Tommi Pajala

Linear model prediction for cognitive skill subgroups in Multiple Criteria Decision Making

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 3.4.2014

Thesis supervisor:

Prof. Hannele Wallenius

Thesis advisor:

Prof. Jyrki Wallenius

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Tommi Pajala

Title: Linear model prediction for cognitive skill subgroups in Multiple Criteria Decision Making

Date: 3.4.2014 Language: English Number of pages:9+82

Department of Industrial Engineering and Management

Professorship: Work Psychology and Leadership Code: TU-53

Supervisor: Prof. Hannele Wallenius

Advisor: Prof. Jyrki Wallenius

This study looks at predicting decision makers' choices in a multi-criteria decision making problem by using a linear value function model. Additionally, I investigate whether the cognitive style of subjects - measured by Frederick's (2005) Cognitive Reflection Test - is related to the predictive power of the linear model.

Subjects completed pairwise comparisons of student apartments defined by four criteria: size, price, distance to university and distance to a leisure activity. The linear model is set up using the first 10 choices and then the model tries to predict the next 10 choices.

I find that subjects notice dominance relations and make quite consistent choices, but consistency decreases towards the end of the test, most likely due to the subjects getting tired. Moreover, a majority of the subjects are not consistent with the importance of criteria they provided in the beginning of the study. This result questions the usefulness of eliciting judgments of importance abstractly, as they have only a limited connection to the choices subjects make. Importance judgments cannot be directly interpreted without knowledge of the scale of criteria values.

On average, the linear model predicts approx. 81,7 % of choices correctly and outperforms comparison models based on equal weights, and a lexicographic model. Participant provided weights were equally good at prediction, when criteria values were scaled to the 0-1 range. The efficacy of the linear model is slightly higher for linear consistent subjects than for nonlinear subjects. However, even with nonlinear subjects 77,8 % of predictions are correct. No differences in predictive power were found between cognitive style groups, gender, or study year.

Keywords: MCDM, Linear value function, Weights, Criteria importance

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Tommi Pajala

Työn nimi: Lineaarisen arvofunktiomallin ennustusvoima kognitiivisen tyylin ryhmien suhteen monikriteerisessä päätöksenteossa

Päivämäärä: 3.4.2014 Kieli: Englanti Sivumäärä:9+82

Tuotantotalouden laitos

Professuuri: Työpsykologia ja johtaminen Koodi: TU-53

Valvoja: Prof. Hannele Wallenius

Ohjaaja: Prof. Jyrki Wallenius

Tässä työssä tutkitaan neljän kriteerin päätösongelmaa ja sitä, voidaanko koehenkilöiden valintoja ennustaa lineaarisella hyötyfunktiolla. Lisäksi tutkitaan, onko koehenkilöiden kognitiivisella tyylillä merkitystä ennusteiden onnistumisen tai henkilöiden lineaarisen konsistenssin kannalta.

Koehenkilöt tekevät parivertailuja hypoteettisista opiskelija-asunnoista. Asunnot on määritelty neljän kriteerin avulla: koko, hinta, etäisyys yliopistosta ja etäisyys harrastuspaikasta. Ensimmäiset 10 vastausta virittävät lineaarisen mallin, josta saadaan painokertoimet neljälle kriteerille. 10 viimeistä vastausta pyritään ennustamaan kertomalla kriteerien arvot näillä painokertoimilla.

Huomataan, että koehenkilöt vastaavat suhteellisen konsistentisti, joskin he tekevät lopun kontrollikysymyksissä virheitä. Tämä tapahtuu luultavimmin keskittymisen herpaantumisen vuoksi. Ennen kaikkea suuri osa koehenkilöistä vastaa epäkonsistentisti verrattuna kokeen alussa antamaansa kriteerien tärkeysjärjestykseen nähden. Tulos kyseenalaistaa kriteerien tärkeyden ja painokertoimien suorasta kysymisestä saatavan hyödyn, koska kriteerien tärkeys on vain heikosti yhteydessä henkilöiden tekemiin valintoihin. Kriteerien tärkeys ei ole tulkittavissa ilman tietoa kriteerien arvojen skaalauksesta.

