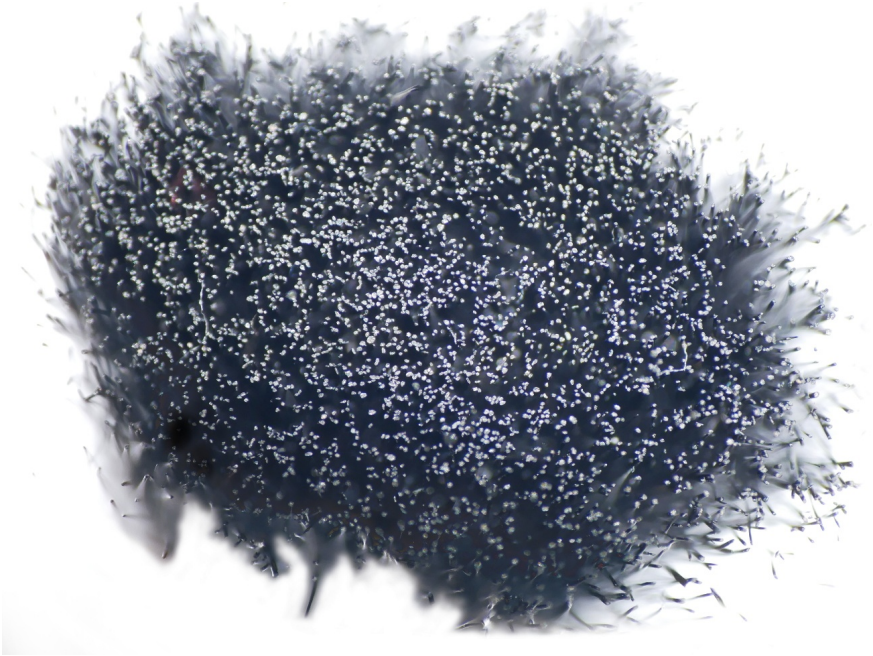
Umeå International School of Public Health
Umeå University



Department of Public Health and Clinical Medicine
Epidemiology and Global Health

In Pursuit of Weights For CALY

Exploring Methods for Measuring And Calculating Capability Adjusted Life Year Weights

Author:

Kaspar Meili

Supervisor/s:

Lars Lindholm

Nr: 36/2017

Master thesis, 15 credits

Master's programme in Public Health, 120 credits

Acknowledgment

I would like to express my gratitude to the following persons for helping me along the way.

Mazen Baroudi, Lisa Haryson, Anne-Karing Hurtig, Lars Lindholm, Anna Månsdotter, Lennart Nilson, Fredrik Norström, Anna Steling, Anni-Maria Pulkki-Brännström, Ka Chun Tsang

External Collaborator: Professor Lars Lindholm, Department of Public Health and Clinical Medicine, Umeå University

Abstract

Background: Meusers that capture outcomes more broadly than currently used paradigms such as QALYs which is geared towards health are necessary to better inform decision making because effects of interventions often apply beyond a single area such as health. Capability Adjusted Life Years (CALYs) is a new Swedish based extra-welfarist approach based on Sen's capabilities. Capabilities refer to the set off possibilities that an individual can choose to be or can do. Similar in application to Quality Adjusted Life Years (QALYs), the measure incorporates the weighing of lifetime. The aim is to investigate the survey and statistical procedures for the calculation of these weights.

Methods: Weights for 5 attributes corresponding to the capability areas economic ressources, health, education, occupation and social relations with each 3 levels, encoding 243 possible states, were evaluated using a web survey that featured paired comparison and time trade-off questions in different framings and question types. Invitations to the survey were administered via post to a sample of 2000 Swedish permanent residents. The statistical methods included probit regression for the estimation of coefficients corresponding to the attribute levels, and interval regression to calculate weights for 5 states that were used to reanchor the probit weights on the 0 to 1 scale with a linear transformation.

Results: A response short rate of 11.7% resulted. The respondents were not representative of the Swedish population. The estimation of probit weights yielded significant estimates that were consistent with the associated level. Trade-off estimates calculated with interval regressions differed depending on the framing (relative versus absolute life expectancy) and question type (ranking versus discrete choice) and was partially inconsistent. Similarly, the distributions of the reanchored weights varied between framing and question types.

Conclusion: The sampling procedure used is inadequate to achieve a representative sample, especially given the low response rate. The web survey format performed reasonably well. More than 5 states should be estimated on the 0 to 1 scale to increase the quality of the reanchoring, and the estimates for these state need to be more precise and consistent. Careful consideration of the desired characteristics of the distribution of reanchored weights is required when choosing between framings and types of trade-off questions.

Table of Contents

Abbreviations

ALE: Absolute Life Expectancy

CALY: Capability Adjusted Life Year

DALY: Disability Adjusted Life Year

DC: Discrete Choice

QALY: Quality Adjusted Life Year

RLE: Relative Life Expectancy

TTO: Time Trade-off

1 Introduction

1.1 Background

Progress has a lot of faces and one may argue welfare, aggregated on a societal level, is an important component. The question of how to define and judge welfare has been discussed long and intensively throughout human history (Sedlacek and Havel, 2013). As an important consequence, an efficient distribution of resources to improve welfare depends on this discussion (Månsdotter et al., 2017b). Furthermore, adequate measures are necessary to describe the current state of welfare and to inform decisions regarding efficient resource distribution, for example in the health care context. As such, they should reflect the prevalent understanding of welfare in the population to be acceptable, they need to enable effective decisions that improve welfare, and they should be well and easily operationalizable.

Welfarism is one such paradigm that offers a definition of welfare. Rooted in utilitarianism, the judgment criteria is confined to the maximization of individual utilities and as such consequentialist. Only affected individuals themselves are a valid source of information to judge the outcomes. Comparing interpersonal utilities or judging the desirability of distributions of utility comes with difficulties, because it is unclear how such criteria can be derived from individual based preferences alone (Brouwer et al., 2008; Drummond et al., 2005, p. 217).

A health economic example of a decision rule according to welfarism would be the maximum net benefit option in a cost-benefit analysis, where the willingness to pay is a proxy for individual utility and deducted solely by individual-specific information, for example with revealed preferences (Brouwer et al., 2008; Drummond et al., 2005, p. 16)

Another paradigm is extra-welfarism. In contrast to welfarism, extra-welfarism shifts away from the focus on utility and allows other – extra - information to be used in the judgment, which does not necessarily have to originate from the affected individuals. Specifically, not only individual-based information, but for example also information about the distribution of outcome measures, or the judgment of experts, and measures other than utility are allowed to be used to derive the level of welfare. Extra-welfarism also enables to trade-off between different measures and between individuals (Brouwer et al., 2008).

Cost-utility analysis, such as cost per quality adjusted life year or cost per disability adjusted life year, are associated to extra-welfarism, because the “utility” measures, namely quality adjusted life years (QALY)(Drummond et al., 2005, pp. 173–175), for example in the form of EQ-5D (Brooks and Group, 1996), or disability adjusted life years (DALY)(Drummond et al., 2005, pp. 187–188), do not represent individual utility alone but rather health as an end in its own, that can be compared across different persons. Furthermore, the weights used to aggregate QALYs or DALYs represent a form of publicly pooled preferences that are imposed on the individuals, as opposed to more individualistic utilities that are difficult to use for interpersonal comparisons (Brouwer et al., 2008; Drummond et al., 2005, p. 188; Lopez and Murray, 1998, pp. 7–13). But, extra welfarism does not exclude the use of individual-preference derived measures such as the QALY- or DALY weights(Brouwer et al., 2008).

Sen’s capabilities approach another extra-welfarist concept (Brouwer et al., 2008; Sen, 1985). Sen argues that welfareis given by the possibility of achieving a certain outcome, not the

actual outcome itself. He uses the term “functions” to describe possible uses of commodities, that is what an individual can be or can do (Sen, 1985, pp. 10–11). For example, an individual can be a fisherman or a carpenter, or one may chose to climb a mountain or to knit a quilt. Capabilities are all possible functions, given restrictions in commodities and personal characteristics, that the individual can choose from (Sen, 1985, p. 13). According to the capabilities approach, neither the endowment in commodities nor the experienced happiness, that is, the utility, are appropriate criteria to judge welfare. Instead, welfare should be judged based on the capabilities that an individual has, because sufficient capabilities enable individuals to freely chose a desirable life (Sen, 1985, p. 28). Furthermore, the amount of commodities or wealth alone does not directly translate into welfare (Sen, 1985, p. 28). Neither is utility-maximization synonymous with achieving a high well-being, as for example a slave may be happy when his living circumstances are sufficiently good, even if the individual judges its state of being a slave not to be desirable(Sen, 1985, p. 21).

1.2 Capability Adjusted Life Years (CALYs)

Recently, a Swedish based initiative put forward a suggestion that is based on Sen’s capabilities: Capability adjusted life years (CALYs) (Månsdotter et al., 2017a). This proposal comes from a public health and health economics related perspective but aims to offer an evaluation tool that captures more wholesomely the impact of decisions on a society level. Thus, in addition to the aforementioned arguments towards a capability oriented approach, the initiators argue that interventions to improve health often affect other areas of life and that policy-relevant evaluations should account for these (Månsdotter et al., 2017a). For example, improved general education can be an effective way of tackling a number of health related issues but other consequences of improved education may be even more desirable than improved health alone, such as an improved ability towards self-sufficiency for women. Wile QALYs in the form of EQ-5D(Brooks and Group, 1996) and DALYs could adequately cover health effects, other effects e.g. an increased earnings potential, are arguably not directly capturable. A more wholesome approach to evaluate policy consequences has the potential to increase the quality of policy-relevant decisions and help to anticipate unwanted side effects regardless of the field that the policy in question targets.

CALYs is such an approach. It is measure that aims to quantify the amount of time lived but also takes into consideration how this time is lived (Månsdotter et al., 2017b, 2017a). It consists of two components: First, the amount of time (normally in years) and second an adjustment in the form of a multiplier, called capability weight that takes into account the quality of a life situation. Each unit of time is multiplied with situation specific capability weight and added up. The resulting sum represents a judgment criteria for the evaluated lifetime; a higher value indicates a higher capability. Subsequently, CALYs can also be summed over groups of individuals.

CALYs also differs from QALYs on another theoretical basis, as the approach is intended to clearly contrast welfarism in comparison to QALYs and thus aims to be preferably associated to non-welfarism instead of extra welfarism (Månsdotter et al., 2017a).

The adjustment weights are between 0 and 1, wiht 1 representing the highest achievable level of capabilities and 0 representing a capability level of no capabilities. However, the interpretation of 0 capability may not be as straight forward as in the health context where the interpretation of the 0 weight as equivalent to death is plausible (Drummond et al., 2005,

pp. 174–175).

This concept is related and inspired by other composite measures that weight a person-time outcome, such as QALYs and DALYs . For example, CALYs could be used in future in the denominator in cost-effectiveness analyses , i.e. cost per CALY, to base funding decisions on minimal costs per CALY or thresholds. That is, similar to how QALYs are used, by calculating cost per CALY for different options.

Given the high number of areas of life, for example wealth, culture, social relations and health that can be relevant for welfare, selecting which to include in an operationalized CALY measure is a challenging task. Different capabilities may be relevant to a varying degree for different individuals or in different countries, as for example in Sweden educational and political rights may be fully developed whereas political rights are less developed in other countries (Månsdotter et al., 2017a). An influential proposal is Nussbaum's (2000, pp. 78–80) list of capabilities.

Furthermore, the chosen capabilities should reflect the values of a society, should be relevant for policy, feasible to be coherently established in a given context (Månsdotter et al., 2017b; SOU, 2015), and be able to explain capability inequalities in the relevant context, e.g. the Swedish one (Månsdotter et al., 2017a). A pilot study among Swedish researches has been conduced to rank different capability areas (Månsdotter et al., 2017b).

The CALY initiative aims to base the choice of capabilities on a group of "fair-minded" people that are representative for the Swedish society and consider the above criteria and constructively contribute according to the CALY paradigm with the goal to further societal welfare (Månsdotter et al., 2017a, 2017b). Alternatively, the choice of relevant capabilities could be based on data from a representative population sample (Månsdotter et al., 2017b).

1.3 Aims

The aims of this thesis are to:

- Pilot a possible survey procedure to estimate CALY weights and assess the appropriateness of different question types and wording used, the sampling and evaluation methods, particularly in regard to response rate, the quality of data and the results, and the overall feasibility.

- Derive initial estimates for the CALY weights, with focus on a plausible value for the worst state.

The first aim is relevant for a future estimation of the weights with higher credibility and reliability. The second aim also serves this purpose as prior information about coefficients can potentially be used adapt the data collection procedure to yield and make it more efficient. Moreover, the second aim potentially allows comparisons between different attributes and levels, and the weights' plausible range can deliver information about the relative relevance of selected capabilities and the applicability of the CALY concept for its intended purpose in policy making.

Before the CALY measure can be employed in a cost-effectiveness framework, it is necessary to obtain weights. This work thus aims to contribute to the development of the CALY

measure by piloting the calculation of CALY weights (Månsdotter et al., 2017a) and thus preparing for a valid estimation of the weights in the Swedish context.

2 Methodology

2.1 Capability Attributes and Levels

Based on the pilot study where 200 Swedish researchers ranked capabilities (Månsdotter et al., 2017b), that previously have been proposed by the Swedish government (SOU, 2015), the following 5 capability attributes were chosen for the scope of this work: economic resources, health, education, occupation (job-related or other occupation), and social relations.

For simplicity in both interpretation and estimation, the CALY measure is limited to 5 attributes with 3 levels each. The 3 levels correspond to full capability, partial capability, or low capability. This structure allows a total of 3^5=243 different configurations, or states. For example, the state (3,3,3,3,3) describes the best possible life situation with the highest capabilities levels in all attributes, whereas the state (1,1,1,1,1) is the worst state with the lowest capability levels in all attributes. Alternatively, states are denoted as sX with X reaching from 1 to 243. s1 corresponds to (1,1,1,1,1), s2 to (2,1,1,1,1), s3 to (3,1,1,1,1), s4 to (1,2,1,1,1), …. , and s243 to (3,3,3,3,3).

2.2 Sampling Procedure and Invitation

On April 12. 2017 we sent 2'000 letters out to a random sample of the Swedish population aged 18 to 75. The subset including the addresses were obtained from Statistics Sweden (Statistika centralbyrån). At least 1 of the used address labels was invalid, resulting in 1999 valid addresses. Because address labels were not provided digitally, they were scanned and the text recognition software Tesseract (Tesseract, 2017) was used to convert all the postcodes to a digital format. Survey participants were asked to provide their postcode in the survey which was then verified against the digitized postcodes before including the responses of the participants into the analysis. In the case of multiple identical postcodes, responses with postcodes that exceeded the number of legible, identical postcodes that were used on the address labels, were excluded. Therefore, responses were considered according to the time submitted: Earlier responses with identical postcodes were preferred over later submitted responses with identical postcodes.

The letters contained a link to participate in the web survey along with a short orientation about the aims of the research project. The web survey was constructed by using HTML, JavaScript, PHP, SQLite, and MySql/MariaDB and hosted on a commercial webspace provider.

2.3 Survey Composition and Types of Questions

Each participant that visited the website was assigned to one of three versions, namely the version that was administered the least amount of times at that time point. The aim was to distribute each survey version an equal amount of times to enable conclusions about the differences in the way participants interacted with the survey. The versions differed in the type of question and the framing of the questions. For the specific composition of the survey versions and the order of the questions, refer to Table 1.

Part	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Question Nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Version 1	1 self assessment for all attributes	general self-assessment with a slider	4 discrete choice (DC) paired comparison questions				1 DC time trade-off (TTO) question in the absolute life expectancy (ALE) framing	1 TTO ranking question in the ALE framing	Background questions	Optional possibility to leave feedback
Version 2	‘‘	‘‘	‘‘				1 DC TTO question in the relative life expectancy (RLE) framing	1 TTO ranking question in RLE framing	‘‘	‘‘
Version 3	‘‘	‘‘	‘‘				1 DC TTO question in the ALE framing	1 DC TTO question in RLE framing	‘‘	‘‘

Table 1: Survey Versions
‘‘ indicates identical cell content as the adjacent cell in the row above. Part refers to categories of the same questions. Question Nr. Indicates the order of which question appeared in the web survey.

The self-assessment question asked the participants to indicate for each of the capability-attributes which option applied to their situation by choosing either “completely true”, “partially true” or “not at all tSaugesrue”. Alternatively, they could select for each attribute to omit to answer.

The general self-assessment presented the participants with a slider where they could mark on a scale from 0 to 100 how good their current life is, where 0 was marked with “a bad life” and 100 with “a good life”. Alternatively, participants could choose not to answer that question.

For both the discrete choice (DC) paired comparison and the discrete choice trade-off (TTO) questions, the two individuals were described to be in a specific life situation that corresponds to a certain state, using a graphical representation and a shorter version of the descriptions used in the self-assessment question. The graphical representation consisted of a bar chart with 5 bars, each representing one of the capability attributes, and the level of this attribute was indicated with a blue bar. If the lowest level 1 was present, the space was empty. See Appendix I for examples. The detailed description was additionally accessible to the participants in the form of a graphical pop-up. Participants marked their selection with a radio button, without possibilities to opt-out and skip the question.

Different terminologies are used in the literature. Paired comparisons, as used here and in Salomon et al. (2012), are also known as pairwise comparisons, for example in Brazier et al. (2012).

In the DC paired comparison questions, the participants were asked to judge which of two hypothetical individuals had a better life situation.

In the TTO questions in the absolute framing, each of the two individuals was described to life until a certain age. The age was fixed at thirty years for the individual with the perfect life situation and randomly chosen from 33, 38, 45, 54, 67, and 90 for the individual with the non-perfect life situation. For the relative framing, the remaining life expectancy was 6 years in the perfect state and chosen randomly for the non-perfect state out of 7, 9, 12, 18,30, and 60 years. The individuals were asked to judge if the first situation was preferable, the second situation was preferable, or if they are approximately equally preferable.

In the TTO ranking questions, the same levels were employed and the 3 options that were combined with the perfect state were randomly selected without repetition from all possible state-level combinations (s1, s37, s113, s131, and s208 combined with one of the respective levels). The short description was omitted for this type of question, and the were only described with their names (economic resources, health, education, occupation, and social relations). The detailed descriptions were still available in the form of a pop-up dialog. The participants were asked to state their ranking by dragging blocks that represented the different states in the according order into a specified area. See Appendix I for examples.

