9; Negative: Δ=2.51); and neither do we find it for thin concepts (Positive: Δ=2.47; Negative: Δ=2.53).

## INTERPRETATION OF EXPERIMENT 5

The results confirm our hypothesis that the normal person is a slightly virtuous person. Behaving in a positive way is more expected than behaving in a negative way. These results provide support for our explanation of the Polarity Effect.

We also find that in general, doing good things is approved of just as much as doing bad things is disapproved of. Thus, people are aware that doing good things is something good, but they are nevertheless unwilling to evaluate the agent positively for actions that are expectable.