Lineaarinen malli ennustaa keskimäärin 81,7 % valinnoista oikein. Mallin ennusteprosentti on parempi kuin käytetyillä vertailumalleilla. Kriteerien tärkeyteen perustuva malli suoriutui ennustamistehtävästä yhtä hyvin, kun kriteerien arvot skaalattiin välille 0-1. Lineaarinen malli ennustaa jonkin verran paremmin lineaaristen kuin epälineaaristen vastaajien valintoja. Eroja kognitiivisen tyylin, sukupuolen tai opintovuoden perusteella ei havaittu.

Avainsanat: MCDM, lineaarinen arvofunktio, painot, kriteerien tärkeys

Preface

Well, this is truly the end of an era. Six years of studies here at Aalto University culminate in this thesis. As I write these last lines and reflect on the previous years, I can only say that this is a fitting end, a strong finish to a period of hard work and learning, but also good times and friendships.

This thesis required a long and arduous process, with several revisions and hours of scratching my head. Fortunately, I didn't have to do it alone. First and foremost I want to thank Professors Jyrki Wallenius and Hannele Wallenius for their fantastic support and advice. Whenever I faced a dead end and saw no way out, you always somehow came to the rescue. Your faith in my ideas and perspective was a major part in why I enjoyed this process so much!

Also, I am extremely grateful to all the peer support I had. Joosef, your academic prowress was instrumental in helping me think straight. Additionally, the fact we managed to drag each other out from the office every once in a while was probably for the good, for both our sakes. Ilkka, Tomi, Lauri, Outi, Terhi, you all pushed and encouraged me to work diligently, yet you also ensured my sanity with the ever so militaristic afternoon break schedule.

However, these past years haven't been just about the thesis. All the friends from the past years deserve a very special thank you. To fit in all of you in this one preface is impossible, so I hope it suffices to say: thank you all, especially Niko P., Rosa L., Satu P., Juho L. and Pyry S. To have you as friends is to be a truly lucky person.

Mum and dad, at this point I'm finally starting to realize how important your example has always been to me. You have always encouraged me to follow my dreams and face the challenges of life head on. You have always believed in me and my capabilities. Without dad's math and sport tutoring I wouldn't have achieved any of this - or be as fit as I do. Without mum's support I wouldn't have had the courage to go to Austria, or realized how much I can learn if I just do my best.

Last but definitely not least: Sonja. You took me to the other side of the world. You showed me that it's possible to be smart and driven, but also to simultaneously be a good person. Day by day, you help me to become a better version of myself. I will always be in awe of your energy and capacity to help those around you. You are the most courageous person I know, and with you, I'm sure there is nothing we can't face head on.

Otaniemi, 4.4.2014

Tommi Pajala

Contents

List of Figures

List of Tables

Symbols and abbreviations

Symbols

λ_j	the weight of attribute j in a linear value function
X	option in choice set
x_j	attribute j of a choice option

Operators

$a \succ b$	a is strictly preferred to b

Abbreviations

AHP	Analytic Hierarchy Process
CRT	Cognitive Reflection Test
CP	Clopper-Pearson estimation
DSS	Decision Support System
LEX	Lexicographic Decision Strategy
MAUT	Multiattribute Utility Theory
MCDM	Multiple Criteria Decision Making
ML	Maximum likelihood

1 Introduction

Decisions are an integral part of our lives. We face daily innumerable situations that demand a decision: should I do this or that? Which shirt should I put on? Should I go to the movies or for a run? Some of our decisions can be solved by evaluating just one criterion: the typical student might choose her or his lunch based solely on the price. However, in a lot of decisions we have several criteria that influence the decision. Situations such as these are called Multiple Criteria Decision Making (MCDM). This thesis is concerned with one such case: choosing a student apartment.

Decision making refers to a situation where the decision maker is faced with a need to answer the questions: "what to choose?". Typically, answering this question is done by choosing the "best option" among some possible alternatives. In this study, the "best option" is defined conventionally: option A is the best option if and only if no other option in the choice set is preferred to A according to the decision maker.

MCDM refers to the situation where a decision maker (DM) has to evaluate an option space regarding several criteria. Buying a car can be thought of as an MCDM problem, with the particular criteria being price, fuel consumption, size, and the colour of the car, for example. Some options clearly dominate other options, for example, you might find the same car from another salesman for a cheaper price. In that case, the cheaper car is clearly much better and dominates the more expensive car.