For all the paired comparison,and TTO questions, the order of display was randomized.

The background questions asked the participant about age (categories 16-19, 20-29, 30-39, 40-49, 50-59, 60-69, and over 70), postcode, gender (categories man, woman, other), education (categories less than 9 years schooling, 9 years obligatory schooling, high-school or similar, and university or similar), living status (categories alone, with parent/parents, with partner/wife/husband, and shared flat) as well as the place of birth (categories Sweden, Europe, and rest of the world). All the background questions included an option to not to answer, with the exception of the field for the postcodes, since it was required for the verification of the results. For examples of each kind of questions, see Appendix I.

Important criteria in constructing the wording of the questions and descriptions were ease of understanding and reflection of the capability approach . The Swedish descriptions of the different attributes were for these reasons written as statements of anonymous persons in the first person. For example, the translated description for the attribute education was:

"*I have the education, experience and skills that are necessary so that i can largely work as what I want and devote myself to what I want.*"

translated from

"*Jag har den utbildning, erfarenhet och skicklighet som krävs för att i stort sett kunna arbeta med och ägna mig åt det jag vill.*"

The formulation for the persons affected was changed. While the first person perspective was used for the self-assessment question, the TTO ranking questions used the wording "this person… " to indicate affiliation to the display block, while the questions with two states to choose from used the wording "the first person… ", and "the second person…" respectively (Appendix I).

2.4 Ethical Considerations

Ethical approval to conduct the study was obtained from the Umeå University ethical board with the decision 2017/53-31. Sensitivity of the collected data is limited as no details regarding personal situations was collected, but only summary measures used for statistical aggregation, for example gender, and a ratings about the respondents life. Nevertheless, it is crucial to handle entered data carefully. Several measures were employed to increase security of the infrastructure of the web survey, such as the HTTPS protocol and prepared database statements. Additionally, the feedback was stored separately so that the associated email addresses, which could be optionally submitted by participants in case they wanted to be informed about the research results, were not able to be associated to the respondents

answers and postcodes.

Similarly, the used address labels that were scanned needed to be protected. Thus, no address information was uploaded to the internet.

2.5 State Selection, Experimental Design and Sample Size

Discrete choice experiments such as the the paired comparison questions can be structured more efficiently by optimizing which states are used in the questions. In other words, the number of necessary questions to estimate an effect with a given precision can be decreased by employing a more efficient experimental design (Reed Johnson et al., 2013; Street et al., 2005).

Here, the approach described in Street et al. (2005) was used to create a design for the paired comparisons. Starting from the full factorial design, a generator was added to obtain the second state for each comparison. The full factorial design refers to all 243 possible combinations of levels of attributes. A generator consists of one number for each attribute that is added to each attribute of a state to obtain the levels for other state. The addition is performed modulo the number of levels and thus it is necessary to first translate the levels to a 0-based representation. The generator used was 2,1,2,1,2. For example, the comparison state for (2,2,2,2,2) would become (2,1,2,1,2): First (2,2,2,2,2) is changed to a zero based representation (1,1,1,1,1), then the generator is added, which results in (3,2,3,2,3), and then the modulo 3 operation performed. Pairs of states where all attributes in one state were higher than in the second state were discarded.

For the TTO questions, 5 states were selected that were thought to represent a broad spectrum of capabilitiy endowments. As a proxy for severity, the sum of all levels for each attribute was used. The used states were: s1 (1,1,1,1,1), s37 (1,1,2,2,1), s113 (2,2,1,2,2), s131 (2,2,3,2,2), s207 (3,3,2,2,3) with proxy severity of 5 ,7, 9, 11, and 13. A lower number means hence less capability endowment.

Minimally 10 observations for each of the 60 possible combinations of level and state for both the ALE and the REL framing in the TTO questions were targeted; including the TTO ranking questions. We expected a response rate of 15%, which translates into 100 participants per survey version, and a total of 400 DC TTO observations and 600 ranking observations, with respective shares for the absolute and the relative framing of 50%.

Hence, with 4 paired comparison questions per administered survey version we were expecting 1200 paired comparison observations.

2.6 Statistical Methods

Thurstone (1927) laid the theoretical foundations for discrete choice decision models where an individual's decision is thought to be based on an internal value. Methods based on this approach were first applied in a utility-theory context (Bradley, 1984; Luce, 1959; McFadden, 1973; Train, 2003, p. 19). Later on, the concept has been applied to health outcomes (Fanshel and Bush, 1970) and more modern approaches involving disability weights (Dolan et al., 1996; Haagsma et al., 2014; Lopez and Murray, 1998)

The statistical analysis was performed in three stages, based on (Salomon et al., 2012). The

underlying assumption is that the decision between two states is guided by an internal value that represents the desirability of a state. Individuals that choose between different options compare the internal values for each of the options and choose the option with the highest internal value. In a different context, this value is called latent utility(Ratcliffe et al., 2009), but due to the paradigm used in the present thesis that focuses on capabilities the term utility will be avoided. The internal value of a possible choice is assumed to be the sum of the individual specific values of the attributes for each of the 5 attributes that correspond to the specific state under evaluation.

2.6.1 First stage

Firstly, the paired comparisons were evaluated using probit regression and dummy coding (Bech and Gyrd-Hansen, 2005; Daly et al., 2016). The probit regression estimates the internal values associated for the levels of the attributes. The random-effects model is given by $Y_{ij}^{*}=x_i'\mathrm{B}+e_{ij}$ assuming normally distributed latent values representing attribute levels with $Pr(Y=1|X=x)=\Phi(x,\mathrm{B})$. B is a vector of the estimated coefficients for the dummy-coded attribute levels, and x_i is a vector representing dummy coding for the i-th observation by individual j with outcome Y_{ij} and internal latent values Y_{ij}^{*}. The resulting values for the attribute levels are on an arbitrary scale in probit space and not anchored on the desired 0 to 1 scale. Thus, in stages 2 and 3 the anchoring is performed by using the weights obtained from the TTO observations and calculating a mapping to the 0 to 1 scale for the values derived with the probit regression. A mapping in the form of a linear transformation seems to be more accurate compared to other approaches (Rowen et al., 2015).

2.6.2 First stage rationale

The reason why not all capability weights are directly calculated using TTO questions is that TTO questions were criticized to be more complex and susceptible to bias in comparison to paired comparison questions (Bleichrodt, 2002; Brazier et al., 2012; Haagsma et al., 2015; Rowen et al., 2015). For example, loss aversion may influence TTO evaluations(Brazier et al., 2007, pp. 159–160). Another reason is the estimation of weights with TTO methods may require considerable resources (Rowen et al., 2015).

2.6.3 Second stage

Secondly, the trade-off observations from both the ranking and the paired TTO questions were evaluated using interval regression. Here, the internal value for each state is directly estimated. The perfect state (3,3,3,3,3) is anchored to 1. A choice in a TTO question implies an interval for the weight of the compared state in relation to the perfect state: For example, in the relative framing, a preference for 60 years remaining life expectancy in state A compared to the perfect state B that is anchored to 1 implies for the internal value and for the weights of A and B: $w_A*60>w_B*6$ and thus $w_A>w_B*\frac{6}{60}>0.1$ since $w_B=1$. Thus, the implied interval for A's weight for this observation is $w_A\in(0.1,1)$, with the assumption that $w_B=1>w_A$. If the participant stated that the life situations are about equal, the implication is that $w_B=0.1$. Similarly, the data from the ranking questions was transformed to binary comparisons: For example, the a ranking of A > B > C > D implies that B < A, B > C, and B > D, if B is the option that represents the perfect state. This conversion requires the assumption

of independence of irrelevant alternatives, where the choice behavior is not altered by the characteristics of alternatives other then the ones converted to pairs(Train, 2003, p. 50).

2.6.4 Second Stage Rationale

Per ranking question with 3 options in addition to the perfect state, 3 interval censor points observations are obtained. Hence, the ranking questions yield more observations in the same amount of questions compared to DC questions with just two options under consideration in each question. No individual specific fixed effects or random effects models were considered for reasons of complexity and lacking sample size.

The reason behind employing different both the RLE and ALE framing was to explore which one is more valid. The ALE framing has the disadvantage of suffering from a limited plausible range of implied intervals. If the best state, that is used as the comparison to trade-off with, has an associated life span of 30 years, and a maximum plausible life expectancy of 90 years, the smallest implied interval is 1/3 which may be too large to adequately capture lower values. We chose the age of 30 years a reasonable duration associated to the best state because it is the age at which the average person has taken the most important decisions regarding the rest of its life, such as the choice of profession, and possibly regarding family.

The different level of years in the TTO questions were chosen to offer a good coverage of discrimination points over the interval 0 to 1 in both the relative- and the absolute life expectancy framing. Furthermore, the ALE framing may be susceptible to bias where equal time spans are not perceived to be equally viable depending at what age they are started. For example, it is reasonable to assume that the age span 70 to 90 may be differently valued by participants than 30 to 50. Thus, while the ALE framing may be more intuitive because the lifetime in years of a person is often handled by an average person, the, the RLE framing has the potential to cover a wider range of censor points for constructing the intervals and be somewhat less susceptible to differences in judgment regarding the achieved age.

Table 2 contains an overview of all combinations of framings and question types used in the TTO questions. The table scheme will be used when reporting the results.

DC ALE	**DC RLE**	**DC Combined**
Ranking ALE	**Ranking RLE**	**Ranking Combined**
ALE combined	**RLE combined**	**All combined**

Table 2: Combinations of TTO Framing and Question Types
DC ALE refers to discrete choice absolute life expectancy observation data.
DC RLE refers to the discrete choice relative life expectancy observation data.
DC combined refers to all the discrete choice observations combined.
Ranking ALE refers to the ranking absolute life expectancy observation data.
Ranking RLE refers to the ranking relative life expectancy observation data.
Ranking combined refers to all the ranking observations combined.
ALE combined refers to all the absolute life expectancy observations combined, irregardless of DC or Ranking question type.
RLE combined refers to all the relative life expectancy observations combined, irregardless of DC or Ranking question type.
All combined refers to all TTO observations combined.

Another way of eliciting TTO values is using a series of questions to delimit intervals and the point of indifference for a state (Dolan et al., 1996). For example, a respondent may indicate in the first question a possible range from 0.5 to 1 by selecting the first option in a choice between 10 years in the state in question and 5 years in the perfect state. In the second

question, the respondent may choose the first option in a choice between 10 years in the perfect state and 7 years in the other state, reducing the possible range from 0.5 to 0.7, and so on. The present methods based on interval regression was chosen because because of the ease of integrating it into the web survey, since a fixed number of questions can be administered per participant. However, the method has been criticized to not yield better fits than other models(Brazier et al., 2007, p. 147).

Because normally distributed weights on the 0 to 1 scale are not likely to be a reasonable assumption (Salomon et al., 2003, pp. 415–430), the interval boundaries were first remapped to probit space by using the quintile function, before performing the interval regression. Using a probit transformation instead of a logit transformation that was used in Salomon et al. (2010) given that the analysis of the paired comparison data was performed with a probit regression and the similarity the of the logit and probit functions.

2.6.5 Third Stage

In the third step, the probit values obtained through the paired comparison observations were rescaled to the desired 0 to 1 scale by using the calculated weights for the states used in the TTO questions. The probit values of the states used in the TTO questions were used to estimate a linear transformation with a linear regression in the probit space that maps the probit weights to the desired 0 to 1 scale. Additionally, weights were also mapped to the 0 to 1 scale with a stretching factor that was calculated by dividing the weight of the worst state s1 resulting from the TTO observations with the weight in probit space for the worst state resulting from the paired comparison questions: $\lambda = \frac{weight_{11111\text{TTO}}}{weight_{11111\text{probit}}}$

This factor was then multiplied with all probit weights for reanchoring, ensuring that the lowest state is mapped to the TTO weight of the lowest state(Rowen et al., 2015; Salomon, 2003). These transformations were then used to calculate weights for all 243 possible states on the 0 to 1 scale.

2.6.6 Third Stage Rationale

The two methods of a linear transformation and a stretch vector were employed to explore their differential effects. While the linear transformation may be more accurate and seems to perform better (Rowen et al., 2015), the stretch factor has the advantage of an exact mapping for the worst state.

2.6.7 Other Methodological aspects

Furthermore, various descriptive statistical methods were used to describe the characteristics of respondents and their response behavior, such as chi square tests and t-tests.

All statistical analysis was performed using R (The R Foundation, 2017). The glm function was used for the probit regression and the package survival (Therneau and Lumley, 2017) was used for the interval regression. The package lme4 was used for random effects probit regression (Bates et al., 2017).

3 Results

3.1 Response Rate

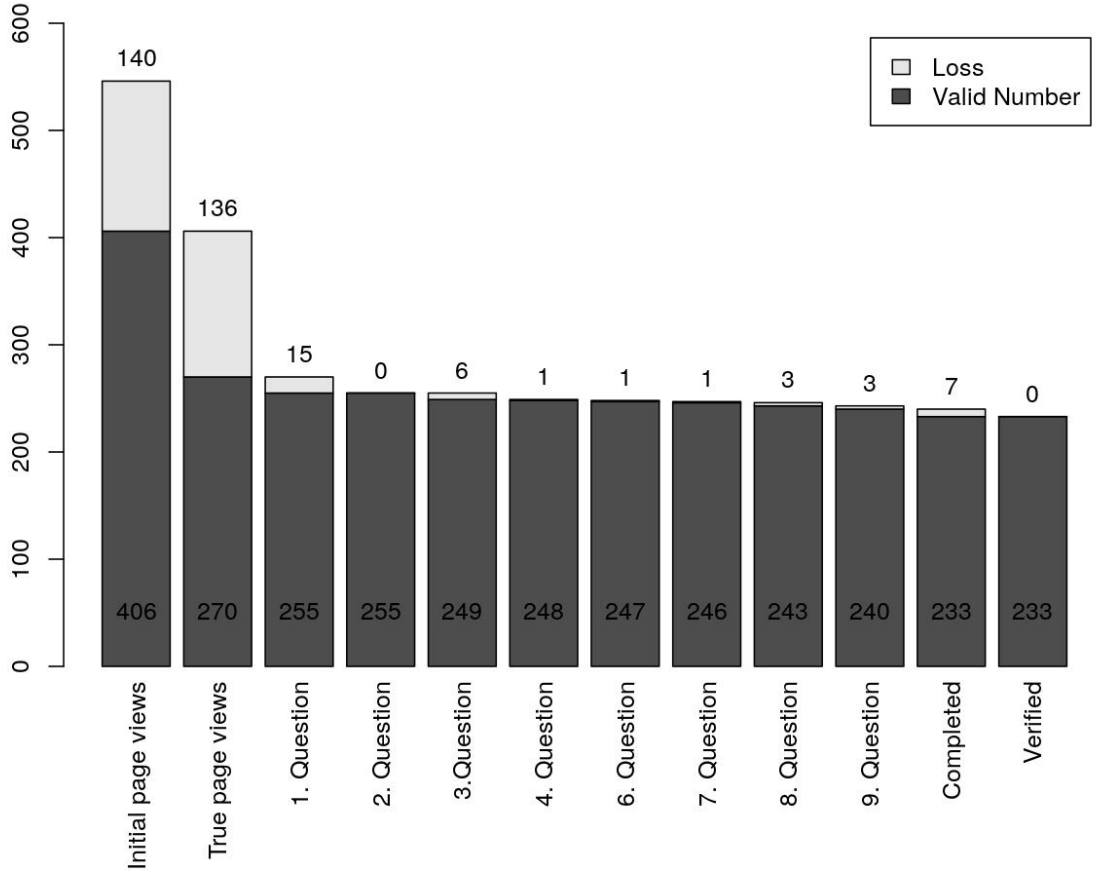


Figure 1: Dropout Numbers
Displays the number of total and discarded respondents grouped by questions. The loss proportion refers to the number of respondents who did not proceed to the next question. .The number of 30 losses from the total 358 unique page views refers to page visits that were estimated to be non-human; that is by web crawlers for example used by search engines. Valid designates the number that passed to the next stage.

Figure 1 depicts the number of respondents who finished the survey starting from April 13. until May 1., and the dropout numbers by question. 140 of the unique page visits for the initial page of the web survey were assumed to be by bots such as web-crawlers used by search engines, based on an approximate classification using the user-agent string.

Because the internal dropout number was low (n=15) without the exception of the first question about the self assessment (n=15) that did not gather information that can be used in calculating the weights, data from respondents that did not complete the whole survey was discarded before statistical analysis. Similarly, the answers of respondents that did not provide a verifiable postcode (n=7) was not included in the data analysis. Figure 2 contains a map with the geographical distribution of all the postcodes.

Thus, the overall response rate of verified answers was 233 out of 1999 or 11.7%. 406 unique page visits were registered, that is 20.3% of all the total number of addressees, and 270 or 13.5% started to fill out the survey. No noteworthy differences regarding dropout numbers between the different survey versions were observed.

In total 139 individuals left an email address to be notified about the research results and 42 individuals left a feedback text. Out of these, 32 provided both an address and a feedback text.

84 individuals completed version 1 of the survey, 72 version 2, and 77 version 3.

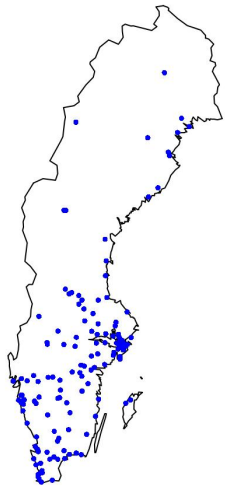


Figure 2: Map of Postcodes
Geographical distribution of verified postcodes using data from GeoNames (2017).

