Affective modulation of the weighting function

Victor Møller Poulsen, Studie Nr.: 201707639

January 10, 2021

1 Description

Code for Simulation & Figures available at: https://github.com/victor-m-p/BayesianDecisionWeights

Both Expected-utility theory (EUT) and prospect theory (PT) posit that humans maximize some version of utility. The theories get there by a combination of two functions (Rottenstreich & Hsee, 2001). A value function v transforms objective value to subjective utility, and a weighting function w distorts probabilities (Gonzalez & Wu, 1999; Rottenstreich & Hsee, 2001). Expected-utility and prospect theory combine these two paramters in the simplest way possible (Rottenstreich & Hsee, 2001)

$$\sum w(p_i)v(i),$$

where p stands for probability and i stands for the i^{th} gamble.
In EUT the weight function w is the identity $w(p) = p$ assuming that people do not distort probabilities (Rottenstreich & Hsee, 2001). In expected-utility the value functin v is proposed to reflect how people feel about end states. This assumes that people should take into account their current state (e.g. of wealth) when evaluating outcomes (Newell et al., 2015).

With regards to both v and w PT (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) advances theorizing, and better reflects empirical results. It is arguably the main model of human decision making (Newell et al., 2015).

PT advances theorizing with regards to v by positing that losses and gains are evaluated as changes in wealth rather than with regards to end states. This means that rich and poor people should show relatively similar behaviour, because both evaluate outcomes based on a neutral starting point (Newell et al., 2015). This leads us to the familiar (non-linear) S-shaped value function v proposed in PT (Kahneman & Tversky, 1979). The value function v is concave for the gains domain and convex for losses. This reflects the fact that small changes in outcome are relatively overweighted. I.e., a monetary increase from $0 - 100\$$ has greater utility than an increase from $1000 - 1100\$$. The same is the case for the domain of losses. Another key stylistic of the value function is that it is steeper for the loss domain, showing loss aversion.

PT advances theorizing with regards to w by proposing a non-linear probability distortion (Kahneman & Tversky, 1979). w is stylized as being reverse S-shaped, meaning that it is concave for low probabilities and convex for high probabilities (Gonzalez & Wu, 1999). This means that people underweight changes in probability in the middle of the spectrum (e.g. $[0.2 - 0.8]$) while overweighting changes in probability close to the end-points (e.g. $[0.0 - 0.2], [0.8 - 1.0]$). These general characteristics of the weighting function are empirically well documented (Tversky & Kahneman, 1992; Wu & Gonzalez, 1996).

1.1 Prior work

There is evidence to support the notion that the affect of outcomes modulates the parameters of both v (Hsee & Rottenstreich, 2004) and w (Rottenstreich & Hsee, 2001).

The S-shape of the weighting function w appears to be more pronounced for high-affect than low-affect outcomes under uncertainty (Rottenstreich & Hsee, 2001). They investigated how much participants were willing to pay for two coupons (worth the same) at different levels of probability. The first item was a trip to Europe (high-affect) while the second item was tuition covering (low-affect). They were able to show a preference reversal in which the high-affect outcome was preferred for low probability (1%) whereas a low-affect outcome was preferred for high probability (99%). If the results are solid, they suggest that people distort probabilities more for high-affect as compared to low-affect outcomes.

For gains, the value function v becomes more concave for high-affect as opposed to low-affect outcomes (Hsee & Rottenstreich, 2004). Hsee and Rottenstreich (2004) showed this by priming participants to evaluate outcomes either based on calculation or based on feeling. Their results clearly suggest a modulatory effect of affect-richness. Taken together the results suggest a consistent picture of modulation of both the value function v and the weight function w. This line of evidence has been pursued elsewhere (Mukherjee, 2010, 2011) with the idea of modeling decision making as an interaction between an affective system and a deliberative system.

1.2 Focus and parameterization

In this article we focus exclusively on the weighting function w while ignoring both the value function v and the combination of the two functions. We also restrict ourselves to the gains domain. In Rottenstreich and Hsee (2001) they propose that the affective modulation can be estimated as an affect paramter a in the form:

$$w(p) = \frac{p^{1-a}}{p^{1-a} + (1-p)^{1-a}}.$$

where $a \in [0, 1]$ and larger a values indicate greater affect and more curvature (Rottenstreich & Hsee, 2001). The issue with this one-parameter formulation is that it does not account for the fact that people generally show low *elevation*. What I mean by that is that the empirically observed weighting function w typically

crosses the diagonal line at around 0.3 rather than 0.5 (Gonzalez & Wu, 1999). This can be interpreted as people generally being pessimistic (i.e. 50% probability is evaluated as being worth less than 50% of the outcome). The one-parameter formulation fixes this point at 0.5, (i.e. $w(.5) = 0.5$), which can be seen from figure 1.