Usually MCDM problems do not have one solution that would dominate all the other options. More often there is an efficient frontier, which dominates other options, but the options on the frontier do not dominate each other. The DM then has to evaluate which trade-offs he is willing to make. For example, consider a case in which I am bying a car. Assume that I think a cheaper price is better than a higher one, and that more seats are better than less. If two cars are otherwise alike but one costs 15000 and has two seats, but the other one costs 25000 and has four seats, which one should I choose? It clearly depends on the trade-off I am willing to make between price and seating capacity (assuming all other criteria are equal).

An MCDM problem is called discrete when the number of potential solutions, ie. options the DM can choose from, is limited. This study considers a discrete problem. This study also features a case where all the criteria are cardinal, ie. they can be measured on a numerical scale. This is not always the case, criteria in an MCDM problem can also be ordinal (for example, [bad, average, good] is a scale in increasing order of preference).

There are several streams of theory available for solving MCDM problems. The best known streams and methods include the Analytic Hierarchy Process (AHP), Multiattribute Utility Theory (MAUT), conjoint analysis, French school, interactive mathematical programming, and several others. Compensatory models estimate all criteria simultaneously so that all criteria contribute to the goodness of the solution. This means that a solution that is the worst on some criterion can still be picked as the best one, if the other criteria have values that are high enough to compensate.

Clearly, such compensations mean that there has to be some way of doing tradeoffs between criteria. A common question regarding this is, how is it possible to analyze tradeoffs when the traded criteria are measured in different units. For example, if one criterion is measured in apples and the other in oranges, how can we compare the levels of those two criteria with each other? In MCDM the MAUT model solves this by treating the weights as scaling constants that transform the criterion value to common utility (Choo et al., 1999). This study features a similar approach.

Historically, linear models have performed remarkably well as models predicting subject's choices (Dawes and Corrigan, 1974; Dawes, 1979). A linear model also proved efficient in predicting pairwise comparison choices when the weights were calibrated with a subset of the choice data (Korhonen et al., 2012). The purpose of this study is to extend the previous research by showing how the increase of criteria from two to four influences the power of predicting subjects' choices with a linear model.

A particular problem in using a linear model is obtaining the values of the weights. It is unclear what the weights actually stand for (Choo et al., 1999; Weber and Borcherding, 1993). Many studies simply ignore this problem, assuming that the interpretation of the weights is clear and shared between the analyst and the subject, which makes eliciting weights directly a feasible option. However, this procedure has been recently questioned (Korhonen et al., 2013).

Korhonen et al. (2013) studied a situation where subjects performed pairwise comparisons between options with two criteria. They concluded that for approximately 30% of the subjects the weights of the value function had a different ordinal relationship than the importance of criteria stated by the subject. This essentially means that for those subjects the weights either do not imply anything about the importance of criteria, or that the subjects changed their opinions about the importance after having stated them. The purpose of this study is to show how increasing the number of criteria from two to four affects the relationship between criteria importance and linear weights.

Moreover, this study is concerned with trying to delineate the classes of subjects whose choices can be explained with the linear model. More exactly, I will look at how the cognitive style of subjects is related to linear consistency and choice predictability of subjects. The motivation for this is that perhaps those with a more analytic and reflective cognitive style either use a decision strategy that is closer to a linear model, or that they are better at predicting their own preferences in the beginning of the experiment. Therefore, I hypothesize that the Cognitive Reflection Test (CRT) score can be used as a proxy variable, ie. that there will be differences between the CRT groups in terms of linear consistency or choice predictability.

To summarize, this thesis is concerned with the following questions:

- What, if any, is the connection between provided judgments of criteria importance and the weights of the linear function?

- Can we predict subjects' choices with a linear value function in this four-criteria problem?

– Is the CRT score connected to the linear consistency or the predictive power of the linear model?

The question regarding the interpretation of the ranking of criteria is highly relevant. It is quite common to hear someone analyze a multiple criteria problem and say that one criterion is more important than some other. However, it does not seem clear what this means. If a subject says that "criterion A is more important than criterion B", what message is she trying to convey? Clearly some message must be behind the utterance, as subjects find it quite easy to rank criteria ordinally. A previous study (Korhonen et al., 2012) did not find a relationship between the rank order of the criteria and the weights of the linear function. This study is meant to explore if the relationship changes with a different problem case and with a larger set of criteria.