3.2 Timing

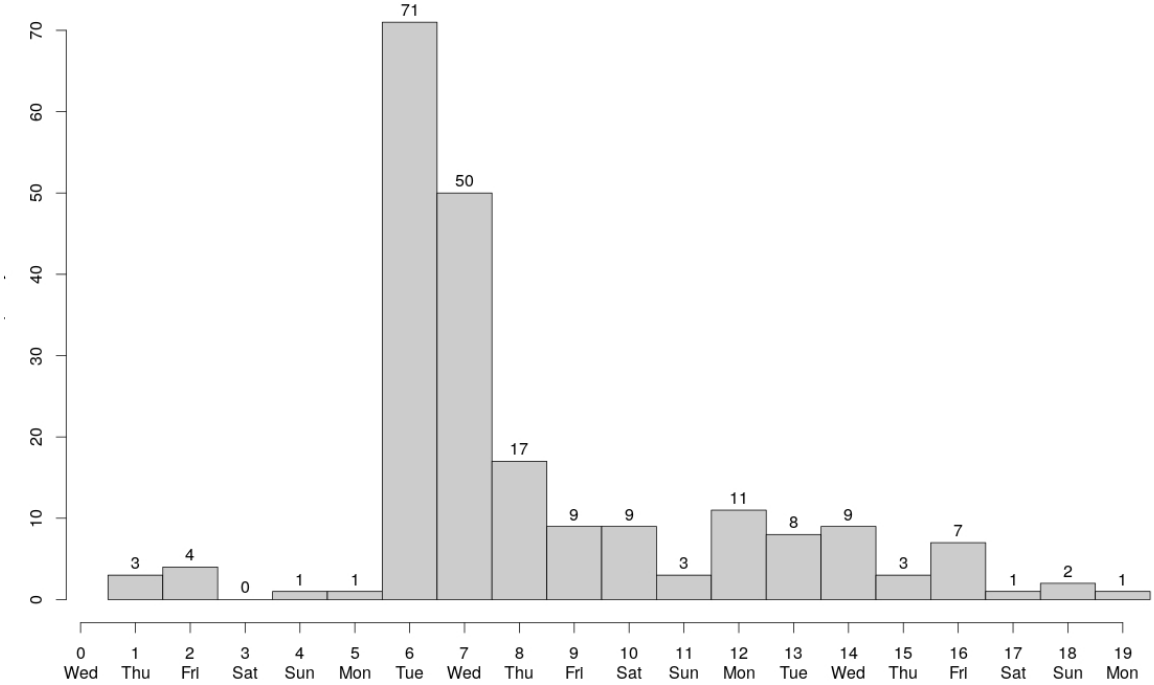


Figure 3: Answer Delay in Days After Sending Letters
Denotes the number of completed surveys after the day indicated on the x- axis. 0 corresponds to the 12.4.2017 when the letters where sent. Day 6 corresponds to Tuesday 18.4.2017, the day after Easter Monday. Only answers submitted within 3 after day 0 are displayed.

Figure 3 depicts the time-span it took in days after sending out the letters for respondents to fill in the survey. Most of the replies were received on day 6 and 7, after sending out the letters, with a decline afterwards.

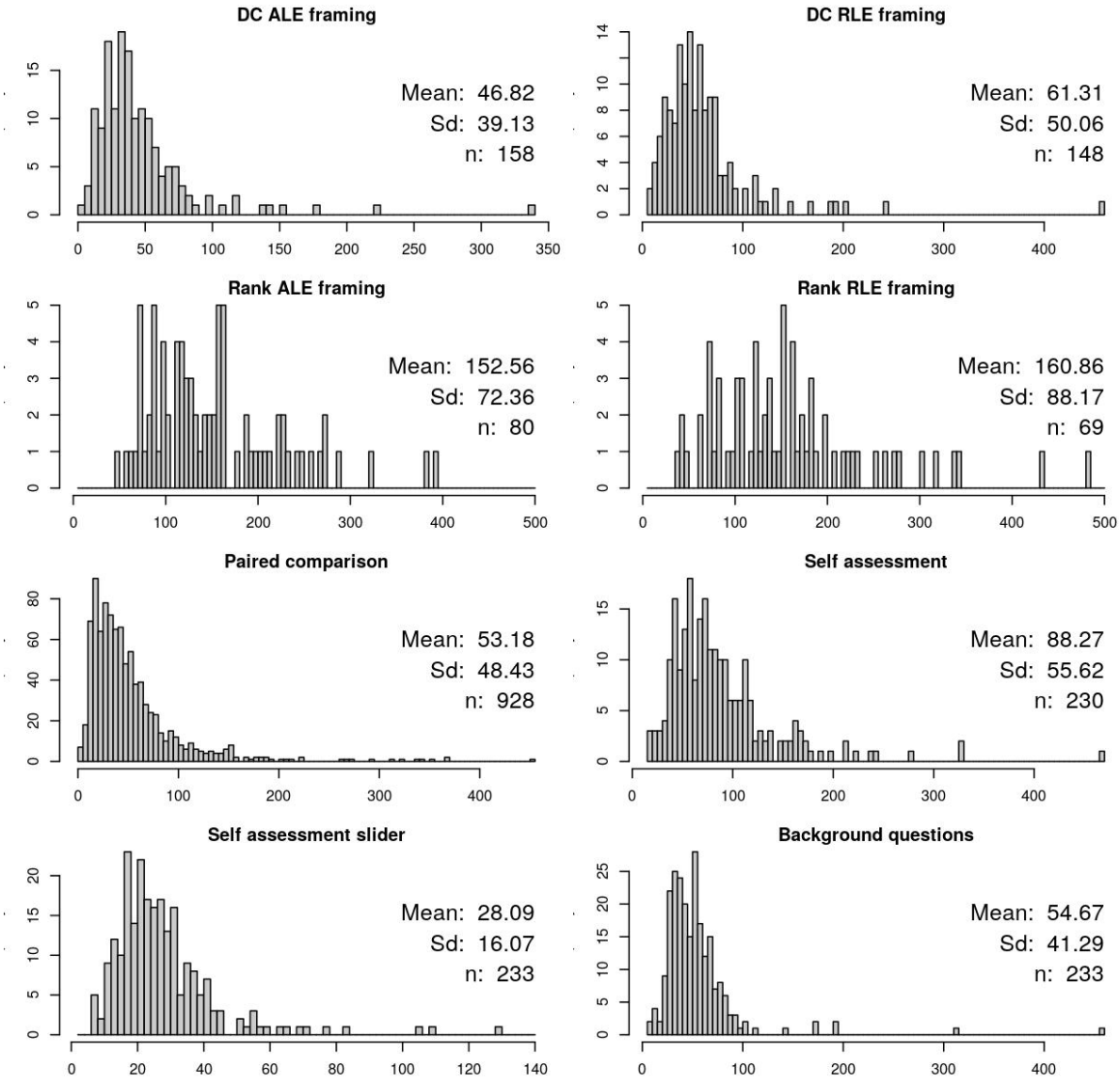


Figure 4: Distribution of Answer Times per Question Type
Time in seconds on the x-axis. Durations >=500 have been discarded. Sd is corrected and refers to the sample. For 8 observations, time intervals could not be determined.

In Figure 4, the distributions of the answer times in seconds, per combination of question type and framing, are displayed. Durations longer or equal to 500 have been discarded, as then respondents are likely to have had interrupted filling out the survey. Differences in means using a two sided t-test resulted in significant differences at $\alpha = 0.05$ between the ranking and DC TTO questions and between DC RLE and ALE framings, but not between the ranking question in RLE and ALE framings. Neither were the differences in mean durations for answering to the paired comparison questions and to the DC TTO questions significant.

3.3 Characteristics of Respondents

Gender			Age			Education			Living-status			Born		
	n	E		N	E		N	E		N	E		N	E
Man	98	117.6	**16-19**	10	7.0	**Less than 9 years schooling**	7	11.4	**Alone**	46	44.8	**Sweden**	214	184.8
Woman	134	114.4	**20-29**	28	44.0	**Primary school or similar**	19	25.5	**Partner**	165	137.7	**Europe**	12	22.3
Other	1		**30-39**	33	40.9	**High school or similar**	62	105.9	**Parents**	11	20.7	**Rest of the world**	7	25.9
Optout	0		**49-49**	37	42.5	**University or similar**	143	88.2	**Shared flat**	2		**Optout**	0	
Tot	233		**50-59**	40	40.6	**Optout**	2		**Other**	7	25.8	**Tot**	233	
			60-69	55	36.9	**Tot**	233		**Optout**	2				
			> 70	29	20.2				**Tot**	233				
			Optout	1										
			Tot	233										

Goodness of fit Chi square

χ^2	**df**	**p**	χ^2	df	p	χ^2	df	p	χ^2	df	p	χ^2	df	p
6.62	1	0.01*	22.09	6	0.00*	55.65	3.00	0.00*	23.67	3	0.00*	23.15	2	0.00*

Table 3: Demographic Characteristics

Demographic characteristics of the verified respondents, compared to the expected frequencies for the Swedish population obtained form the Swedish statistical office (Statistika Centralbyrån, 2016). N is the number of observed occurrence, E refers to the number of expected values given the relative frequencies of the whole population. On the bottom the results for a goodness of fit chi square test are displayed. The category shared flat is was combined into other for the chi square test. Individuals whose country of birth was marked "unknown" in the government statistics were disregarded. Only two categories, man and woman, existed in the government statistic. The opt out categories were not included in the chi square test. P-values significant at α = 0.05 are indicated with *. χ^2 is the test score for the chi square test and df refers to the degree of freedom.

Table 3 contains the characteristics of the verified respondents, in comparison to the composition of the Swedish population. The sample appears to be significantly different composed than the Swedish population with high probability: For all categories the null hypothesis of no difference is significantly rejected. The statistics for the whole Swedish population were obtained from the Swedish statistical bureau (Statistika Centralbyrån, 2016). for the age categories of 18-75.

Figure 6 contains the distribution of answers to the self assessment question and Figure 5 the distribution of the answers in the slider self assessment question, excluding the 2 respondents who chose not to answer. A peak is observable at 100.

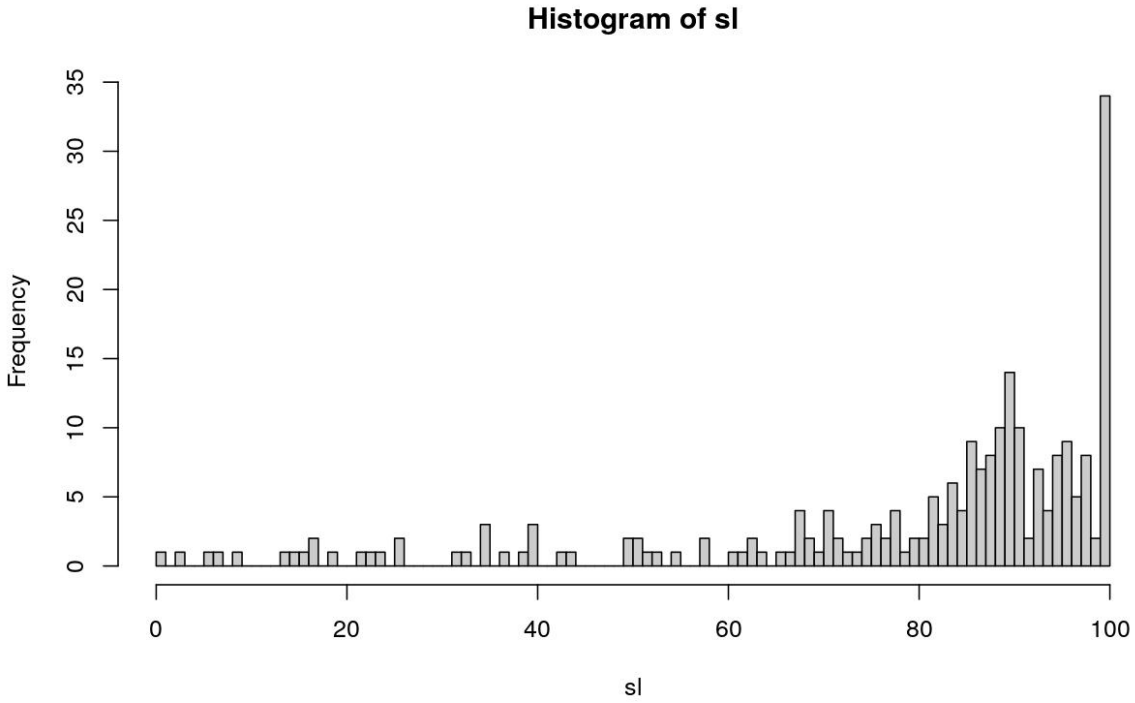


Figure 5: Distribution of Answers in the Slider question
Frequency is the number answers per value on the X axis. 2 opt out answers are excluded.

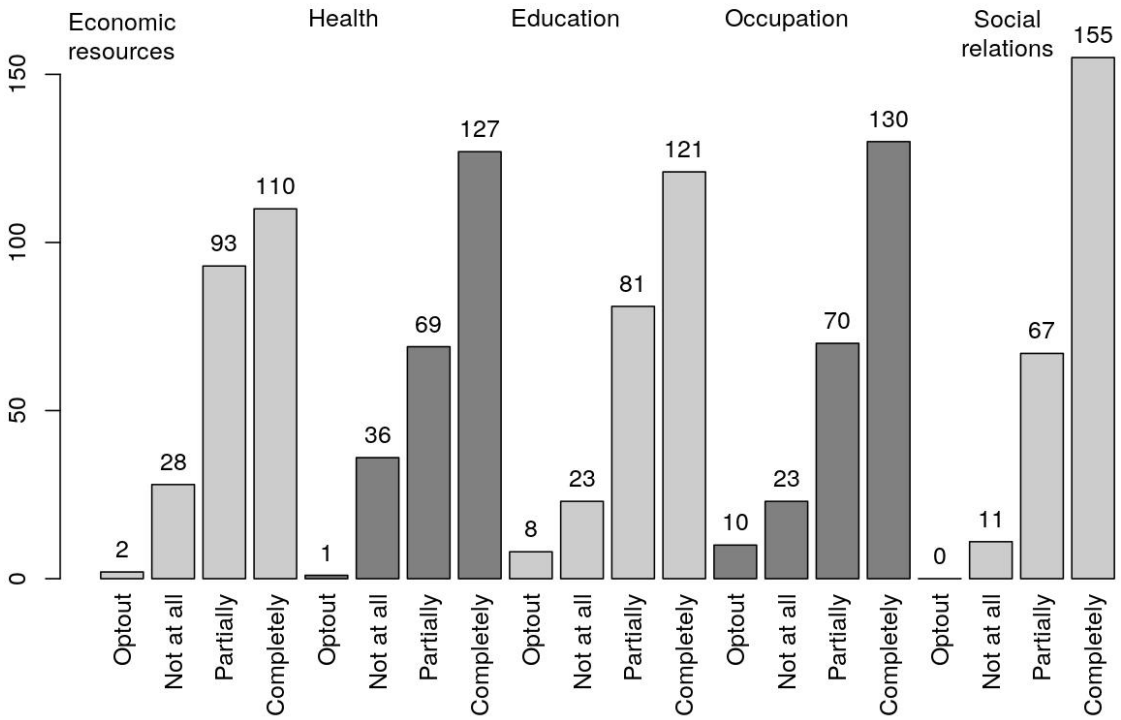


Figure 6: Distribution of Self Assessment Answers
The number of answers per category and per attribute in the self rating question that asked for the participants assessment of their capabilities.

3.4 Weight Calculation

3.4.1 Probit Regression for Paired Comparison Data

	Dummy	Difference
eco2	0.487*** (0.083)	0.487*** (0.083)
eco3	0.637*** (0.082)	0.150* (0.081)
health2	0.960*** (0.083)	0.960*** (0.083)
health3	1.340*** (0.087)	0.380*** (0.082)
educ2	0.180** (0.082)	0.180** (0.082)
educ3	0.428*** (0.082)	0.247*** (0.081)
occu2	0.545*** (0.082)	0.545*** (0.082)
occu3	0.840*** (0.085)	0.295*** (0.081)
socrel2	1.165*** (0.085)	1.165*** (0.085)
socrel3	1.618*** (0.089)	0.453*** (0.080)
Constant	-2.774*** (0.141)	-2.774*** (0.141)
Observations	1,864	1,864
Log Likelihood	-946.008	-946.008

Table 4: Probit Regression Coefficients
The coefficients refer to attribute levels of economic resources, health, education, occupation, and social relations in the order listed (1,2,3,4,5). The attribute level increase from low to high and correspond to "not at all" (represented by the constant), "partially true" (level 2), and "completely true" (level 3). The standard errors are reported in brackets. * indicates significance at α =0.1, ** at α =0.05, and ***at α =0.01. Dummy coding has been used in the left column. Thus, the intercept corresponds to the value for (1,1,1,1,1). In the difference coding in the right column, the next level represents the difference to the previous level, while the intercept still corresponds to (1,1,1,1,1).

Table 4 contains the probit regression coefficients. Because the estimation did not detect a random effect, a normal probit regression is reported instead. All coefficients including the constant are significantly different from 0 at $\alpha=0.05$ for the dummy coding. However, this information is of little value as the location of the coefficients in relation to zero is arbitrary, depends on the value for the intercept, and can for example vary with the coding scheme used. Instead, the standard error is more informative, which refers to the probit space, that is in space corresponding to the z-score from the cumulative normal distribution A standard error of 0.09 in probit space around 0 where the cumulative normal distribution is the steepest corresponds to a probability difference of approximately 0.036. More towards the outer ends of the scale at z=1.96 or z=-1.96, a standard error of 0.09 corresponds to a probability difference of 0.005.

The model on the right hand side in Table 4 contains the estimates for the differences compared to the previous level where all the coefficients are estimated significantly for α =0.05, with the exception of education.

The coefficients for health and social relations are larger than the coefficients for economic resources and education. Across all 5 attributes, the step from level 1 to level 2 is larger than the change from level 2 to level 3, with the exception of education where the estimate was not highly significant for the difference of level 1 to 2. This may be an indication that it is more important to have a partially fulfilled capabilities rather perfect capabilities.

Another probit regression was run that included the answers to the self assessment slider question as a coefficient, scaled to the interval [0,1] with a division by 100. However, this coefficient was not estimated significantly.