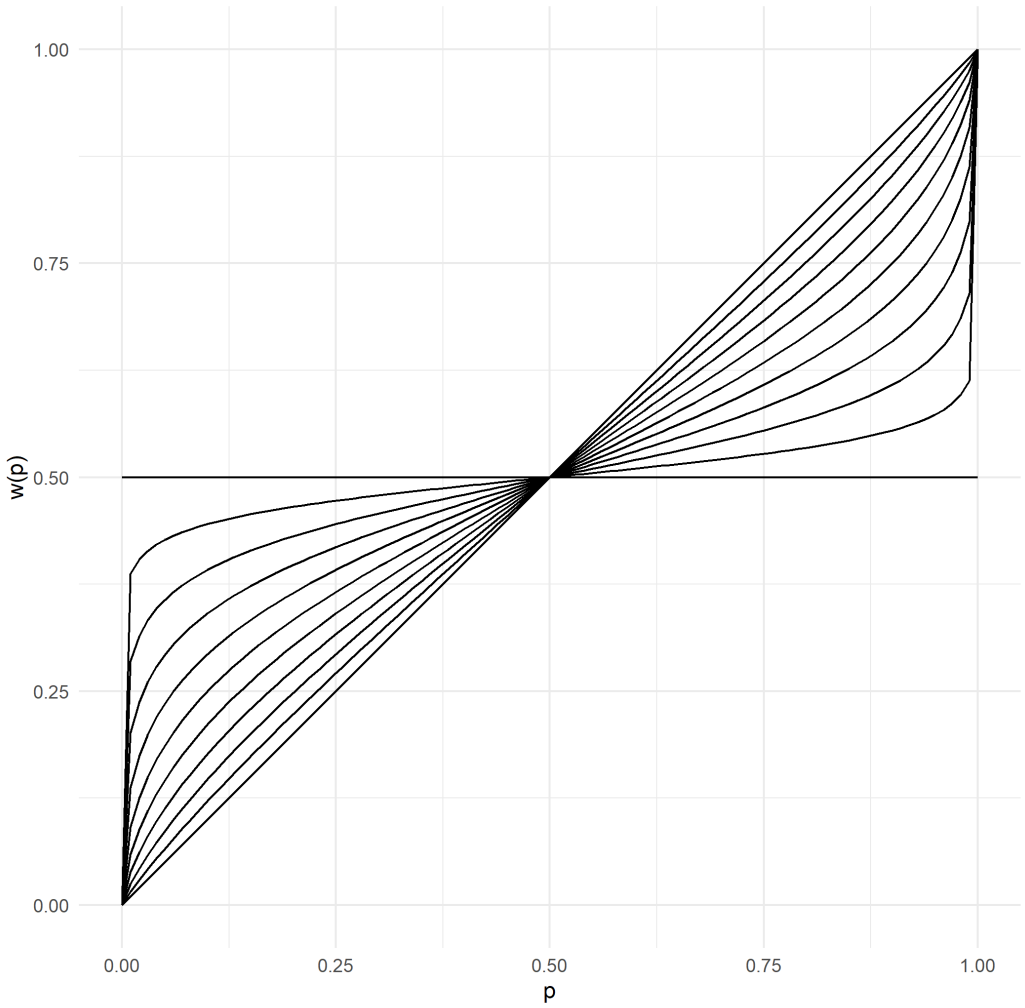

Figure 1: Data simulated from the model $w(p) = \frac{p^{1-a}}{p^{1-a}+(1-p)^{1-a}}$ with $a \in [0, 1]$. Diagonal line has $a = 0$, and the horizontal line has $a = 1$. Intermediate curves are generated for 0.2 increments of a. All values beside $a = 0$ show a probability distortion as compared to the objective probability. Note that all curves meet at $w(p) = 0.5, p = 0.5$. This is not empirically supported.

Instead of using the parameterization proposed in Rottenstreich and Hsee (2001) this paper will use the parameterization of w proposed in Gonzalez and Wu (1999).

They parameterize w with two parameters; δ and γ.

The δ parameter will vary based on *elevation* (intercept) (Gonzalez & Wu, 1999), which here simply refers to the overall perceived attractiveness of outcomes under uncertainty.