Secondly, this study is aimed to explore the feasibility of predicting subjects' choices with a linear value function in a case where not all criteria are independent of each other. In our case the problem is that of choosing a suitable student apartment. The criteria are size, price, distance to University and distance to leisure activity location. The distance to University and distance to leisure location can both be thought to reflect the availability and convenience of public transport from the apartment, and are in that sense related to each other.

Third, this study tries to look at why the linear value function model is better at predicting the choices of some subjects than of others. My hypothesis is that this is related to the cognitive style of the subject, ie. the fact whether the subject is more analytic or heuristic. I measure the cognitive style with Frederick's (2005) Cognitive Reflection Test and analyze whether this has any relation to the predictive power of the linear model or the linear consistency of the subject.

2 Theoretical background

A multiple criteria decision making problem has the following qualities: there are more than one criteria, which define the solution space. Additionally, the set of possible solutions most commonly does not have a single best option, ie. an option that would be better than all the other options on every criterion. This means the decision maker has to somehow determine, which criteria are more important for him and choose the best alternative so that it maximally fulfills his preferences.

A classification of multiple criteria problems can be given with the following distinguishing concepts (Korhonen et al., 1992):

Discrete vs. continuous A problem is called discrete when the number of elements in the set of possible solutions is countable. When the number of solutions is uncountable, the problem is continuous. For example, our example of choosing an apartment from two options is discrete: there are only three possible choices (choosing A,choosing B, or being indifferent). In contrast, a continuous problem might be the production of concrete in a factory.

Deterministic vs. nondeterministic A problem is called deterministic, when the values of the criteria for an alternative are known with certainty. When there is uncertainty regarding the criteria values, the problem is nondeterministic or stochastic.

In this study, the choosing of student apartments is a discrete deterministic problem. Next, let us look at the more exact mathematical representation of this problem.

2.1 MCDM from a mathematical perspective

Even though there are several alternative methods for solving an MCDM problem that differ considerably, there is nevertheless a shared problem representation that these different methods have in common.

An MCDM problem can be defined mathematically so that the DM has a set of alternatives, which is denoted by A. Most commonly A is finite, and henceforth it will be treated as such. These alternatives $a \in A$ can be related to a set of criteria, so that each alternative is a vector $[x_1(a), \ldots, x_k(a)]$, where $x_j : A \mapsto R, j = 1, \ldots, k$. The DM is faced with the following problem: how to choose an alternative $a \in A$ so that the vector $[x_1(a), \ldots, x_k(a)]$ is maximized?

A mathematical representation of this situation is the following:

$$
\begin{aligned}
&\max q \text{ so that} \\
&q \in Q = [f(x) | x \in X], \\
&\text{where} \\
&f(x) = [f_1(x), f_2(x), \ldots, f_k(x)]
\end{aligned}
$$

is the k-dimensional vector of criteria functions, Q is the set of available alternatives, x is a vector of decision variables and X is the proper set of the values of decision variables. The solution depends on the subjective preferences of the DM, but conceptually we can think that every DM is trying to maximize a value function $u = u(q)$ that

is monotonically increasing with respect to all criteria $f_j(x), j = 1, \ldots, k$. Clearly, if the value function u were known, the problem could be reduced to maxizing just one objective function.

That is, of course, hardly ever the case, and a substantial literature has emerged around this problem. There are several ways that one can try to either solve or circumvent this problem. Korhonen, Moskowitz & Wallenius (1992) present in their review three basic approaches to the problem:

1. Assume the existence of some value function and assess it explicitly
2. Assume the existence of some value function, but do not assess it explicitly
3. Do not assume any value function

In this study, the approach is to assume that a linear value function to some extent can be used to represent the preferences of the subject. I do not claim that the subject's value function would necessary in itself be linear, only that the linear value function is a good enough approximation. In this respect, the approach of this study falls into the second category.

2.2 Definitions

In this chapter I will introduce the basic concepts and proofs necessary as background information for the study.

This study focuses on a discrete, finite, and deterministic multiple criteria evaluation case. In the experimental setting the DM compares two alternatives with respect to four criteria. Let us define the set of alternatives as

$$S = \{X_i \in \Re^4, \quad i \in N = \{1, 2\}\}$$

For the purposes of exposition in this theory section and without loss of generality, I assume more for every criterion is better.

Dominance in the set is defined in the conventional way as follows.