3.4.2 Interval Regression Time Trade-Off Data

	DC ALE	DC RLE	DC combined
s37	-0.345 (0.316)	-0.465 (0.413)	-0.512* (0.299)
s113	0.491* (0.259)	0.356 (0.376)	0.486* (0.262)
s131	0.542** (0.254)	0.515 (0.387)	0.611** (0.262)
s207	1.095*** (0.272)	0.898** (0.381)	1.106*** (0.269)
Constant	-0.492** (0.209)	-1.033*** (0.315)	-0.868*** (0.215)
Observations	161	149	310
Log Likelihood	-82.874	-85.960	-178.039
	Rank ALE	**Rank RLE**	**Rank combined**
s37	-0.220 (0.583)	-1.322 (1.134)	-0.900 (0.806)
s113	2.045*** (0.782)	2.131* (1.114)	2.931*** (1.066)
s131	2.432*** (0.808)	2.670** (1.279)	3.724*** (1.218)
s207	3.294*** (1.007)	3.611** (1.430)	4.923*** (1.480)
Constant	-1.890*** (0.701)	-3.192*** (1.213)	-3.554*** (1.124)
Observations	252	216	468
Log Likelihood	-114.406	-104.196	-226.061
	ALE Combined	**RLE Combined**	**All combined**
s37	-0.262 (0.268)	-0.699* (0.412)	-0.579** (0.282)
s113	0.964*** (0.252)	0.884** (0.370)	1.085*** (0.261)
s131	1.154*** (0.245)	1.116*** (0.393)	1.390*** (0.267)
s207	1.759*** (0.277)	1.682*** (0.399)	2.041*** (0.288)
Constant	-0.915*** (0.210)	-1.616*** (0.336)	-1.489*** (0.230)
Observations	413	365	778
Log Likelihood	-205.396	-197.824	-420.786

Table 5: Interval Regression Coefficients
Resulting estimates for the TTO states from observations of the different question types and framings: DC ALE, DC RLE, Rank ALE and rank RLE data. DC combined and rank combined contain the estimates for all the data from the discrete choice and ranking TTO questions respectively, irregardless of ALE or RLE framing. ALE and the RLE combined contain the estimates for the pooled data from all the ALE framing TTO questions and RLE framing TTO questions respectively, irregardless of discrete choice or ranking question type. All combined are the estimates based on all data from TTO questions. Dummy coding has been used. Thus, the intercept corresponds to the value for state s1. The standard errors are reported in brackets. * indicates significance at α =0.1, ** at α =0.05, and ***at α =0.01.

Table 5 contains the coefficients of four separate regressions for the data from the ranking and the TTO questions and for the absolute and the relative framing. Table 5 also contains models for pooled combinations of DC, ranking, ALE and RLE data. The value of the coefficients is inconsistent as the value for state s1 which corresponds to the intercept, should be lower than the one for s37, because (1,1,1,1,1) is expected not to be preferred to (1,1,1,2,2,1) by a rationale choice agent.

	DC combined	Rank combined	All combined
s37	-0.387 (0.258)	-0.593 (0.582)	-0.440* (0.240)
s113	0.456** (0.225)	2.322*** (0.700)	1.023*** (0.223)
s131	0.546** (0.224)	2.765*** (0.762)	1.217*** (0.225)
s207	1.027*** (0.231)	3.722*** (0.906)	1.826*** (0.242)
irle	-0.575*** (0.135)	-1.094*** (0.308)	
idc_rle			-0.534*** (0.186)
irank_ale			0.247 (0.170)
irank_rle			-0.583*** (0.176)
Constant	-0.513*** (0.192)	-2.169*** (0.655)	-1.139*** (0.228)
Observations	310	468	778
Log Likelihood	-170.436	-219.554	-404.965

	ALE combined	RLE combined
s37	-0.272 (0.270)	-0.699* (0.413)
s113	0.986*** (0.256)	0.888** (0.372)
s131	1.171*** (0.248)	1.118*** (0.395)
s207	1.771*** (0.281)	1.688*** (0.400)
irank	0.198 (0.146)	-0.062 (0.201)
Constant	-1.043*** (0.239)	-1.585*** (0.350)
Observations	413	365
Log Likelihood	-204.429	-197.775

Table 6: Interval Regression Coefficients for Combined Data with Estimated Differences
Resulting estimates for the TTO states from observations of the different question types and framings: DC ALE, DC RLE, Rank ALE and rank RLE data. DC combined and rank combined contain the estimates for all the data from the discrete choice and ranking TTO questions respectively, irregardless of ALE or RLE framing. ALE and the RLE combined contain the estimates for the pooled data from all the ALE framing TTO questions and RLE framing TTO questions respectively, irregardless of discrete choice or ranking question type. All combined are the estimates based on all data from TTO questions. irle contains the estimated difference of the RLE framing compared to the ALE baseline. idc_rle is the estimated difference to the baseline discrete choice in the ALE framing. irank_ale and irank_rle contain the estimation for the difference between the ranking question in the RLE framing and the ALE framing respectively compared to the ALE discrete choice baseline. irank is the estimated difference for the ranking questions compared to the baseline discrete choice questions. Dummy coding has been used. Thus, the intercept corresponds to the value for state s1.The standard errors are reported in brackets. * indicates significance at α =0.1, ** at α =0.05, and ***at α =0.01.

Table 6 contains regressions with combined data where the difference between the RLE and ALE framings, as well as ranking versus discrete choice format were estimated. Differences between the ALE and RLE framing were significant, whereas the difference between TTO ranking questions and discrete choice TTO questions did not result in a significant estimate. The significant difference in the combined regression for discrete choice TTO answers between the ALE and RLE framing disappeared when observations "about equal" observations were removed. Again, the interpretation of the p-values for main coefficients in Tables 5 and 6 is limited because the location of 0 in relation to the scale depends on the estimate of the intercept.

	DC ALE	DC RLE	DC combined
s37	-0.345 (0.316)	-0.465 (0.413)	-0.512* (0.299)
s113	0.836*** (0.309)	0.821** (0.366)	0.998*** (0.281)
s131	0.051 (0.228)	0.158 (0.314)	0.125 (0.224)
s207	0.552** (0.235)	0.383 (0.308)	0.494** (0.224)
Constant	-0.492** (0.209)	-1.033*** (0.315)	-0.868*** (0.215)
Observations	161	149	310
Log Likelihood	-82.874	-85.960	-178.039

	Rank ALE	Rank RLE	Rank combined
s37	-0.220 (0.583)	-1.322 (1.134)	-0.900 (0.806)
s113	2.264*** (0.847)	3.453** (1.352)	3.831*** (1.238)
s131	0.388 (0.459)	0.539 (0.705)	0.792 (0.580)
s207	0.861* (0.488)	0.941 (0.709)	1.199* (0.615)
Constant	-1.890*** (0.701)	-3.192*** (1.213)	-3.554*** (1.124)
Observations	252	216	468
Log Likelihood	-114.406	-104.196	-226.061

	ALE Combined	RLE Combined	All combined
s37	-0.262 (0.268)	-0.699* (0.412)	-0.579** (0.282)
s113	1.226*** (0.283)	1.583*** (0.378)	1.664*** (0.282)
s131	0.190 (0.202)	0.232 (0.290)	0.304 (0.205)
s207	0.605*** (0.203)	0.567** (0.287)	0.652*** (0.205)
Constant	-0.915*** (0.210)	-1.616*** (0.336)	-1.489*** (0.230)
Observations	413	365	778
Log Likelihood	-205.396	-197.824	-420.786

Table 7: Interval Regression Coefficients for Combined Data With Difference Coding
Resulting estimates for the TTO states from observations of the different question types and framings: DC ALE, DC RLE, Rank ALE and rank RLE data. DC combined and rank combined contain the estimates for all the data from the discrete choice and ranking TTO questions respectively, irregardless of ALE or RLE framing. ALE and the RLE combined contain the estimates for the pooled data from all the ALE framing TTO questions and RLE framing TTO questions respectively, irregardless of discrete choice or ranking question type. All combined are the estimates based on all data from TTO questions. Difference coding has been used. Thus, the coefficients correspond to the difference to the previous state, while the intercept still corresponds to the worst state s1. The standard errors are reported in brackets. * indicates significance at α =0.1, ** at α =0.05, and ***at α =0.01.

	DC ALE
state37	-0.364 (0.318)
state113	0.542** (0.262)
state131	0.589** (0.257)
state207	1.129*** (0.273)
sliderval	-0.574* (0.300)
Constant	-0.072 (0.282)
Observations	159
Log Likelihood	-80.445

Table 8: Interval Regression Adjusted for Answer from Slider Self Assessment Question
Observations from the DC questions in the ALE framing. * indicates significance at α =0.1, ** at α =0.05, and ***at α =0.01.

Another regression was run where the coding scheme represented differences between the levels, the results are depicted in Table 7. Differences between s1 and s37 as well as s113 and s131 seem to be less pronounced than between s131 and s207.

In order to test the influence of stated well-being on the responses, regressions were run that included a coefficient that reflected the answers of the slider self assessment question, which were rescaled to the scale from 0 to 1 by dividing it by 100. The coefficient was only significantly at α =0.1 for the regressions based on data from the discrete choice

questions in the ALE framing, displayed in Table 8, as well as the regression for the combined data from the discrete choice questions.

	DC ALE			DC RLE			DC combined		
	Weight	**95% CI**		**Weight**	**95% CI**		**Weight**	**95% CI**	
s1	0.31	0.18	0.47	0.15	0.05	0.34	0.19	0.1	0.33
s37	0.2	0.09	0.38	0.07	0.02	0.18	0.08	0.03	0.18
s113	0.5	0.37	0.63	0.25	0.14	0.4	0.35	0.24	0.47
s131	0.52	0.4	0.64	0.3	0.17	0.47	0.4	0.28	0.52
s207	0.73	0.61	0.82	0.45	0.3	0.6	0.59	0.47	0.71
	Rank ALE			**Rank RLE**			**Rank combined**		
	Weight	**95% CI**		**Weight**	**95% CI**		**Weight**	**95% CI**	
s1	0.03	0	0.3	0	0	0.21	0	0	0.09
s37	0.02	0	0.28	0	0	0.06	0	0	0.03
s113	0.56	0.3	0.8	0.14	0.02	0.48	0.27	0.07	0.58
s131	0.71	0.48	0.87	0.3	0.06	0.68	0.57	0.28	0.83
s207	0.92	0.72	0.99	0.66	0.31	0.91	0.91	0.65	0.99
	ALE combined			**RLE combined**			**All combined**		
	Weight	**95% CI**		**Weight**	**95% CI**		**Weight**	**95% CI**	
s1	0.18	0.09	0.31	0.05	0.01	0.17	0.07	0.03	0.15
s37	0.12	0.05	0.24	0.01	0	0.05	0.02	0.01	0.06
s113	0.52	0.4	0.63	0.23	0.13	0.36	0.34	0.24	0.45
s131	0.59	0.49	0.69	0.31	0.18	0.47	0.46	0.35	0.57
s207	0.8	0.71	0.87	0.53	0.38	0.67	0.71	0.61	0.8

Table 9: Weights Based on TTO Data
Resulting estimates for the weights of the TTO states from observations of the different question types and framings: DC ALE, DC RLE, Rank ALE and rank RLE data. Discrete choice combined and ranking combined contain the estimates for all the data from the discrete choice and ranking TTO questions respectively, irregardless of ALE or RLE framing. ALE and the RLE framing combined contain the estimates for the pooled data from all the ALE framing TTO questions and RLE framing TTO questions respectively, irregardless of discrete choice or ranking question type. All combined are the estimates based on all data from TTO questions. 95% CI contains the lower bound on the left and the upper bound on the right of a 95% confidence interval. Estimates and confidence interval limits were transformed from probit space with the cumulative normal distribution function.

In Table 9 the weights based on the TTO data are depicted. The data from ranking questions in the ALE framing results in estimates that are more evenly spread over the 0 to 1 scale than from weights resulting from the RLE results. The weights based on the ranking data tend more towards 0, especially for the RLE framing and cover a wider range than the weights resulting from the DC TTO questions. Also already recognizable in Table 6, the weights of s1 and s37 are inconsistent for all estimations. s1 is the worst state and should not be preferred over s37. Similar results are visible in Table 10 with the frequencies of selections that implied a left censored interval: s37 is tendentially less often preferred over the perfect state than s1, which is inconsistent.

	Frequency per level											
	ALE observations						RLE observations					
	33	38	45	54	67	90	7	9	12	18	30	60
s1	0.89	1	0.83	0.85	0.92	0.76	0.93	0.83	1	0.77	0.5	1
s37	1	1	0.83	0.95	0.8	0.85	1	1	1	0.95	0.69	0.88
s113	0.67	0.53	0.7	0.5	0.33	0.56	0.82	0.71	0.65	0.6	0.53	0.44
s131	0.92	0.62	0.61	0.38	0.2	0.21	0.8	0.75	0.6	0.56	0.36	0.25
s207	0.64	0.27	0.23	0.16	0.33	0.09	0.6	0.56	0.27	0.4	0.25	0.22
	Observations per level											
	ALE observations						RLE observations					
	33	38	45	54	67	90	7	9	12	18	30	60
s1	16	14	10	11	11	13	13	5	11	10	3	3
s37	19	19	10	18	8	11	11	13	13	19	11	7
s113	2	8	7	3	5	9	14	5	11	6	9	7
s131	12	8	14	5	2	3	4	6	9	5	5	3
s207	7	4	3	3	5	1	9	5	4	6	3	4

Table 10: Frequency of Choices Implying a Left Censored Interval
On top the frequency of choices in the discrete choice and ranking TTO questions, stratified by state and years for each possible answer option, that implied a left censored interval; that is where the perfect state was preferred over the state level combination in question. Below is the number of observations per combination of level and state. The level for the perfect state was 30 in the ALE framing and 6 in the RLE framing.

In order to overcome the issue of inconsistency, the data for s1 and s37 was pooled and assigned to s37 and another set of weight was calculated, which are depicted in Table 11.

	DC ALE			DC RLE			DC combined		
	Weight	95% CI		Weight	95% CI		Weight	95% CI	
s37	0.26	0.16	0.39	0.1	0.04	0.2	0.13	0.07	0.21
s113	0.5	0.37	0.63	0.25	0.13	0.4	0.35	0.24	0.47
s131	0.52	0.39	0.64	0.3	0.17	0.47	0.4	0.28	0.52
s207	0.73	0.61	0.82	0.45	0.3	0.61	0.59	0.47	0.71
	Rank ALE			**Rank RLE**			**Rank combined**		
	Weight	95% CI		Weight	95% CI		Weight	95% CI	
s37	0.02	0	0.25	0	0	0.07	0	0	0.04
s113	0.56	0.3	0.8	0.14	0.02	0.48	0.26	0.07	0.58
s131	0.71	0.48	0.87	0.3	0.06	0.69	0.57	0.27	0.83
s207	0.92	0.72	0.99	0.67	0.3	0.92	0.92	0.65	0.99
	ALE combined			**RLE combined**			**All combined**		
	Weight	95% CI		Weight	95% CI		Weight	95% CI	
s37	0.15	0.08	0.25	0.02	0	0.07	0.04	0.01	0.08
s113	0.52	0.4	0.63	0.23	0.13	0.36	0.34	0.24	0.45
s131	0.59	0.49	0.69	0.31	0.18	0.47	0.46	0.35	0.57
s207	0.8	0.71	0.87	0.53	0.38	0.67	0.71	0.61	0.8

Table 11: Weights Based on TTO Data With s1 and s37 Pooled
Resulting estimates for the weights of the TTO states from observations of the different question types and framings: DC ALE, DC RLE, Rank ALE and rank RLE data. Discrete choice combined and ranking combined contain the estimates for all the data from the discrete choice and ranking TTO questions respectively, irregardless of ALE or RLE framing. ALE and the RLE framing combined contain the estimates for the pooled data from all the ALE framing TTO questions and RLE framing TTO questions respectively, irregardless of discrete choice or ranking question type. All combined are the estimates based on all data from TTO questions. Observations for s1 and s37 were combined as s37 because of inconsistencies. 95% CI contains the lower bound on the left and the upper bound on the right of a 95% confidence interval. Estimates and confidence interval limits were transformed from probit space with the cumulative normal distribution function.

Furthermore, the influence of the ability for respondents to choose the “about equal” in the

discrete choice TTO questions was explored resulting in wider ranges and higher standard errors. These weights are displayed in Table 12. Notably another inconsistency appears between s113 and s131 in the results from the DC ALE framing.

	DC ALE			DC RLE			DC combined		
	Weight	**95% CI**		**Weight**	**95% CI**		**Weight**	**95% CI**	
s37	0.06	0	0.41	0.02	0	0.15	0.02	0	0.13
s113	0.39	0.15	0.69	0.22	0.08	0.44	0.27	0.11	0.48
s131	0.34	0.11	0.67	0.26	0.09	0.53	0.28	0.11	0.51
s207	0.75	0.45	0.93	0.54	0.28	0.79	0.66	0.42	0.85
	ALE combined			**RLE combined**			**All combined**		
	Weight	**95% CI**		**Weight**	**95% CI**		**Weight**	**95% CI**	
s37	0.04	0	0.19	0	0	0.03	0	0	0.02
s113	0.49	0.3	0.67	0.18	0.06	0.37	0.26	0.13	0.45
s131	0.57	0.39	0.73	0.28	0.11	0.52	0.41	0.24	0.61
s207	0.87	0.72	0.96	0.61	0.38	0.81	0.82	0.65	0.93

Table 12: Weights based on TTO Data with s1 and s37 Pooled Without Equal Choices
Resulting estimates for the weights of the TTO states from observations of the different question types and framings: DC ALE, DC RLE, Rank ALE and rank RLE data. Discrete choice combined and ranking combined contain the estimates for all the data from the discrete choice and ranking TTO questions respectively, irregardless of ALE or RLE framing. ALE and the RLE framing combined contain the estimates for the pooled data from all the ALE framing TTO questions and RLE framing TTO questions respectively, irregardless of discrete choice or ranking question type. All combined are the estimates based on all data from TTO questions. Observations for s1 and s37 were combined as s37 because of inconsistencies. Observations from discrete choice questions where respondents selected “about equal” were discarded. 95% CI contains the lower bound on the left and the upper bound on the right of a 95% confidence interval. Estimates and confidence interval limits were transformed from probit space with the cumulative normal distribution function.