The γ parameter will vary based on *curvature* (slope) (Gonzalez & Wu, 1999) and is what we are primarily interested in for our purposes. It follows as a direct prediction from Rottenstreich and Hsee (2001) that the curvature (γ) should be modulated by changes in the affective level of outcomes.

$p(w) = p$ for $\gamma = 1, \delta = 1$ with this parameterization. Higher δ corresponds to higher elevation, and higher γ corresponds to *less* curvature (unintuitively, and as opposed to the α parameter in the one-parameter w function). See figure 2 for an illustration of how the δ and γ parameters independently modulate different aspects of the weighting function w.

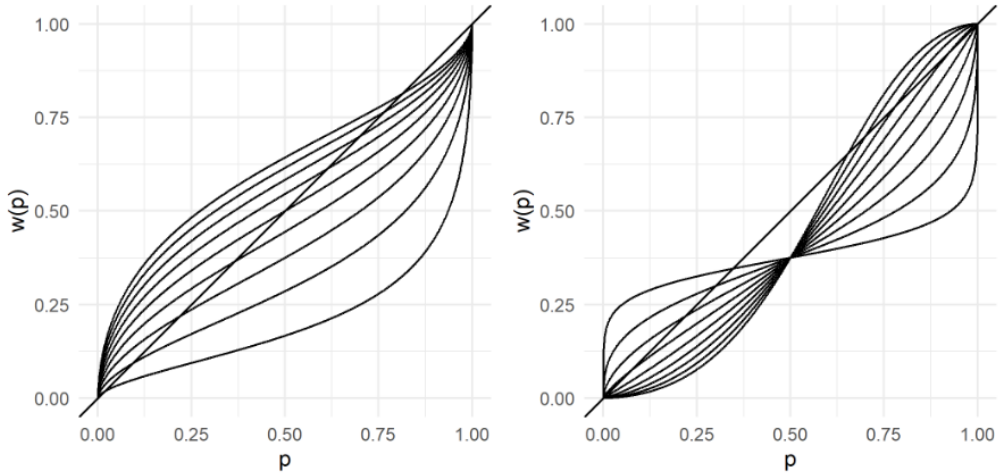

Figure 2: Data simulated from the model $w(p) = \frac{\delta \cdot p^\gamma}{\delta \cdot p^\gamma + (1-p)^\gamma}$ similarly to figure 4 of Gonzalez and Wu (1999). On the left: γ fixed at 0.6 and δ varied between 0.2 and 1.8. On the right: δ fixed at 0.6 and γ varied between 0.2 and 1.8. Shows that γ controls curvature and δ controls elevation. The identity function $w(p) = p$ is achieved for $\delta = 1, \gamma = 1$. Note.. gamma low has the opposite interpretation as compared to rottenstreich?

The model proposed in Gonzalez and Wu (1999) is:

$$\log \frac{w(p)}{1 - w(p)} = \gamma \log \frac{p}{1 - p} + \tau.$$

where solving for $w(p)$ and setting $\delta = \exp(\tau)$ gives us

$$w(p) = \frac{\delta \cdot p^\gamma}{\delta \cdot p^\gamma + (1 - p)^\gamma}.$$

1.3 Methodology

Two studies are proposed to properly test the robustness of affect-level on the curvature (γ) of the weight function w.

In the first study, subjects will be asked to rate the affect-richness of 10 different items. All outcomes consist of coupons redeemable for various items, all worth $\$500$. The 10 items are designed to cover the full spectrum from affect-rich to affect-poor.

Example of expected high-affect item:
"If you won a $\$500$ coupon redeemable for a vacation abroad with a friend/partner how emotionally affected would you be?"

Example of expected low-affect item:
"If you won a $\$500$ coupon redeemable for insurance covering how emotionally affected would you be?"

For the full list of items see *Appendix A*. Participants will indicate how affect-rich each outcome is with a slider. Participants will see "not affected at all" (left), "somewhat affected" (middle) and "very affected" (right). We will receive continuous ratings from 0 (affect poor) to 1 (affect rich). A mean affect rating across participants for each item will rank them from least affective to most affective. Three items are then selected: The least affective item (A), the most affective item (C) and the item in between these two extremes (B) which separate them best (follow up). Welch two sample t-tests between A and B, and between B and C will indicate whether they differ significantly.

In the second study, subjects will be presented with the three items (A, B, C) which have been validated for affect-richness in the prior study. All subjects rate items

in all three conditions, making the study a within-subject design. The formulation around the items is that of a gamble. The formulation is the same for all items:

"You can buy a lottery ticket with an $[x]$ percent chance of winning a $\$500$ coupon redeemable for $[y]$ with a $[1 - x]$ percent chance of winning nothing. How much are you willing to pay for the lottery ticket?"