Definition 1. A vector $X^* \in \Re^4$ is dominated if and only if there exists another $X \in \Re^4$ so that $X \geq X^*$ and $X \neq X^*$. From here on, I shall use "$\succ$" to indicate the relationship of preference, meaning that when $X \succ X^*$ then X "is preferred to" X^*

Definition 2. Any strictly increasing function $\nu : \Re^4 \to \Re$ is called a value function.

Definition 3. Assume $P = \{(X_r, X_s) | X_r \succ X_s, r, s \in N\}$ consists of preference information available about alternatives $X_i, i \in N$.

Definition 4. If $\nu(X_r) > \nu(X_s)$ for all $(X_r, X_s) \in P, \nu$ is said to be consistent with P. If there exists a weight vector $\lambda > 0$ such that $\sum_{j=1}^{4} \lambda_j x_{rj} > \sum_{j=1}^{4} \lambda_j x_{sj}$, for all $(X_r, X_s) \in P$, then a linear value function $\sum_{j=1}^{4} \lambda_j x_{ij}, i = 1, 2$ is consistent with the DM's preferences in P.

Lemma 1. *The consistency property of a linear value function is invariant under a linear transformation of the criteria.* $x_{ij} \to \alpha x_{ij} + \beta_j, i \in N, j = 1, 2 \ldots 4$, *and* $\alpha > 0$.

Proof: See Appendix A.

Lemma 2. *If a linear value function* $\sum_{j=1}^{4} \lambda_j x_{ij}$ *with vector* $\lambda \geq 0$, *is consistent with the DM's preferences, then the linear value function* $\sum_{j=1}^{4} \mu(\lambda_j \alpha x_{ij} + \beta_j), i \in N$ *and* $j = 1, \ldots 4$, *where* $a_j > 0, \mu_j = \frac{\lambda_j}{\alpha_j}$, *is consistent with all preferences* $(X_r, X_s) \in P$.

Proof: See Appendix A.

Corollary 1. *If there exists no vector* $\lambda > 0$ *such that the linear value function* $\sum_{j=1}^{p} \lambda_j x_{ij}, i = 1, 2 \ldots n$, *is consistent with preferences* $(X_r, X_s) \in P$, *then there exists no vector* $\mu \geq 0$ *such that the linear function of the form* $\sum_{j=1}^{p} \mu_j(\alpha_j x_{ij} + \beta_j), \alpha_j > 0$, *is consistent with all* $(X_r, X_s) \in P$.

Proof: Follows directly from Lemmas 1 and 2.

What corollary 1 essentially tells us is that if we cannot find a linear value function that is consistent with the subject's preferences, then we cannot find such a function even by scaling the criteria.

2.2.1 Efficiency

We can assume that a rational DM is only interested in efficient alternatives. An alternative x^o is efficient, if $\nexists x \in X$ so that$f(x) \geq f(x^o)$ and $x \neq x^o$. Clearly the set of efficient alternatives is a subset of all possible alternatives.

Rarely will the formation of the efficient set result in a situation where only one alternative remains. If this happens, it is straightforward to pick the only efficient alternative, because it dominates all other solutions. If the efficient set contains several alternatives, the DM has to make choices regarding which option to choose. The problem can be approached via several possible routes. For example, the DM can define aspiration levels and consider only options that fulfill these aspiration levels, which is analoguous to constraining the choice set. Alternatively, the DM can average different criteria together, or pick the alternative with the least worst value on some criteria, or simply pick the alternative with the best value on some criterion. Quite clearly, at this point alternative solution strategies abound. The review of alternative methods is outside the scope of this thesis. For comparison purposes, I will include the lexicographic choice model as one of the comparison models. The lexicographic model is presented in section 2.6. Otherwise, the methodology used in this study simplifies the problem by assuming that the DM has a linear weighted value function. Intuitively, this is compelling. It would be plausible that students put more emphasis on rent than distance to centre, for example. In the next section, I will review some of the literature regarding linear models.

2.3 Linear models

The usage of linear models is quite commonplace in Multiple Criteria Decision Making, due to the ease of use and the simple form of the value function. These linear

models imply a value function, which takes the form

$$f_v(x) = \sum_{j=1}^{p} \lambda_i x_{ij}$$

where p is the number of criteria and $i = 1, 2 \ldots n$ is the index of a particular alternative. The dimensionality of the linear value function can be thought of so that the weight is not dimensionless but has a unit of $\frac{1}{u_x}$, where u_x is the unit of the attribute x (Choo et al., 1999). This way we can avoid calculating apples and oranges together, as each product of weight and criterion results in dimensionless utility.