3.4.3 Anchoring on the 0 to 1 Scale

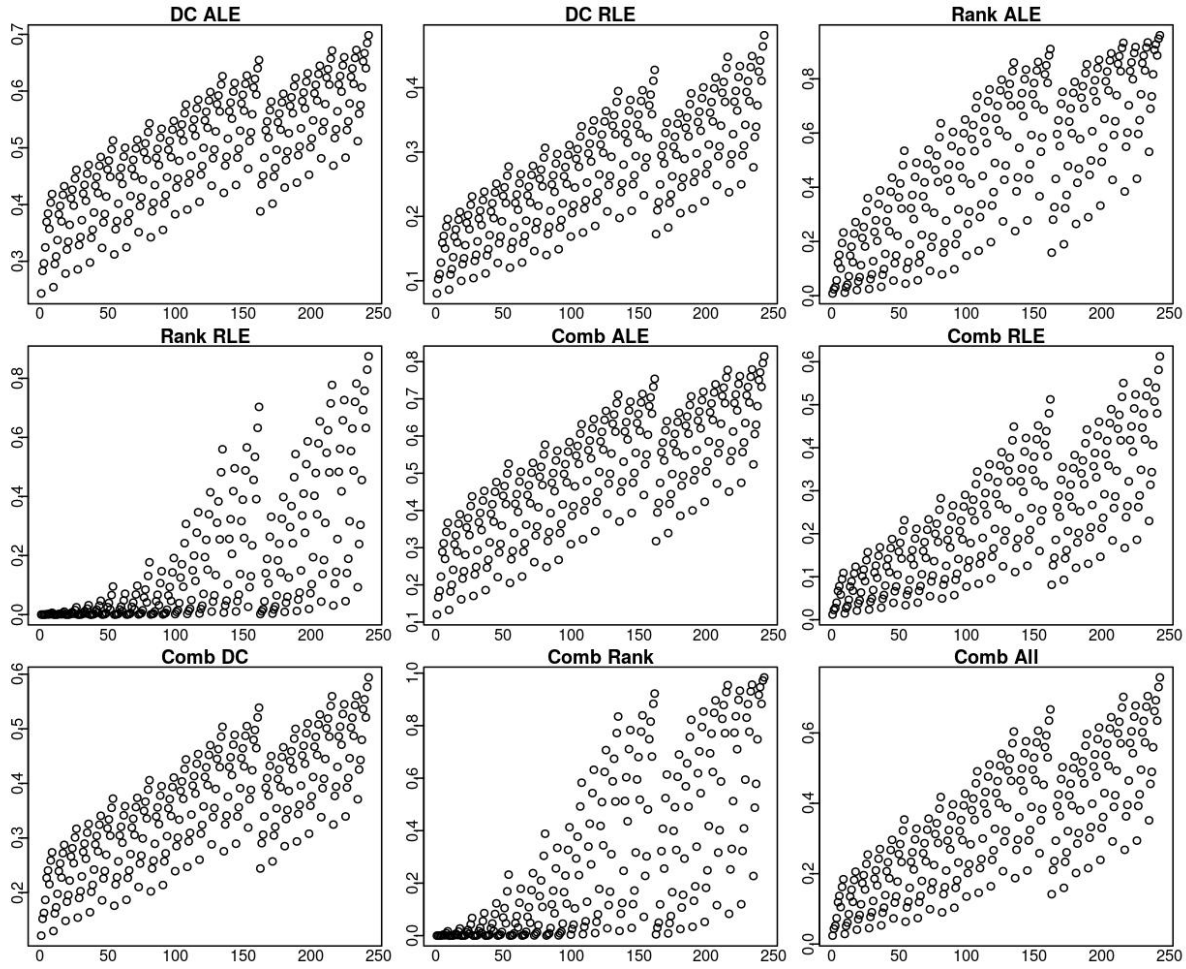


Figure 7: Reanchored Weights on the 0 to 1 Scale
Contains the distribution of the weights, according to the index of the states on the x axis from 1 to 243 and with the weight on the y axis. The titles of the graphs refer to to the weights based on the trade-off data that were used to reanchor the probit coefficients. For example, Rank ALE denotes that the distribution of weights in that graph results from reanchoring based on the weights that were calculated with the observations from the TTO ranking questions in the absolute life expectancy framing. The weights used to reanchor were calculated with all the observations and s1 and s37 were not pooled.

Figure 7 contains the distribution of the reanchored weights on the 0 to 1 scale. The range and distribution vary based on which weights have been used to estimate the slope and intercept for the linear transformation. Reanchoring with weights based on the RLE framing discrete choice results in a lower range across all states. Reanchoring with ranking RLE weights results in a compressed distribution of the lower values.

More weight distributions were explored. The exclusion of equal options resulted generally in a wider range of the distributed range (Figure 8). Pooling s1 and s37 also generally increased the range across the weights in all combinations of reanchoring weights and simultaneously also increased the frequency of weights in the lower ranges slightly.

Furthermore, the weights in probit space were also reanchored by calculating a stretch factor, for both scenarios where s1 pooled with s37 or not. The resulting distributions of weights

stretched higher upwards and were the differences between single states were more pronounced. The differences in curvature is due to For more weight distributions of weights resulting from different weights used in the anchoring mechanisms, refer to the Appendix II.

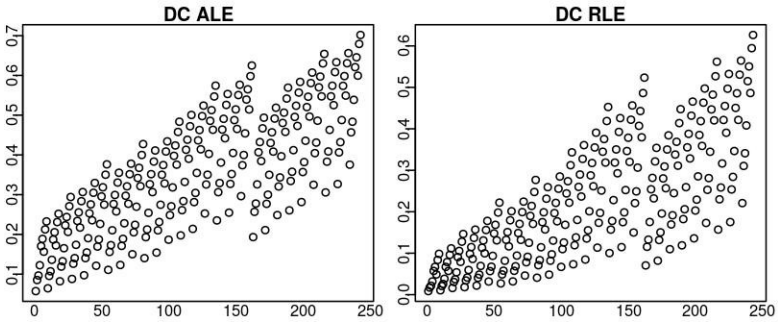


Figure 8: Reanchoring Weights Without Equal Choices

4 Discussion

4.1 Web Survey

4.1.1 Survey Format

The employed survey procedure offered several advantages. By administering all the questions digitally a high degree of flexibility was achieved with relatively low additional effort. For example, it was easy to use 3 different versions and different question formats.

Electronic screens allow sophisticated display and input formats compared to paper and true interaction between the subject and the medium. For example, the ranking questions gave visual feedback about which situations were already ranked. The ability to examine the answers of participants and react to it, for example by forcing participants to answer a question before they can continue the survey, may result in higher data quality and quantity.

On the other hand, not all members of the Swedish population may find it equally easy to leave a survey reply using electronic media, particularly in comparison to more traditional methods with pen and paper. Therefore, additional bias may be introduced when respondents who are more likely to use electronic media are overrepresented in the sample.

Moreover, while the present format may have increased the response rate by lowering the hurdles the effort required for participants compared to a format where participants would need to send back the answers with the post, it may also had an negative impact on the response rate because of higher complexity. Also, other negative associations with internet surveys may have had a negative impact, such as a perceived higher importance of a paper survey.

Another advantage is that data collection, transformation into different formats, and backing up can be automated and changes in this processes are quick and straight-forward. For example, the input of the participants does not need to be converted to a digital format before analyzing it with statistical software, as in paper-based surveys, resulting in time savings and decreased risks for mistakes.

Similarly, it is easy to handle feedback that is given by participants. It is directly available digitally. Also, the procedure where participants have the opportunity to leave an email address to receive information about the research results allows for efficient answering in terms of time and cost.

Another limitation was the lack of an explicit age confirmation before participants started to fill in the survey. As the questions that asked for age and other demographic background information were asked at the end of the survey, participants answered already all the other questions before and their data was recorded. Additionally, the age input field on the last page only asked participants to select their matching age category but was not explicitly tied to an age requirement of at least 15. Under the age of 15 parental consent is necessary by Swedish law if children are research participants (CODEX, 2016). Thus, the present survey made it possible for persons under 15 years of age to fill in and complete the survey and provide an false age category or opt-out of the age question without being aware they should not participate in the survey due to their age. On the other hand, younger participants were

not targeted as the address labels were drawn from a population sample aged 18 to 75 only. This procedure may not have free of errors though, as one participant indicated in the voluntary feedback that he was only aged 17. A future study should hence include some indication that the participation agreement is coupled to an minimal age requirement, for example in the introductory text.

Furthermore, while offering high flexibility, implementing and setting up the web survey required a substantial amount of time in addition to being prone to mistakes and thus requiring a testing. Off-the shelf commercial solution such as 1000minds (Hansen and Ombler, 2017; Sullivan and Hansen, 2017) may be more efficient and reliable, with the disadvantages of higher financial costs, less transparency and control, and decreased independence.

4.1.2 Question Order

The order of the question was chosen to maximize response rates. Demographic background questions were asked last because this type of question may be less exciting for participants to fill in and they may be more likely to complete the survey if they already answered all the other questions, given that they are aware that the this is the last set of inputs before the survey is completed (for example in the present survey where an indicator marked the progress on each page of the survey). Furthermore, the aim was to obtain more important information first; in the present case the preferences regarding different states (Teclaw et al., 2012). Another positive effect when asking demographic questions, such as education or place of birth, at a latter point in time is to avoid stereotype threat (Spencer et al., 1999) where participants may alter the response behavior based on stereotypes that they are confronted with. For example, highly educated people may value education more when they get remembered about their education level initially.

Ideally, to avoid any kind of priming effect (Kahneman, 2013, pp. 52–58), the questions where participants stated their own life situation should also have been placed at the end of the survey. However, the self-assessment question where participants rated their own life situation in each of the 5 attributes were suitable to build up the participants' familiarity with the descriptions of the attributes, because the question required the participants to more carefully read the attribute descriptions than the comparison and ranking questions, where the descriptions were shortened or the attributes only indicated by a title, and the detailed descriptions were only accessible in a separate pop-up dialog. Instead, to increase overview and the ease for participants to make TTO conclusions, the attribute levels were indicated graphically.

4.1.3 Timing

As a conclusion based on the number of received replies per day shown in Figure 3, most of the letters arrived probably 6 days after sending them out, on Tuesday the 18.4.2017 after Easter Monday, whereas some letter arrived earlier. Therefore, the postal service did not deliver all letters on the same day. Furthermore, the majority of respondents that completed the survey did so most likely in a time span of two days after receiving the letters. Regarding the response rate, the short proximity to Easter may have had some influence, to what effect however is difficult to estimate.

The mean durations (Figure 4) to answer the DC TTO questions differed significantly

between the ALE and the RLE framing, as well as between the DC TTO and ranking TTO questions. A higher cognitive burden may be the cause for these differences, as hypothesized.

4.1.4 Characteristic of Respondents

As displayed in Table 3, the sample of respondents of the web survey was not representative of the Swedish population. Particularly women, older, and well educated individuals seem to be over-represented. Although some inaccuracies regarding the way the categories were aggregated lessen the validity of the comparison (See legend Table 3), it seems likely that the sample is not representative in the compared characteristics. Thus, the external validity of the other study results is also lowered, especially results of the self assessment and slider questions. The weight calculations may have been less impacted as the TTO and the paired comparison questions did adopt an outsider perspective. This is of concern as the CALY weights should be representative for the general population and hence the here used sampling method seems inadequate.

The self assessment question showed that most of the participants rate themselves well in both on both a general scale and in all capability attributes. Social relations received the most "completely true" ratings and economic resources the least. Health received the most answers for the "Not at all" option which could be an indication for a more unequal distribution among the respondents. Remarkable is the peak in the distribution of answers at the value of 100 in the slider self assessment question, while the rest of the shows a normal-distribution like pattern around 90. This pattern could be an indication for an effect opposed to the end of scale bias and may correspond to a truncated normal distribution. On the other hand, the maximal value of 100 may had an anchor like impact where respondents that would actually have chosen values between 90 and 100 chose 100 instead because of higher salience.

Relatively few respondents choose to opt out in the demographic background questions, and slightly more in the self-rating questions. Most opt out answers were registered in the economic resources dimension, which could be an indicator that some people may be reluctant to disclose information about their wealth.

4.1.5 Response Rate

A serious limitation of the survey is the low response rate of 11.1% which fell short of the targeted 15% which was rather low in the first place. The chances of bias increase if certain subgroups of an originally representative sample are more likely to respond (concrete reference to data, gender, age, if visible). Achieving higher response rates, respectively increasing the representatives of the sample, could for example be achieved by using commercial research panels (Tsuboi et al., 2015) or by employing live interviewers (Brazier et al., 2002; Salomon et al., 2012). However, these options may be more costly, especially personal interviews, and may not completely avoid bias. Commercial panels fore example may not succeed in replicating population characteristics (Tsuboi et al., 2015).

Ways to improve the response rate within the currently used procedure may be limited. The biggest potential lies in increasing the number of page visits, that is the number of people who decide to visit the survey page based on the invitation. Ideally, an email invitation could significantly decrease the hurdle to do so. Alternatively, it may be possible to improve the design and content of the invitation letter increase the attractiveness of filling out the survey. Likewise, lowering the number of dropouts between the start page and the first question

could potentially be achieved by altering design and content of the landing page to increase attractiveness. Furthermore, gradually increasing the effort required by participants throughout the survey may decrease the dropout rate. Hence, altering the question order to start with the slider question as the first question may be beneficial, as 0 respondents stopped answering the survey during that question, and the self-assessment question demands more effort by the respondents.

The differences in the number of survey versions completed can be explained by dropped out respondents and by the verification procedure using postcodes. However, as described above, most of the respondents dropped out before the survey versions differed; only 4 dropouts were registered in questions 7 and 8 which were the only questions that differed between the versions. New survey versions were assigned based on which survey version was started the least amount of times at the point in time when a responded started the survey. The variation in the number of completed survey versions may be random, or may be influenced by factors such as higher cognitive strain in the survey versions with less completions. However, the difference in relation to the number of total survey versions administered is too low to derive a conclusion.

4.1.6 Sampling

The sampling procedure used to assign survey version to respondents may not have been fully ideal. The prior intention was to detect differences in drop out rates between survey version, without wasting responses unnecessarily. Thus the version was assigned based on the lowest number of previously assigned versions. However, since few drop outs were observed and consequently no difference between the versions was apparent, it may be better to assign survey versions based on the lowest number of completed surveys. By using such an approach the amount of data gained from different versions in relation to each other can be better controlled and equalized, which would be advantageous when looking for differences in the data and not completion rate between the different versions.

Similarly, the possible answer options in the paired comparison and the TTO questions were chosen using a random procedure. To avoid large variance in the number of observations per pair of states from the pregenerated design in case of the paired comparison questions or per state-level combination in the case of the TTO questions, the assignment procedure could take into account the number of already administered combinations and then choose from the least frequent one. Then, a targeted minimal number of observations can be better controlled to achieve an approximately equal amount of observations per combination. However, such an implementation may be complex as all previous data collected must be analyzed and no additional source of bias should be introduced.

4.1.7 Verification Procedure

The verification procedure worked reasonably well, as displayed in Figure 1. Only 7 out of 240 respondents did not provide a verifiable postcode. Two possibilities exist: Either the participant entered a postcode that was not on the address labels or the text-recognition failed to recognize the correct postcode. The postcodes are likely to be not unique as soon as a certain number of participants is reached and they only allow for 10’000 different values. In the current sample about 12% of the postcodes were not unique. Consequently, the chances increase that randomly generated numbers are equal to a valid postcode and thus making it easier to deceive the validation mechanism. Furthermore, mistakenly entered postcodes

could still be valid if they accidentally match another postcode, or more participants could submit survey answers repeatedly while still being verified. Thus, a better approach should make use of a unique identifier per participant.

We decided against a unique key in the present survey to lower the effort required by participants to complete the survey; as presumably most participants know their postcode by heart and the number of participants was low enough to make limitations due to duplicated postcodes manageable. Ideally, the survey invitations should be administered electronically, e. g. via email, and contain unique, clickable links to the survey, therefore participants would not need to enter the verification codes on their own. Another option would be to include printed QR codes (International Organization for Standardization, 2017) that contain the link to the survey, including a unique identifier, in the letters that are sent out to participants and that can be scanned with appropriate devices such as smartphones to directly open the web survey. However, it seems unlikely that all members of a representative sample would have the possibility to use this procedure; another backup option to access the survey and to manually type in the identification key would be needed.

4.1.8 Participant Feedback

Out of the 42 people who left a written feedback, 16 mentioned a positive impression regarding the survey or taking part in it, whereas 6 mentioned negative impressions regarding the survey. The other impression were neutral or not directly related to the survey per se. Several participants remarked that the questions were difficult or the instructions unclear; one feedback suggested the use of different colors for the each of the bars, corresponding to different capabilities that were used as a graphical display. Some explained choices and assumptions they made when answering the questions. Some also took the opportunity to tell about their own life situation, particularly about difficulties they endured. Furthermore, some stated own reflections about values in society and the state of affairs.

Notable other remarks include a notion that wishes for more levels to choose from when answering, presumably in the self-assessment question, or the possibility to write down specific thoughts for each question.

Interestingly, one person also wished to rate the different areas in their importance, and others reflected on the selection of the included capabilities. Additionally, one person missed an alternative corresponding to living single with a child in the question about the living status.

Another statement concerned detected spelling mistakes in the invitation letter and the survey. Furthermore, one person criticized the inability to use the back button in the browser to change an already stated choice.

In summary, the feedback was helpful because it primarily helped to raise awareness about two issues that are relevant from the participants perspective. Firstly, as anticipated, some respondents encountered difficulties with the survey format and instructions, but also to answer the questions themselves. Regarding these later difficulties, there are limits on how much questions can be simplified without compromising the ability to gain the relevant data necessary for evaluation. On the other hand, it is worth to also select evaluation methods that require lower cognitive efforts if the overall quality of estimates can be increased. For example, the TTO questions used for anchoring on the 0 to 1 scale are more challenging

compared to the pairwise comparison questions and better results may be achieved by relying more on pairwise comparisons. The inconsistencies regarding the coefficients gained from the TTO questions can also be interpreted in that regard. Moreover, difficult questions are also intended by design: In order to estimate trade-offs more efficiently, it is necessary to chose set of answers that may be more similar and cognitively demanding (Lancsar and Louviere, 2008). For example, choosing between (1,1,1,1,1) and (2,2,2,2,2) yields no viable information about trade offs, but is cognitively very easy.

While such difficulties may be harder to overcome, difficulties for participants that originate from unclear descriptions and representation should be avoided as far as possible. Given the feedback, more work is required in this direction. Mainly, the visual representation of the capability levels should be improved and the ability to invoke a pop-up dialog with the detailed descriptions should be communicated clearly in all the questions.

Secondly, the nature of the survey may be more sensitive than initially expected, because some participants seemed to strongly reflect on their own life situation while answering the survey. Indeed, some influence of participants ratings of their own life was detected in the analysis, and a connection thus cannot be excluded. Furthermore, this reinforces the need for confidential handling of data, careful wording of questions, and considering potential bias in connection to sampling.

4.2 Weight Calculation

4.2.1 Probit Regression for Paired Comparison Data

The estimated probit coefficient reflected a logical pattern. Based on the size of the coefficients, capabilities relevant for health and social relations seem to be weighted higher than capabilities surrounding economic resources, education, and occupation. These results may reflect the context of Sweden as a high-income country where basic needs around wealth are more likely to be fulfilled and thus of less importance than in other countries (Månsdotter et al., 2017b).