The three selected items are inserted as $[y]$ and 50 different probability levels: $x = 0.01, 0.03, \ldots, 0.99$ will be inserted as $[x]$ and the negation $[1 - x]$. With all possible combinations, this means that all participants will rate the items of the 3 conditions at 50 different levels of certainty each. As in experiment 1 participants will rate with a slider. This time ranging from $\$0$ to $\$500$ as it is neither logical to assign a value below $\$0$ or above $\$500$ to any of the gambles. The approach is somewhat different from Gonzalez and Wu (1999) but ultimately we estimate the same thing that they do; participants' certainty equivalence (CE). This simply is the amount of money they think that the gamble is worth.

Note that we are not directly measuring either δ or γ. What we do measure is the dependent variable $w(p)$ and the independent variables p for three conditions (ranked based on affect-richness).

In order to infer the unmeasured paramters a bayesian (non)linear mixed effects model is proposed. The model is fitted in R (R Core Team, 2020) with the $brms$

package (Bürkner, 2018). Here we can specify the previously mentioned formula:

$$w(p) \sim \frac{\exp(\tau) \cdot p^{\gamma}}{\exp(\tau) \cdot p^{\gamma} + (1-p)^{\gamma}}.$$

It is extremely important that we specify that we want to measure $\exp(\tau)$ rather than δ as this tells the models that this parameter must be positive and thus limits the flexibility of the model in an appropriate way. δ can be inferred afterwards by taking $\exp \tau$. Additionally, we have to specify that the model should be nonlinear - as we believe that the weighting function is non-linear. We can further specify that we would like to estimate specifically the value for τ and γ with random intercepts (partial pooling) for participants (ID) and with item (condition) as a main effect.

$$\tau \sim 0 + item + (1|ID),$$

$$\gamma \sim 0 + item + (1|ID).$$

As mentioned, estimated τ is converted to δ by exponentiating the estimated τ value after model fitting.

Results are reported as $.66$ and $.95$ credibility intervals for the δ and γ distributions for each condition. Posterior samples are drawn from the distributions, allowing for a nice visualization of effects.

2 Hypotheses

2.1 Study 1

Hypothesis 1: As explained earlier, three outcomes are selected from the 10 investigated outcomes. The most affective (C), the least affective (A) and the question which best separate the two (B). It is hypothesized that pairwise t-tests (Welch Two Sample) between A and B and between B and C will result in significant differences. This has to be achieved before conducting study 2, as that study relies on this effect. As such, if this effect is not achieved, another study should be conducted to validate questions before proceeding with study 2. With that said however, it does appear reasonable that the 10 different questions should cover the spectrum of affect-richness pretty well (see *Appendix A*) and as such it is expected that three items which differ significantly can be extracted.

2.2 Study 2

Hypothesis 1: A directional effect is predicted for the γ parameter of the function:

$$w(p) = \frac{\delta \cdot p^{\gamma}}{\delta \cdot p^{\gamma} + (1-p)^{\gamma}}.$$

Recall that study 2 uses the three items from study 1 as three conditions. We will refer to these as conditions low-affect A, medium-affect B, and high-affect C. It hypothesized that posterior credibility intervals ($.95$) for the γ parameter will not overlap between conditions, and that γ highest for A, lower for B and lowest for C

(recall that high affect is predicted to result in high curvature, which is offered by low γ). This effect would replicate and seriously strengthen the results of Rottenstreich and Hsee (2001). A weaker replication would consist of credibility intervals ($.66$) showing the same effect. This would still be an interesting result, but would indicate a weaker and less reliable effect.

Hypothesis 2: Additionally, Gonzalez and Wu (1999) report population $\gamma = 0.44$ (median). They use monetary gambles (low-affect), and as such it is hypothesized that $\gamma = 0.44$ will be within the $.95$ credibility interval of our low-affect condition (A). This would serve as a replication of the findings of Gonzalez and Wu (1999). Again, a less convincing, but interesting result would be to observe $\gamma = 0.44$ within the 0.66 credibility interval. Minimally interesting effects for γ are shown in figure 3, where γ values are $A = 0.44, B = 0.34, C = 0.24$.