Similarly, a level change from 1 to 2 was estimated to be lower in all capability areas with significant estimates than a level change from 2 to 3 with the exception of education, which could be an indication of loss aversion for individuals with an capability of two, or diminishing marginal returns where an additional gains are valued less than initial gains (Kahneman, 2013; Tversky and Kahneman, 1992, pp. 278–288). The reasons for why education does not display this pattern may be that education level is high in Sweden on average and that a medium level education, for example high school, is in today's environment expected to be of similar limited value as no schooling, and a high level of education is expected. Furthermore, higher educated people were overrepresented in the sample.

A further limitation regarding the accuracy of the probit model concerns the estimated parameters. The used models only estimated the main effects but no interactions specific to state estimates or in the case of probit, between attribute levels. While the inclusion of first or higher order interactions may increase the accuracy of the models, the required sample size would also be higher and the overall complexity would increase.

4.2.2 Interval Regression for Ranking and Trade-Off Data

It is of interest regarding the weights obtained from the TTO questions, both from the comparison and the ranking versions, that the state (1,1,1,1,1) is consistently valued higher than the state (1,1,2,2,1) (Table 9). This constitutes an unexpected inconsistency, as the later state dominates the former. The same inconsistency is visible when calculating the selection frequencies of unique state-level combinations (Table 10), i.e. the fraction of times the alternative with the perfect state was preferred, deferred, or participants selected that the alternatives were approximately equal.

The reasons for this result remain unclear. A true preference for the worst state compared to (1,1,2,2,1) seems unlikely, at least in direct comparison. However, one cause could be that the state (1,1,2,2,1) was displayed in the survey questions with the third and fourth bar half filed with blue , whereas the worst state was displayed with all the bars empty, that is, filled with white. Participants may have misinterpreted this display, for example they may have judged the color white to be used to indicate a higher capability level, instead of blue, which could have bee interpreted as the "empty" filling. That is, some participants may mixed up the colors, causing them to indicate the other state as better. However, this interpretation would also require participants to interpret the bars of "growing" from top to bottom instead of bottom to top. An alternative explanation, that relies on the display mode used, would be that the display of (1,1,1,1,1), featuring only white bars, may have caused some confusion. Nevertheless, the way of graphically displaying the states needs to be improved, for example by always having a small bottom portion of each bar filled instead of completely empty bars to indicate the base level, and by using a shaded filling texture.

Furthermore, the used censor points only covered a range from 0.33 upwards for the ALE framing and 0.1 upwards for the RLE framing respectively. If weights of s1 and s37 lie far below these points, the difficulties of estimating them may increases as this range is not well discriminated by the intervals implied by the choices of respondents.

The pooling of observations for s1 and s37 to s37 was chosen to deal with this issue because the resulting coefficients for the linear transformation reanchoring may be more consistent with the estimated observed in probit space. s37 was chosen as the overriding state for s1 because, given the discussion above about the presentation of s1 which consisted of empty bars, its could be likely that inconsistencies originated in respondents evaluation of s1 as opposed to s37.

Estimates varied between the ALE and RLE framings varied. Even though the significant difference disappeared for the discrete choice TTO questions when discarding the "about equal" options, this may just be a result of the reduced sample size. While there was no significant difference between discrete choice and ranking TTO found, a difference cannot be out ruled, as the sample size may have been insufficient and the ranking questions resulted in weights that covered a wider range. Additionally, the influence of the self-rating of respondents on the responses in the discrete choice questions in the ALE framing further hints that nature of the estimates originating from the TTO questions is fragile.

When comparing the range of covered weights that results from the interval regression of the trade off data, the ALE framing consistently covers a wider range than the RLE framing which results in weights situated on the lower half of the 0 to 1 scale (Tables 9, 11, and 12). Hence, the ALE framing may be more suitable but limitations given the large variance,

plausible ranges for implied intervals, and possible bias due to differing judgment depending on the age remain.

Another possible way of wording TTO questions could be to drop fixed age numbers and let respondents simply choose between different time spans associated to the states, for example between living 10 years in state A vs 20 years in state B. However, the respondents may then themselves encounter problems of choosing the concerning age the lifespan relates to; possibly they would apply it to their own age.

Similarly, the range of the resulting weights in the interval regression based on the ranking questions is larger than the weights based on the discrete choice questions (Tables 9, 11, and 12). However, this difference decreases when equal choices are excluded from the analysis (Table 12). Thus, it cannot be excluded that ranking type questions be a valid alternative to discrete choice type of questions but the large variance again prevents more definite conclusions. Even though the cognitive effort may be perceived to be higher for a single ranking questions, the overall lower number of questions to obtain the same number of observations compared to the DC questions may be advantageous to not discourage participants.

The segregation and resulting variance between framing and question types further aggregated problems due the low sample size, and therefore the weights based on the TTO data were not precise enough. Possibly other, more traditional approaches to used in surveys to estimate weights with the time TTO or standard gamble methods, where participants answer questions to gradually determine the point of indifference between two states that corresponds to the weight (Dolan et al., 1996), could yield more usable results. Yet another possible approach is to use visual analog scales (Herdman et al., 2011).

4.2.3 Anchoring on the 0 to 1 Scale

The number of states used to anchor the probit weights on the 0 to 1 scale was only 5. Due to the low number, the mapping precision has been insufficient, and that problem was aggravated due to inconsistent state choice behavior in the TTO questions. Other efforts that used similar methodology used more states (Rowen et al., 2015; Salomon et al., 2015) . At least 10 states to be used for the reanchoring procedure seems to be the minimum (Rowen et al., 2015), but more than 10 are maybe more likely to achieve satisfying results.

Pooling s1 and s37 resulted in weight distributions with lower weights at the lower end of the scale and thus steeper distributions, while discarding the “about equal” option in the DC TTO questions resulted in weight distributions that stretched over a larger span of the 0 to 1 scale.

Using a stretch factor to map remap the weights, the distribution of the weights generally covered the whole range from the lowest weight to 1, estimates resulting from RLE and ranking data tended to accumulate at the lower and upper end of the covered range. Thus, employing a stretch factor had the advantage that the highest weights tended to be close to 1. However, since effectively only the lowest weight was anchored to the corresponding TTO weight, the weights for the other states do not align with the weights of the corresponding TTO states.

The distributions of weights differed considerably after reanchoring, depending on which weights were used to estimate the linear transformation or to calculate the stretch factor, and

whether s1 and s37 were pooled or the “about equal” options discarded Which methodological choices regarding framing and question type result in the most appropriate results is difficult to decide based on the obtained results, and may ultimately be at least partially be influenced by considerations about the desired characteristics of the values for the worst and best state and the overall distribution of the final weights (Salomon, 2003). Criteria may be how level changes and relative sizes of the paired comparison coefficients are reflected and the characteristics of the distributions, such as the range, in relation to the interval 0 to 1. For example, it may be undesirable that the best state is not assigned to 1.

4.2.4 Experimental Design

The employed experimental design based on (Street et al., 2005) is geared only towards estimation of main effects. Furthermore, it assumes that a logit model will be used in the estimation. A better design may be obtained by using other approaches such as described in (Huber and Zwerina, 1996; Reed Johnson et al., 2013) that are implemented in commercial statistical software solutions.

5 Conclusions

The response rate was below the targeted level. The low response rate gives raise to concerns about representativeness of the sample. Women, older age groups, and individuals with higher education were indeed overrepresented in the sample of the population that completed the survey. The mostly positive feedback received could be an indication that primarily individuals who are interested in the subject would fill out the survey and this interest does not seem to be evenly distributed in the population, which negatively affects the representativeness of the sample. The representativeness for weights that are used in a concrete application of CALYs should be higher and thus a different sampling method is required.

The survey procedure in form of a web survey otherwise seems to have performed reasonably well including the employed experimental design and the verification procedure based on postcodes, as indicated by the the low internal drop out rate. Increased visual attractiveness of the landing page, general design, and the invitation letter potentially could increase response rates. The feedback of respondents was mostly positive, indicating interest for the subject of societal values, although some the TTO questions were described to be difficult.

While the paired comparison question analyses with probit regression yielded appropriate and seemingly robust estimates, results between the TTO questions in the absolute life expectancy and relative life expectancy framing varied partially significantly and showed inconsistencies. Also characteristics of the participants and their life situation may be an influence factor for the response behavior in the TTO questions. A definite independent value for the worst state could not be derived. Consequently the distribution of reanchored weights differed considerable depending on which method and which observations were used for the reanchoring. These differing characteristics of the resulting distribution of weights impact the usability of the weights. Careful consideration of the desired characteristics of both the distribution of the reanchored weights and the estimates of the weights used for reanchoring should be relevant when selecting a method in a concrete application. Other methods to elicit weights on the 0 to 1 scale to reanchor the probit estimates should also be examined. More states that can be used in the reanchoring mechanism and preciser estimates thereof are necessary.

Overall, the results in this work show that the conceptualization of the CALY framework in the form of 5 different capability attributes with 3 levels each was able to capture consistent effects that provide a sound interpretation in connection to welfare. Furthermore, it can be seen as a first indication that the CALY measure may be operationizable and be used in a similar manner as existing QALY measures to effectively inform decision making, although a lot of work needs to be done, for example regarding anchoring or representatives in the calculation of weights.

Bibliography

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., Dai, B., Grothendieck, G., Green, P., 2017. lme4: Linear Mixed-Effects Models using "Eigen" and S4.

Bech, M., Gyrd-Hansen, D., 2005. Effects coding in discrete choice experiments. Health Econ. 14, 1079–1083. doi:10.1002/hec.984

Bleichrodt, H., 2002. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ. 11, 447–456. doi:10.1002/hec.688

Bradley, R.A., 1984. Paired comparisons: Some basic procedures and examples. Handb. Stat. 4, 299–326.

Brazier, J., Ratcliffe, J., Salomon, J.A., 2007. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford University Press, Cary, GB.

Brazier, J., Roberts, J., Deverill, M., 2002. The estimation of a preference-based measure of health from the SF-36. J. Health Econ. 21, 271–292.

Brazier, J., Rowen, D., Yang, Y., Tsuchiya, A., 2012. Comparison of health state utility values derived using time trade-off, rank and discrete choice data anchored on the full health-dead scale. Eur. J. Health Econ. 13, 575–587. doi:10.1007/s10198-011-0352-9

Brooks, R., Group, E., 1996. EuroQol: the current state of play. Health Policy 37, 53–72.

Brouwer, W.B.F., Culyer, A.J., van Exel, N.J.A., Rutten, F.F.H., 2008. Welfarism vs. extra-welfarism. J. Health Econ. 27, 325–338. doi:10.1016/j.jhealeco.2007.07.003

CODEX, 2016. CODEX - regler och riktlinjer för forskning [WWW Document]. URL http://www.codex.vr.se/manniska1.shtml (accessed 4.27.17).

Daly, A., Dekker, T., Hess, S., 2016. Dummy coding vs effects coding for categorical variables: Clarifications and extensions. J. Choice Model., Standalone technical contributions in choice modelling 21, 36–41. doi:10.1016/j.jocm.2016.09.005

Dolan, P., Gudex, C., Kind, P., Williams, A., 1996. The time trade-off method: results from a general population study. Health Econ. 5, 141–154. doi:10.1002/(SICI)1099-1050(199603)5:2<141::AID-HEC189>3.0.CO;2-N

Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., Stoddart, G.L., 2005. Methods for the Economic Evaluation of Health Care Programmes, 3 edition. ed. Oxford University Press, Oxford.

Fanshel, S., Bush, J.W., 1970. A Health-Status Index and its Application to Health-Services Outcomes. Oper. Res. 18, 1021–1066. doi:10.1287/opre.18.6.1021

GeoNames, 2017. GeoNames [WWW Document]. URL http://download.geonames.org/export/zip/SE.zip (accessed 5.2.17).

Haagsma, J.A., Maertens de Noordhout, C., Polinder, S., Vos, T., Havelaar, A.H., Cassini, A., Devleesschauwer, B., Kretzschmar, M.E., Speybroeck, N., Salomon, J.A., 2015. Assessing disability weights based on the responses of 30,660 people from four European countries. Popul. Health Metr. 13. doi:10.1186/s12963-015-0042-4

Haagsma, J.A., Polinder, S., Cassini, A., Colzani, E., Havelaar, A.H., 2014. Review of disability weight studies: comparison of methodological choices and values. Popul. Health Metr. 12, 1.

Hansen, P., Ombler, F., 2017. 1000Minds - Decision-making software [WWW Document]. 1000Minds. URL https://www.1000minds.com/ (accessed 5.5.17).

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., Bonsel, G., Badia, X., 2011. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual. Life Res. 20, 1727–1736. doi:10.1007/s11136-011-9903-x

Huber, J., Zwerina, K., 1996. The importance of utility balance in efficient choice designs. J. Mark. Res. 307–317.

International Organization for Standardization, 2017. ISO/IEC 18004:2015 - Information technology -- Automatic identification and data capture techniques -- QR Code bar code symbology specification [WWW Document]. URL https://www.iso.org/standard/62021.html (accessed 4.27.17).

Kahneman, D., 2013. Thinking, Fast and Slow, Reprint edition. ed. Farrar, Straus and Giroux,

New York.
Lancsar, E., Louviere, J., 2008. Conducting discrete choice experiments to inform healthcare decision making. Pharmacoeconomics 26, 661–677.
Lopez, A.D., Murray, C.C., 1998. The global burden of disease. Nat Med 4, 1241–1243.
Luce, R.D., 1959. Individual choice behavior : a theoretical analysis. J. Wiley, New York, N.Y.
Månsdotter, A., Ekman, B., Feldman, I., Hagberg, L., Hurtig, A.-K., Lindholm, L., 2017a. We Propose a Novel Measure for Social Welfare and Public Health: Capability-Adjusted Life-Years, CALYs. Appl. Health Econ. Health Policy. doi:10.1007/s40258-017-0323-0
Månsdotter, A., Ekman, B., Hagberg, L., Hurtig, A.-K., Lindholm, L., 2017b. Towards capability-adjusted life-years (CALYs) in public health and social welfare: results from a pilot study on ranking capabilities (in preparation).
McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior, in: Frontiers in Econometrics. Academic Press, New York, p. 252.
Nussbaum, M.C., 2000. Women and Human Development The Capabilities Approach. Cambridge University Press.
Ratcliffe, J., Brazier, J., Tsuchiya, A., Symonds, T., Brown, M., 2009. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Econ. 18, 1261–1276. doi:10.1002/hec.1426
Reed Johnson, F., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D.A., Bresnahan, B.W., Kanninen, B., Bridges, J.F.P., 2013. Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value Health 16, 3–13. doi:10.1016/j.jval.2012.08.2223
Rowen, D., Brazier, J., Van Hout, B., 2015. A Comparison of Methods for Converting DCE Values onto the Full Health-Dead QALY Scale. Med. Decis. Making 35, 328–340. doi:10.1177/0272989X14559542
Salomon, J.A., 2010. New disability weights for the global burden of disease. Bull. World Health Organ. 88, 879–879.
Salomon, J.A., 2003. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Popul. Health Metr. 1, 12. doi:10.1186/1478-7954-1-12
Salomon, J.A., Haagsma, J.A., Davis, A., de Noordhout, C.M., Polinder, S., Havelaar, A.H., Cassini, A., Devleesschauwer, B., Kretzschmar, M., Speybroeck, N., others, 2015. Disability weights for the Global Burden of Disease 2013 study. Lancet Glob. Health 3, e712–e723.
Salomon, J.A., Murray, C.J.L., Evans, D.B., Üstun, B.T., Chatterji, S., 2003. Health State Valuations in Summary Measures of Population Health, in: Murray, C.J.L., Evans, D.B. (Eds.), Health Systems Performance Assessment: Debates, Methods and Empiricism. World Health Organization, Geneva, pp. 409–4936.
Salomon, J.A., Vos, T., Hogan, D.R., Gagnon, M., Naghavi, M., Mokdad, A., Begum, N., Shah, R., Karyana, M., Kosen, S., Farje, M.R., Moncada, G., Dutta, A., Sazawal, S., Dyer, A., Seiler, J., Aboyans, V., Baker, L., Baxter, A., Benjamin, E.J., Bhalla, K., Bin Abdulhak, A., Blyth, F., Bourne, R., Braithwaite, T., Brooks, P., Brugha, T.S., Bryan-Hancock, C., Buchbinder, R., Burney, P., Calabria, B., Chen, H., Chugh, S.S., Cooley, R., Criqui, M.H., Cross, M., Dabhadkar, K.C., Dahodwala, N., Davis, A., Degenhardt, L., Díaz-Torné, C., Dorsey, E.R., Driscoll, T., Edmond, K., Elbaz, A., Ezzati, M., Feigin, V., Ferri, C.P., Flaxman, A.D., Flood, L., Fransen, M., Fuse, K., Gabbe, B.J., Gillum, R.F., Haagsma, J., Harrison, J.E., Havmoeller, R., Hay, R.J., Hel-Baqui, A., Hoek, H.W., Hoffman, H., Hogeland, E., Hoy, D., Jarvis, D., Karthikeyan, G., Knowlton, L.M., Lathlean, T., Leasher, J.L., Lim, S.S., Lipshultz, S.E., Lopez, A.D., Lozano, R., Lyons, R., Malekzadeh, R., Marcenes, W., March, L., Margolis, D.J., McGill, N., McGrath, J., Mensah, G.A., Meyer, A.-C., Michaud, C., Moran, A., Mori, R., Murdoch, M.E., Naldi, L., Newton, C.R., Norman, R., Omer, S.B., Osborne, R., Pearce, N., Perez-Ruiz, F., Perico, N., Pesudovs, K., Phillips, D., Pourmalek, F., Prince, M., Rehm, J.T., Remuzzi,

G., Richardson, K., Room, R., Saha, S., Sampson, U., Sanchez-Riera, L., Segui-Gomez, M., Shahraz, S., Shibuya, K., Singh, D., Sliwa, K., Smith, E., Soerjomataram, I., Steiner, T., Stolk, W.A., Stovner, L.J., Sudfeld, C., Taylor, H.R., Tleyjeh, I.M., van der Werf, M.J., Watson, W.L., Weatherall, D.J., Weintraub, R., Weisskopf, M.G., Whiteford, H., Wilkinson, J.D., Woolf, A.D., Zheng, Z.-J., Murray, C.J.L., Jonas, J.B., 2012. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. Lancet Lond. Engl. 380, 2129–2143. doi:10.1016/S0140-6736(12)61680-8

Sedlacek, T., Havel, V., 2013. Economics of Good and Evil: The Quest for Economic Meaning from Gilgamesh to Wall Street, Reprint edition. ed. Oxford University Press, Oxford.