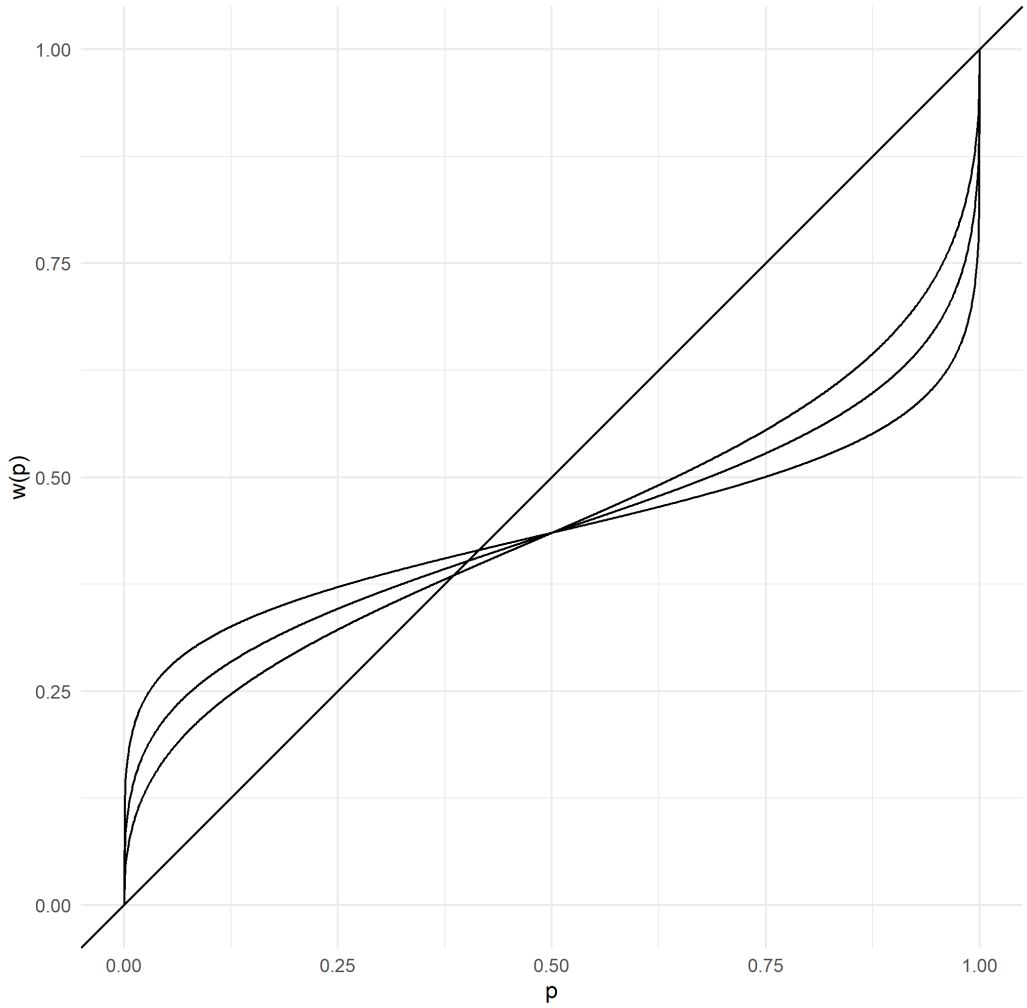

Figure 3: Three curves shown, all with $\delta = 0.77$ as reported in Gonzalez and Wu (1999). γ levels $0.24, 0.34, 0.44$. The least curved line corresponds to $\gamma = 0.44$, as reported in Gonzalez and Wu (1999). For high-affect items γ should be lower, and as such we suggest 0.24 (for high affect) and 0.34 (for medium affect) as minimally interesting effects to detect

.

Hypothesis 3: No direction of effect for the δ parameter by condition is hypothesized. The δ parameter is not of interest to the main hypothesis (replicating and

extending Rottenstreich and Hsee (2001)) and is mainly included in the analysis in order to control for elevation and properly estimate γ. If the three items differ in perceived overall value the δ parameter should capture this. This means that our γ distributions should still be interpretable even if the δ parameter differs by condition. The same analysis pipeline will be applied to δ as for γ (i.e. credibility intervals estimated) but as suggested, it is not clear whether an effect would be interesting. The δ parameter is expected to have a value close to .77 which is the population median found for this parameter is Gonzalez and Wu (1999).

2.3 Simulation

In order to test the pipeline for the bayesian analysis, data simulation was conducted. Unfortunately, Gonzalez and Wu (1999) does not exactly report the values (i.e. distributional properties of τ and γ) that we need to generate data consistent with what they gathered. As such, it does not make sense to calculate power based on our simulations, and the simulation serves only the purpose of making clear how analysis on eventual data will be conducted.

Data is generated for 50 probability levels, $p = 0.01, 0.03, \ldots, 0.99$ crossed with 3 conditions, corresponding to the actual data that will be collected. Data is generated for 30 simulated subjects (ID).

Note that standard deviations vary between γ and δ, and between population level and individual variation. This qualitatively follows the results of Gonzalez and

Wu (1999). Data is generated as a distribution of γ and δ for each condition. We generate 30 values (i) for each, corresponding to the number of participants. As we do not hypothesize that δ is modulated by condition this can simply be generated as once.

$$\begin{aligned} \gamma_{A_i} &\sim norm(n = 30, m = 0.24, sd = 0.1) \\ \gamma_{B_i} &\sim norm(n = 30, m = 0.34, sd = 0.1) \\ \gamma_{C_i} &\sim norm(n = 30, m = 0.44, sd = 0.1) \\ \delta_i &\sim norm(n = 90, m = 0.77, sd = 0.2) \end{aligned} \tag{1}$$

Based on these γ and δ values for participants per condition, we generate the final γ and δ values by adding individual noise for each probability level (j)

$$\begin{aligned} \gamma_{A_{ij}} &\sim norm(n = 50, m = \gamma_{A_i}, sd = 0.1) \\ \gamma_{B_{ij}} &\sim norm(n = 50, m = \gamma_{B_i}, sd = 0.1) \\ \gamma_{C_{ij}} &\sim norm(n = 50, m = \gamma_{C_i}, sd = 0.1) \\ \delta_{ij} &\sim norm(n = 150, m = \delta_i, sd = 0.3) \end{aligned} \tag{2}$$

As such, each condition will contain two levels of noise around a true signal. The simulated data, and the best fit $w(p)$ curves are shown in figure 4. As can be seen, the simulated data shows the expected pattern, where low values of γ exhibit more curvature. The preference reversal shown in Rottenstreich and Hsee (2001) is also

seen in the plot.

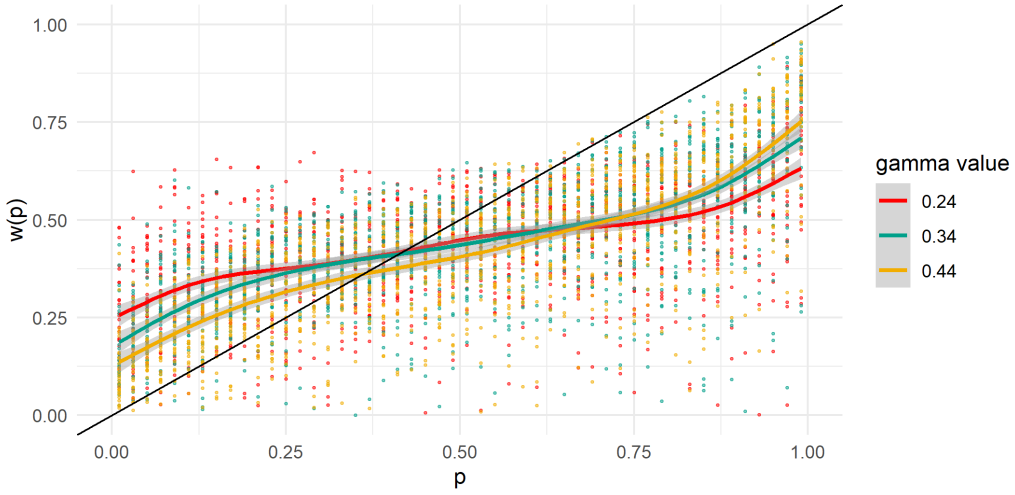

Figure 4: Plot of simulated data in three conditions ($A : \gamma = 0.24$, $B : \gamma = 0.34$, $C : \gamma = 0.44$). In all conditions the true population mean of $\delta = 0.77$. Shows the preference reversal observed in Rottenstreich and Hsee (2001). Note that the yellow curve corresponds roughly to what was found in Gonzalez and Wu (1999). The population effect is a weak, but true signal, which is what we expect from the real data.

Next, the model described earlier is fitted to the data. Regularizing riors are specified:

$$\tau \sim normal(0, 1)$$

$$\gamma \sim normal(0.3, 0.5)$$

Reflecting our knowledge of reasonable values for these parameters. The same priors will be used for modeling the actual data. Various characteristics, such as R-hat and pp_checks (from *brms*) indicate a good model fit.