Sen, A., 1985. Commodities and Capabilities. North-Holland, Amsterdam.

SOU, 2015. Får vi det bättre?: om mått på livskvalitet : betänkande. Fritze, Stockholm.

Spencer, S.J., Steele, C.M., Quinn, D.M., 1999. Stereotype Threat and Women's Math Performance. J. Exp. Soc. Psychol. 35, 4–28. doi:10.1006/jesp.1998.1373

Statistika Centralbyrån, 2016. Statistiska centralbyrån SCB [WWW Document]. Statistikdatabasen. URL http://www.statistikdatabasen.scb.se (accessed 5.5.17).

Street, D.J., Burgess, L., Louviere, J.J., 2005. Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. Int. J. Res. Mark. 22, 459–470. doi:10.1016/j.ijresmar.2005.09.003

Sullivan, T., Hansen, P., 2017. Determining Criteria and Weights for Prioritizing Health Technologies Based on the Preferences of the General Population: A New Zealand Pilot Study. Value Health 20, 679–686. doi:10.1016/j.jval.2016.12.008

Teclaw, R., Price, M.C., Osatuke, K., 2012. Demographic Question Placement: Effect on Item Response Rates and Means of a Veterans Health Administration Survey. J. Bus. Psychol. 27, 281–290. doi:10.1007/s10869-011-9249-y

Tesseract, 2017. tesseract-ocr [WWW Document]. GitHub. URL https://github.com/tesseract-ocr (accessed 4.27.17).

The R Foundation, 2017. R: The R Project for Statistical Computing [WWW Document]. R Proj. Stat. Comput. URL https://www.r-project.org/ (accessed 4.14.17).

Therneau, T.M., Lumley, T., 2017. survival: Survival Analysis.

Thurstone, L.L., 1927. A law of comparative judgment. Psychol. Rev. 34, 273.

Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press.

Tsuboi, S., Yoshida, H., Ae, R., Kojo, T., Nakamura, Y., Kitamura, K., 2015. Selection Bias of Internet Panel Surveys: A Comparison With a Paper-Based Survey and National Governmental Statistics in Japan. Asia Pac. J. Public Health 27, NP2390–NP2399.

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. J. Risk Uncertain. 5, 297–323.

AppendixI

Figure 9: Web Survey Screenshots
1: Landing page. 2: Self assessment question. 3: Self assessment with slider. 4: Paired comparison.

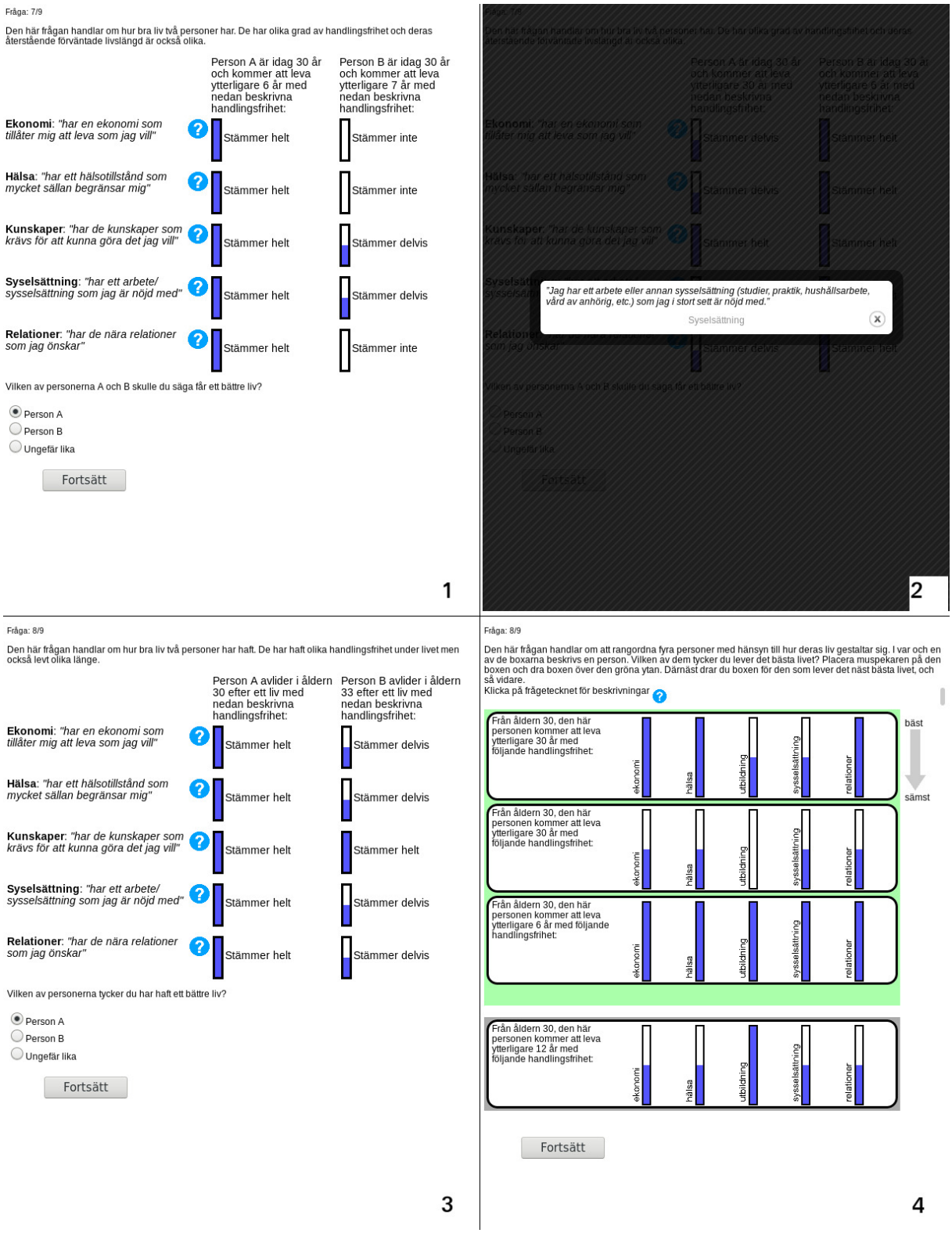


Figure 10: Web Survey Screenshots
1: TTO DC question in the RLE framing. 2: Pop up description. 3: TTO DC question in the ALE framing. 4: TTO ranking question in the RLE framing.

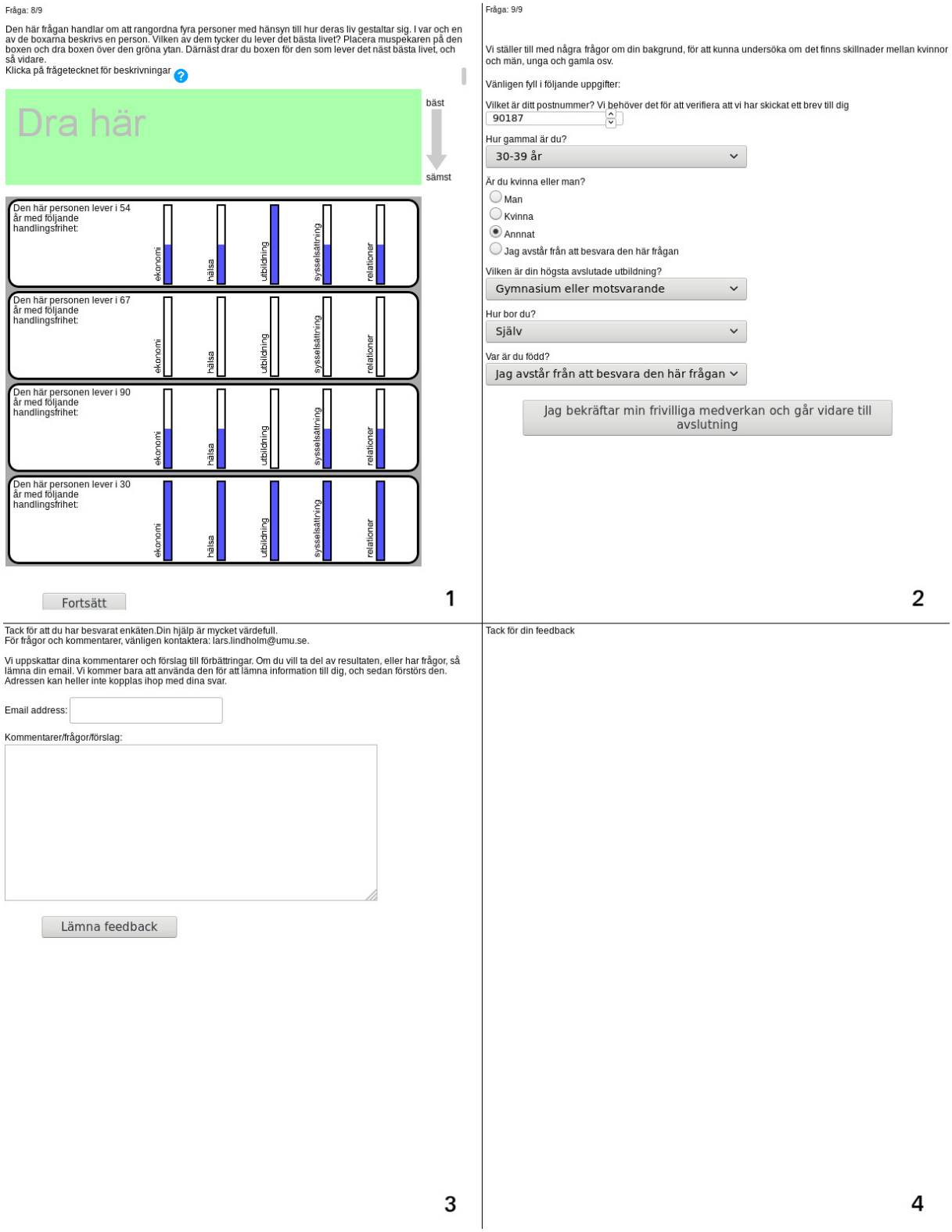


Figure 11: Web Survey Screenshots
1: TTO ranking question in the ALE framing. 2: Demographic background question. 3: Page with optional possibility to leave feedback and that indicates the end of the survey. 4: Confirmation page for the feedback.

AppendixII

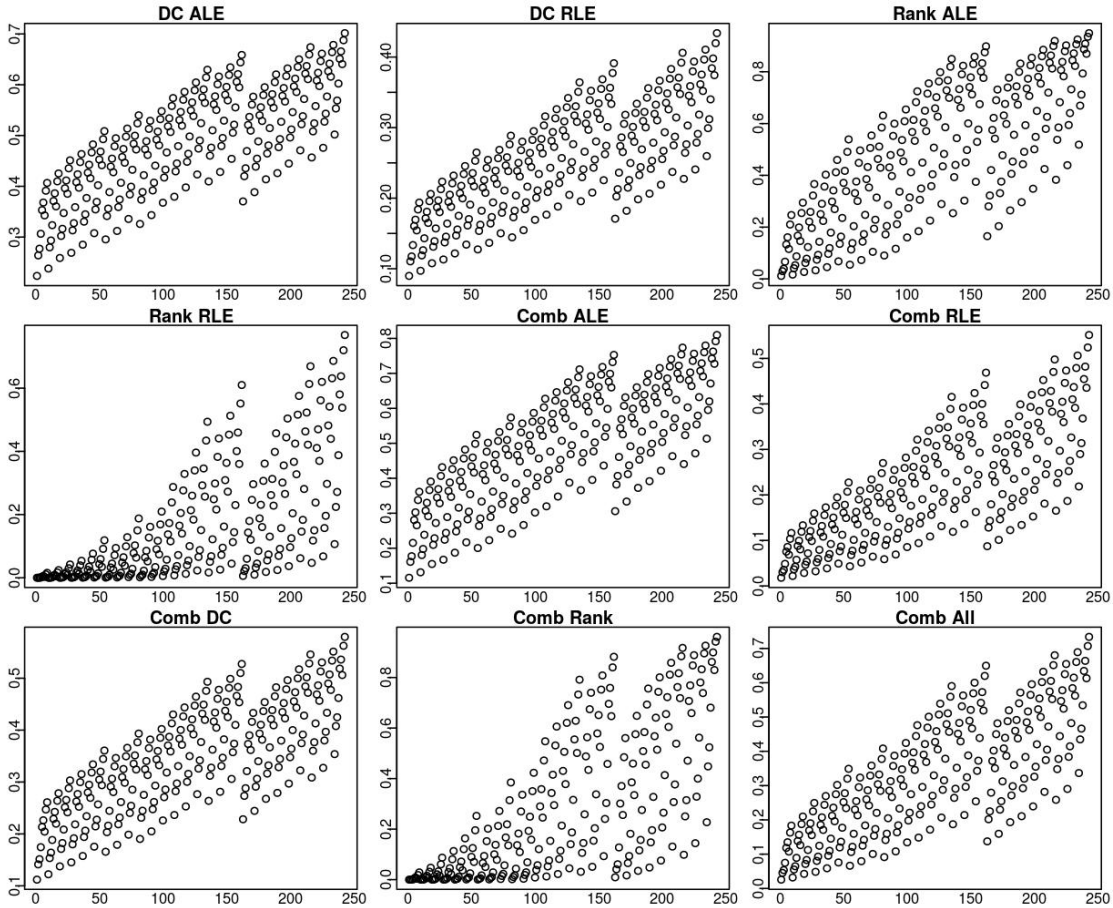


Figure 12: Reanchored Weights on the 0 to 1 Scale

Reanchoring with linear transformation based on TTO weights without pooling s1 and s37 and without excluding equal observations in the DC trade-off questions.

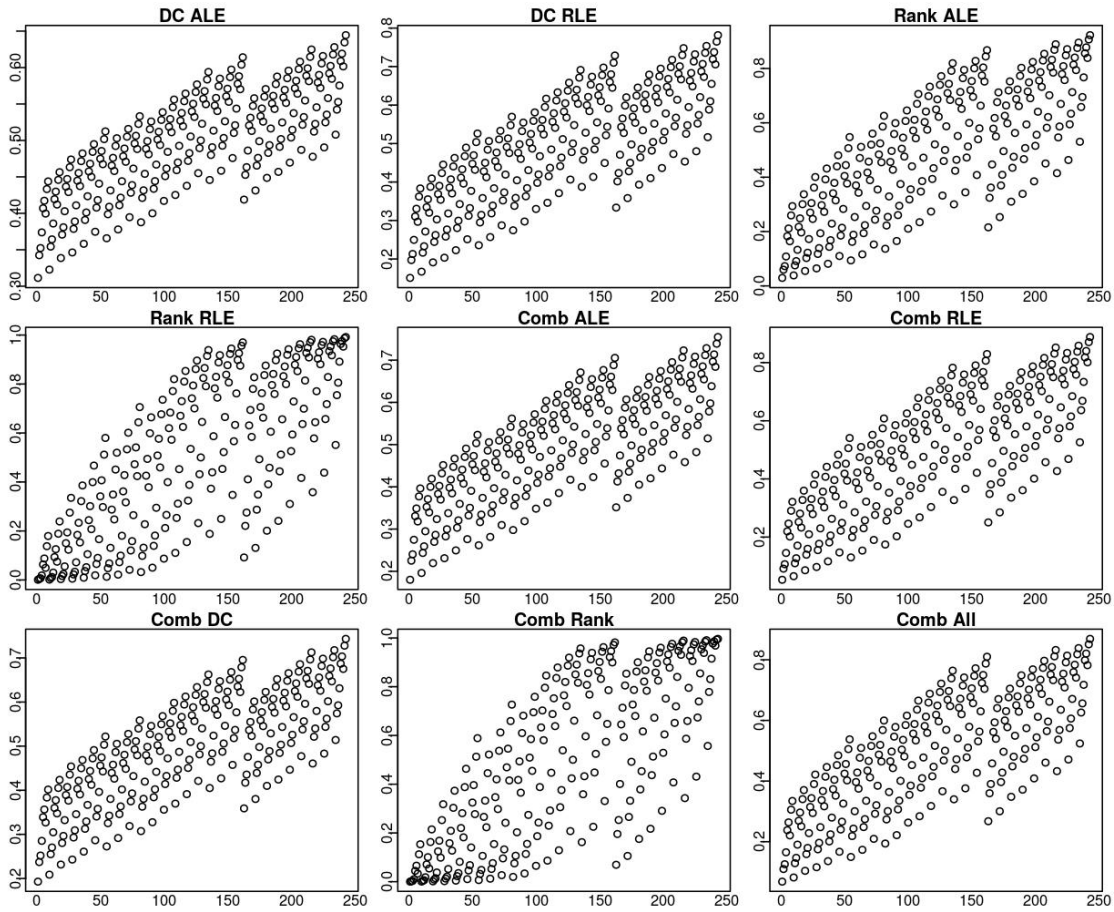


Figure 13: Reanchored Weights on the 0 to 1 Scale

Reanchoring with stretch factor based on TTO weights without pooling s1 and s37 and without excluding equal observations in the DC TTO questions.