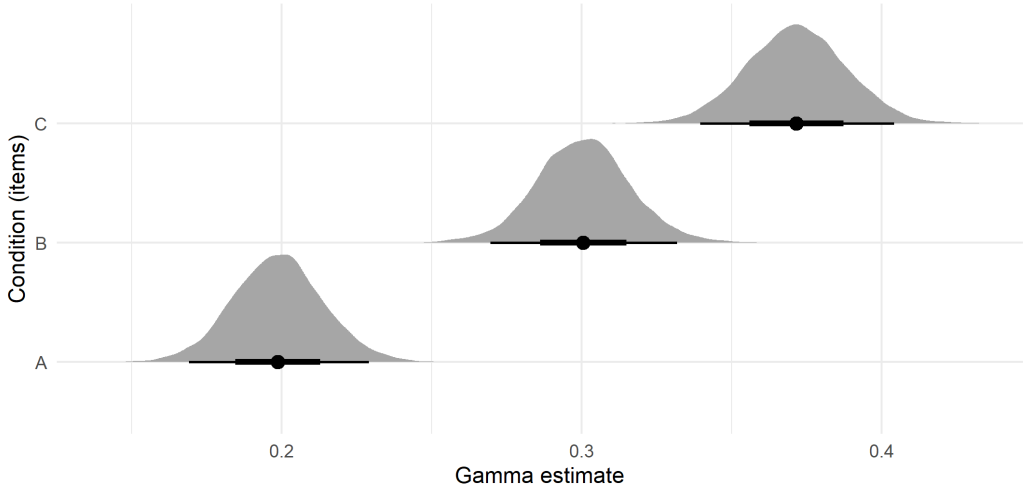

Figure 5: showing the estimated γ population distributions. The thick black line is .66 credibility intervals whereas the thin black line is 0.95 credibility intervals. Note that the conditions are ordered as expected with A having the lowest γ and C having the highest γ. Note that the 0.95 credibility intervals do not overlap.

We extract credibility intervals with regards to γ and δ distributions for each condition (A, B, C). With the simulated data, we note that the model is capable of recovering these unobserved (unmeasured) parameters, and that for γ the $.95$ credibility intervals do not overlap between conditions (see figure 5). The estimates and credibility intervals are, $\gamma_A = 0.37, CI : [0.34, 0.40]$, $\gamma_B = 0.30, CI : [0.27, 0.33]$ and $\gamma_C = 0.20,\ CI : [0.17, 0.23]$. This slightly underestimates the true effect, which we know because we simulated the data. However, it is reasonably close. This shows that with 30 participants, and 50 probability levels crossed with 3 conditions it is possible to detect the effect that we specified. This of course assumes specific distributional characteristics and noise-levels that we cannot know in ad-

vance. It does however, show that the model works as intended.

3 Design Plan

Study type: Study 1 might be characterized as an observational study, since it does not really have an experimental manipulation. It resembles a survey of questions (e.g. the 10 outcomes). Study 2 is an experiment using a within-subjects design, in which all participants participate in all three conditions (A, B, C). This is important because within-subject designs have better power to detect effects than between-subject designs (Charness et al., 2012). Power is a primary concern because effects are likely to be small, and variance is likely to be high (Gonzalez & Wu, 1999). A between-subjects design is not necessary in our case, because we do not induce an effect by priming, as in e.g. Hsee and Rottenstreich (2004).

Blinding: No blinding is involved in this study.

3.1 Study Design

Study 1: All subjects will rate all 10 items (see Appendix 1) as to the level of affect they feel with regards to them.

Study 2: All participants indicate their certainty equivalence (CE) for all certainty levels $p = 0.01, 0.03, \ldots, 0.99 (n = 50)$ and in all three conditions (A, B, C). This results in 150 observations per participant, and 50 observations per participant for each condition.

4 Sampling Plan

Existing Data: Registration prior to creation of data. Data from simulation does exist.

Data collection procedures: Participants will be recruited through online channels (e.g. facebook, student groups, etc.). Participants must be at least 18 years old to participate. In the first experiments subjects will be payed 30 DKK for agreeing to participate in an approx. 10 minute online survey. In the second experiment subjects will be payed 150 DKK for agreeing to participate in an approx. 60 minute online experiment

Sample size:

Study 1: 30 participants are recruited.

Study 2: 30 participants are recruited.

Sample size rationale:

Study 1: Data was simulated to estimate the approximate sample size needed for a Welch two sample t-test to dissociate the three most different items (questions). Plotting and common sense was used to arrive at best guesses for reasonable values. Three distributions were generated (assumed normal, although that is not generally true for slider data):

$$
\begin{aligned}
item_A &\sim norm(0.2, 0.4) \\
item_B &\sim norm(0.5, 0.4) \\
item_C &\sim norm(0.8, 0.4)
\end{aligned} \tag{3}
$$

With $n = 30$ participants we have extremely high power to detect these effects. However, as this is a cheap experiment to run, and the results are critical for the second study (it is important that the three best items are used as the conditions in experiment 2) this is deemed to be reasonable.

Study 2: Choice of sample size is naturally related to power. Typically, $.8$ power is considered reasonable (Cohen, 1992), although this is just convention. Power reflects the ability to detect a effect and is influenced by effect size and number of participants. Unfortunately, a power simulation was not possible to carry out since reasonable estimates for the distributions (and effect sizes) are not present. The simulation presented earlier likely underestimates individual variation, which will lessen power. The best comparison that we have is Gonzalez and Wu (1999) who estimate both parameters of the value function v as well parameters γ and δ of the weighting function w. They do so with 10 participants and collect data with respect to 15 gambles crossed with 11 probability levels (Gonzalez & Wu, 1999). The fact that they are able to reasonable recover the unobserved parameters (of v as well) with a sample size of only 10 participants suggests that it is not so much the number of participants, but rather the number of trials for each participant that

is important. More noise is observed within participants than between (Gonzalez & Wu, 1999). Based on their results and on the simulation carried out, it is argued that 30 participants should probably give us reasonable power to detect a minimally interesting effect.

5 Variables

5.1 Manipulated variables

Study 1: No manipulated variables. The study is an observational survey, where participants use a slider to indicate the affect-richness of unordered outcomes.

Study 2: 50 levels of uncertainty are crossed with 3 conditions (different gambles). These are the two manipulated variables of study 2.

5.2 Measured variables

Study 1: The single outcome variable will be the rating of affect level. This will be measured on a scale of $0 - 1$. Participants will rate this using a slider (and will not see the same scale that we measure).

Study 2: The single outcome variable is the price that subjects indicate that they are willing to pay for a ticket in a lottery. This measures the certainty equivalence (CE) of participants, and can be thought of as $w(p)$. This will be measured on a

scale of $0 - 500$ dollars using a slider. The max is 500 dollars since the lottery tickets by definition cannot be worth more than this.

5.3 Indices

No indices are used.

6 Analysis Plan

All analysis is performed in the programming language *R* (R Core Team, 2020) using *Rstudio* IDE (RStudio Team, 2020). A key package used for bayesian model fitting is *brms* (Bürkner, 2018).

Study 1: The affect ratings will be ordered based on group-level means. The three questions that best separate the are validated as being statistically significant with a Welch two sample t-test.

Study 2: A bayesian generalized nonlinear mixed effects model is fit to the data using the *R* package *brms* (Bürkner, 2018). This is done to estimate the unobserved parameters δ and γ from the independent variables (1) probability level and (2) category, and their relation to the dependent variable $w(p)$ which is the observed certainty equivalence (CE). Regularizing priors are specified for both γ and δ as described in the "simulation" section.

7 Discussion & Future Work

The two-part study presented is important for several reasons. Firstly, it is part of a movement to formalize psychology and decision making (DM) which the author believes is both important and generally done too rarely. The effect that this study tries to replicate (**rottenstreich2001**) is an interesting study. However, it bases the conclusion of a modulation of the weighting function by affect based on a measurement of end-points (1% and 99%) only. As such, it is rather stylized; i.e. they are not able to estimate the modulation of the actual parameters of the weighting function. It is surprising that no-one (to the knowledge of me) has actually followed the suggestion of (**rottenstreich2001**) in extending their stylized effect and testing it across more than two items, and across enough uncertainty levels to estimate parameters of the weighting function w.

The present study also attempts to facilitate cumulative science more generally, by providing all code, simulation and eventual data at: https://github.com/victor-m-p/BayesianDecisionWeights. By making the code and data accessible future experimenters can use the knowledge gained in this experiment to motivate stronger priors - effectively pooling information across studies (by using our posteriors as priors). Sadly, most of the research in this area is carried out in a frequentist framework, and often code and data is not accessible which makes it impossible to properly integrate previous work.

8 Appendix

Appendix A

As all items in study 1 follow the same template:

"If you won a $\$500$ coupon redeemable for/at $[x]$ how emotionally affected would you be?"

The 10 proposed $[x]$ outcomes are:

- for a vacation abroad with a friend/partner.
- at a local shopping mall.
- for donation to a charity of your choice.
- for a cultural experience in your city.
- for insurance covering.
- for investing in the stock market.
- for job training.
- for $\$500$.
- for spending on a present to someone you love.
- at your favorite restaurant.