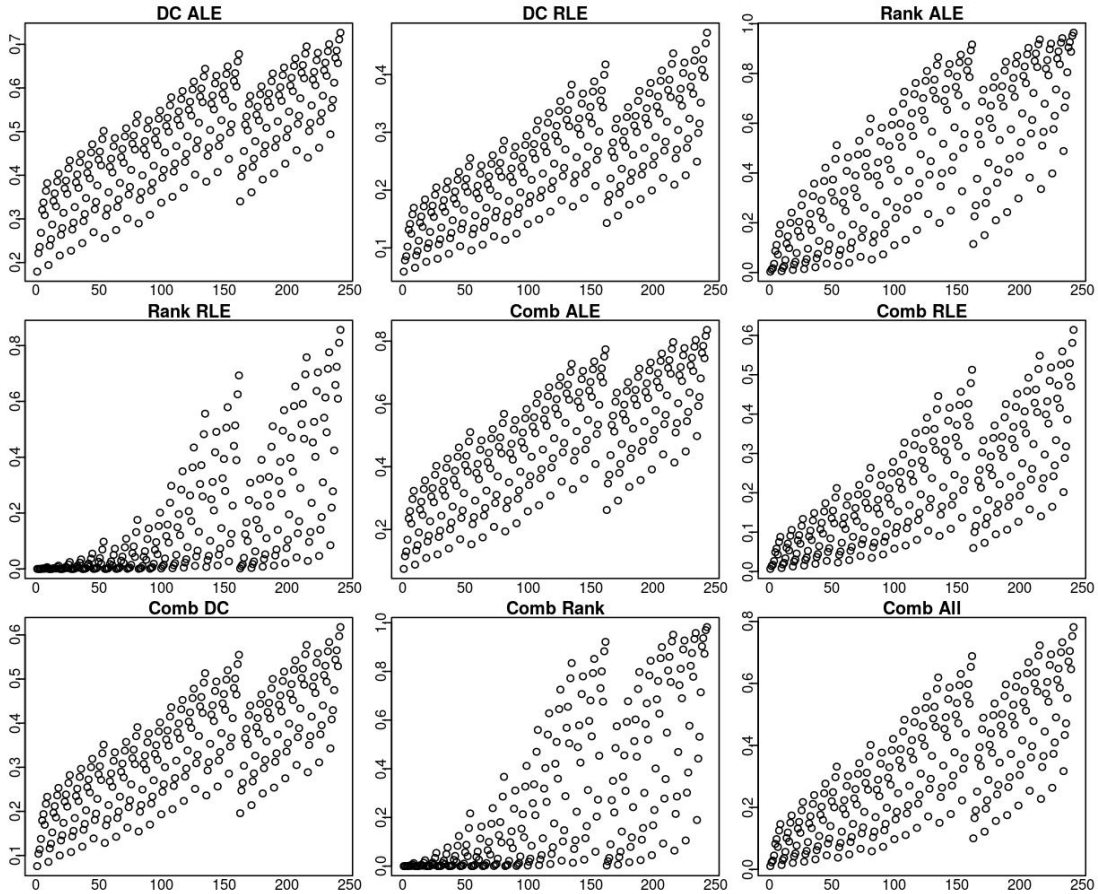


Figure 14: Reanchored Weights on the 0 to 1 Scale

Reanchoring with linear transformation based on trade-off weights with pooling s1 and s37 and without excluding equal observations in the DC trade-off questions.

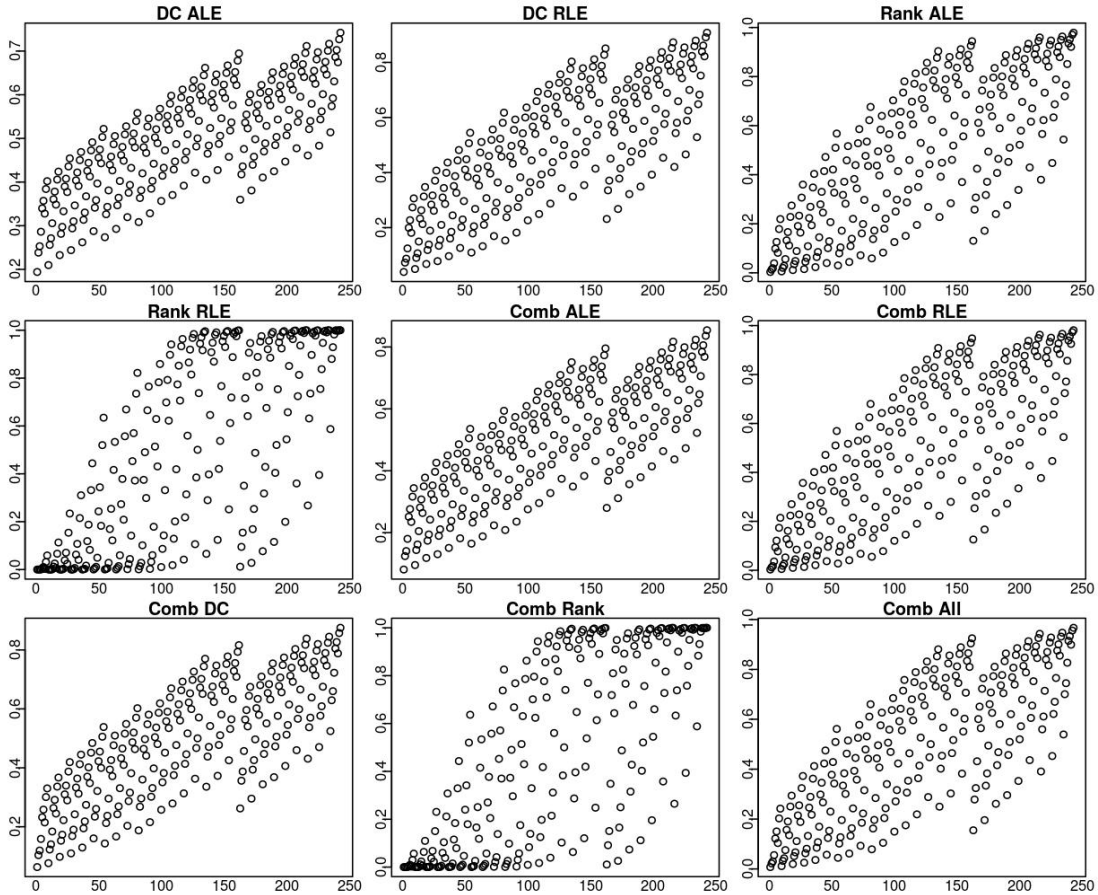


Figure 15: Reanchored Weights on the 0 to 1 Scale

Reanchoring with stretch factor based on TTO weights with pooling s1 and s37 and without excluding equal observations in the DC TTO questions.

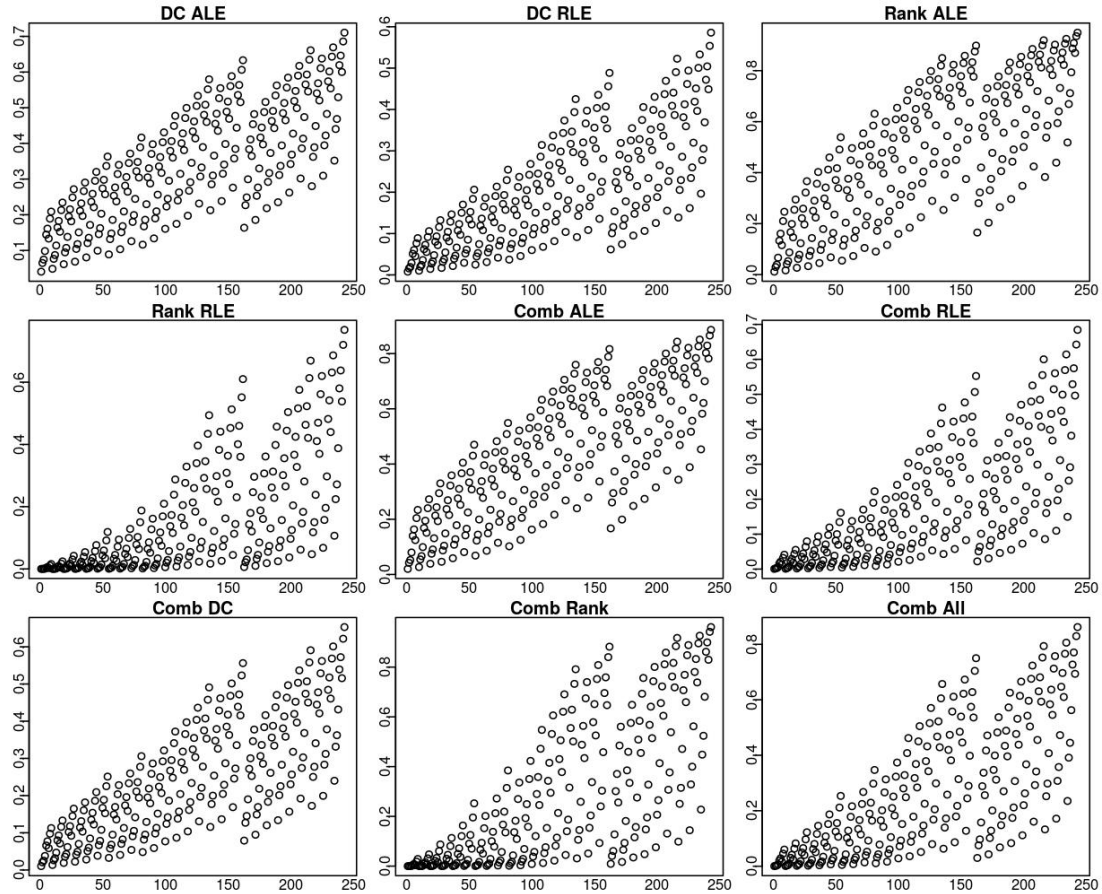


Figure 16: Reanchored Weights on the 0 to 1 Scale

Reanchoring with linear transformation based on trade-off weights without pooling s1 and s37 and with excluding equal observations in the DC TTO questions.

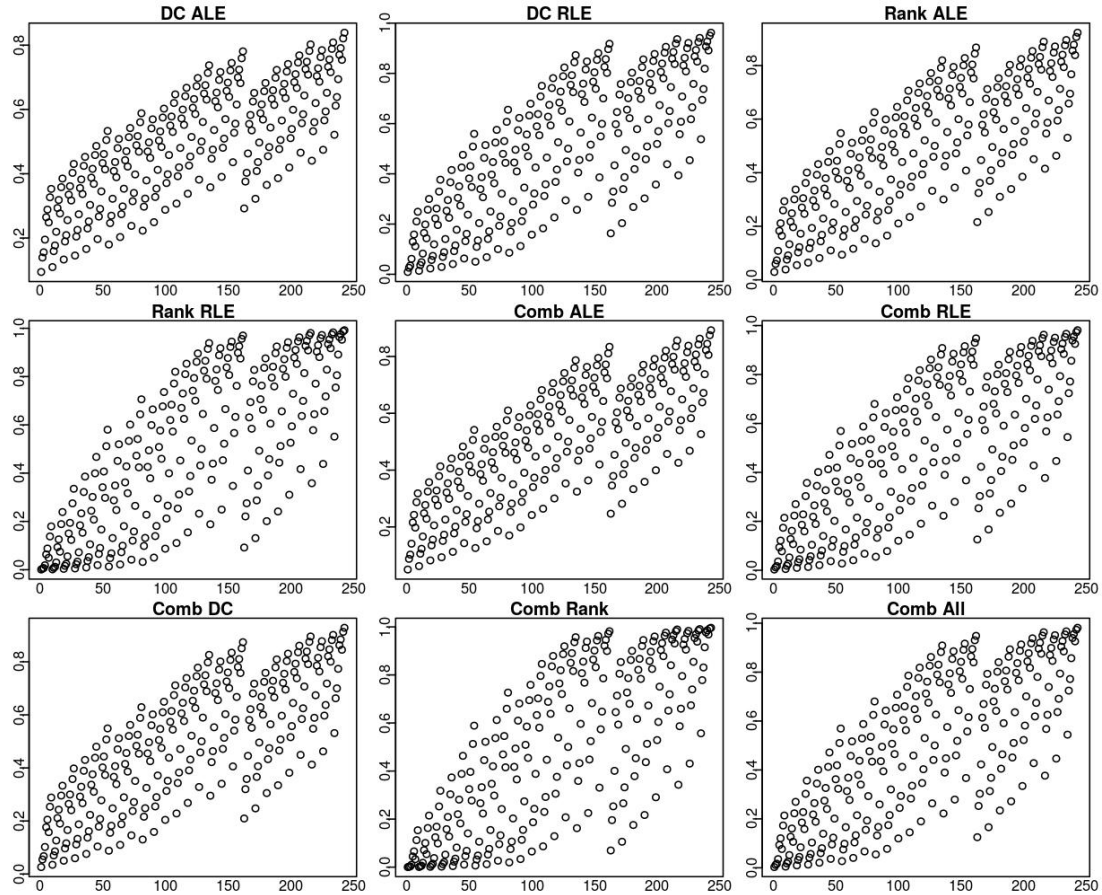


Figure 17: Reanchored Weights on the 0 to 1 Scale

Reanchoring with stretch factor based on TTO weights without pooling s1 and s37 and with excluding equal observations in the DC TTO questions.

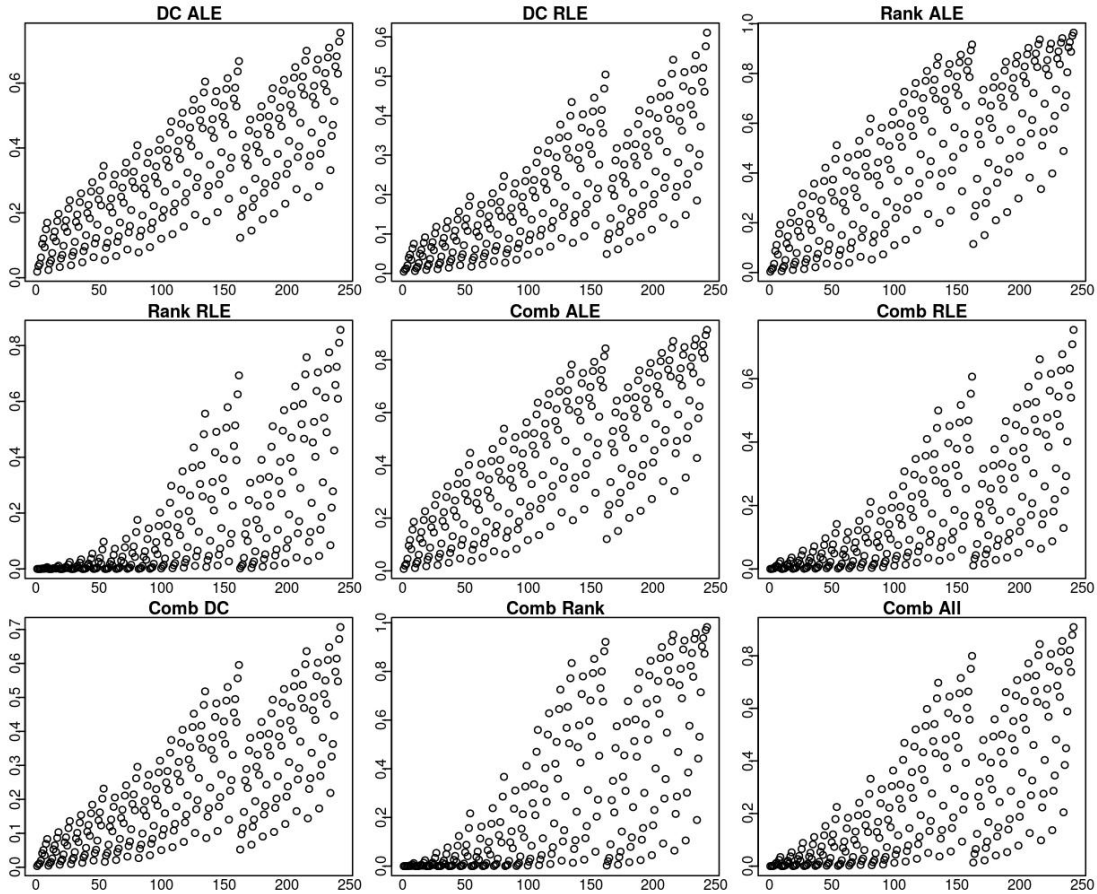


Figure 18: Reanchored Weights on the 0 to 1 Scale

Reanchoring with linear transformation based on TTO weights with pooling s1 and s37 and with excluding equal observations in the DC trade-off questions.

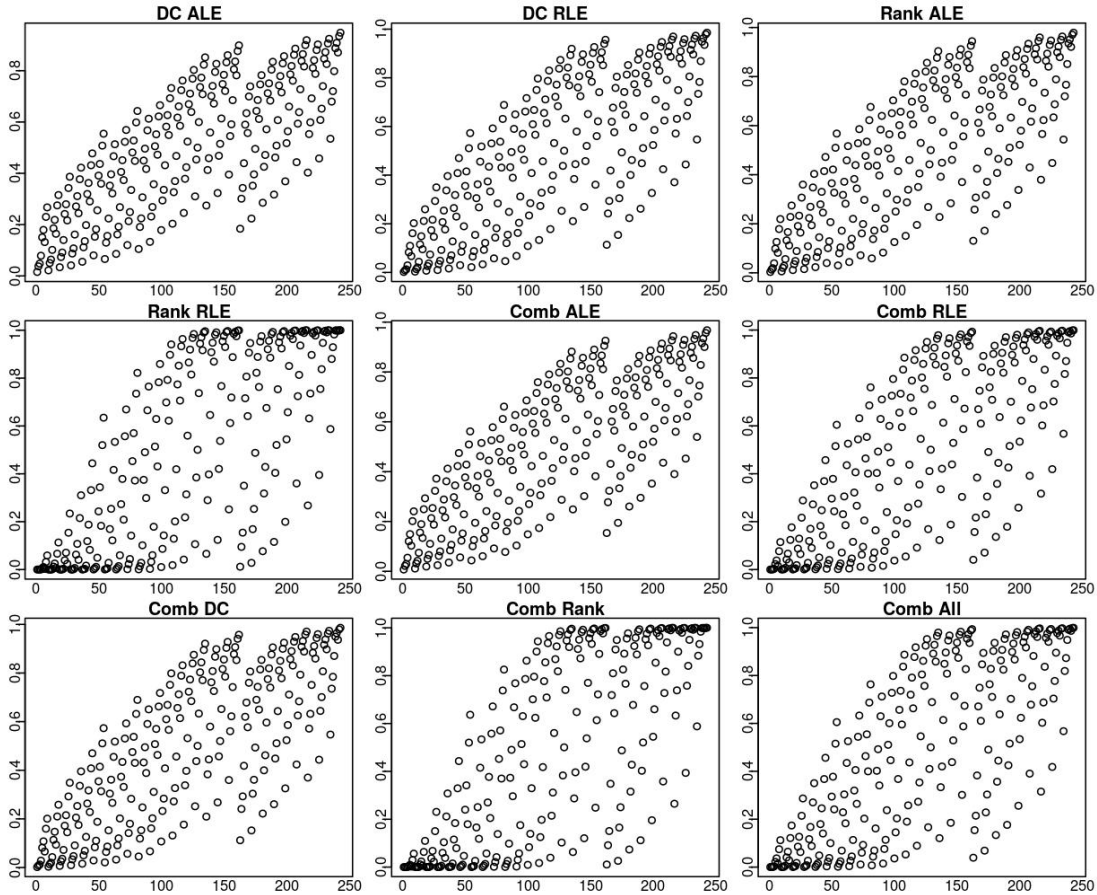


Figure 19: Reanchored Weights on the 0 to 1 Scale

Reanchoring with linear stretch factor based on TTO weights with pooling s1 and s37 and with excluding equal observations in the DC TTO questions.