The linear model can be either normative or descriptive. A normative model is meant to aid the decision maker to reach a good decision, whereas a descriptive model is meant to describe how a decision maker actually behaves. Historically, normative linear models have had many successes in decision making (Dawes and Corrigan, 1974). It seems that humans are quite adept at recognizing the relevant criteria for a good decision, but lack the cognitive ability to properly integrate information necessary to make the right decision. In cases like these, a linear model can handle the integration and possibly produce a better result than a human DM.

However, linear models do not enjoy the same degree of success in all circumstances. According to Dawes & Corrigan (1974), the linear model is likely to fare well when each output variable has a conditionally monotone relationship to the criterion. According to Krantz, Luce, Suppes & Tversky (1971) this is a combination of independence and monotonicity. In simple terms, this means that criteria are not affected by the values of other criteria and that the ordinal relationship of variables is monotone (a value with a higher ordinal ranking has a higher numerical value) (Krantz et al., 1971). This leads naturally to the question what defines a good set of criteria for a linear value function?

2.3.1 Definition of the criteria set

Keeney and Raiffa (1976, p. 50) have proposed the following requirements for a good set of criteria:

1. Completeness
2. Operational
3. Decomposable
4. Nonredundancy
5. Minimum size

Completeness means that the level of achievement of the overall objective can be adequately explained in terms of the attributes, ie. all the most relevant attributes belong to the set. For example, this means that if we analyze the act of buying a

car but leave out the fuel expenditure attribute, the set is not going to be complete. Operational attributes are attributes, which are meaningful for the decision maker and can help in making a good, reflective decision. The DM must understand the meaning of the attributes and their effects to be able to make a good decision. A good set of attributes is decomposable, meaning that a decision problem involving N attributes can be quantified in several sets with less than N attributes in each. This is highly important in cases where the dimensionality of the problem is high. Keeney and Raiffa (1976) regard five dimensions as a limit, after which decomposition is necessary. A good set of attributes should avoid redundancy, for surely we do not want that two attributes count the same thing, for that would then mean double-counting some evidence in favor of (or against) a particular decision option. Minimum size means that as long as the above criteria are fulfilled, we ought to opt for the smallest possible set of attributes.

2.3.2 Previous usage of the epsilon model

A similar setting that is the antecedent of this study was used in Korhonen et al. (2013, 2012). They used an equivalent linear value function model in a bi-criteria setting and investigated the connection of linear function weights and AHP weights, which were obtained from subjects using the standard AHP methodology (Saaty, 1980). They found that 67,7 % of subjects had consistent linear weights, compared to the importance of criteria (ie. AHP weights). Consistency here means the same ordinal relationship. Transforming the weights to different scales did not have a significant effect in any way. This is quite likely due to the fact that the criteria in their study were already very similarly scaled on the original scale. As Korhonen et al. (2013) only had a bi-criteria setting, it may be predicted that in our case the consistency will be even lower. As in this study the setting has four criteria, it is expected that subjects will exhibit more inconsistencies than in a simpler setting.

Korhonen et al. (2013) also looked at the prediction power of different methods. In their study, estimated weights outperformed participant provided AHP weights. The same was true when assuming that the weights represented a scaled individual range. A comparison of using equal weights and estimated weights that were forced to be inconsistent with subjects' AHP weights did not show a large difference, and they concluded that inconsistent estimated weights were as powerful in prediction as equal weights.

The study of Korhonen et al. (2013) questions two common assumptions. It is often assumed that criterion weights reflect criterion importance. That was true only for a subset of the participants. Moreover, it is also commonly assumed that individuals are capable of providing criterion weights directly. The model with directly provided weights did not prove to be a maximal predictor, but was outperformed by the weights estimated from the linear value function. Their study therefore raises the question of how should we then interpret statements of importance. People seem to make statements such as *"Leisure time is more important than a good wage"*. But if these statements are not consistent with the weights of choices people make, it is an open question what they then mean instead.

In another study, Korhonen et al. (2012) found that even though only 38,9 % of subjects were consistent with a linear model, the model nevertheless was able to predict the choices made by participants. In their study, the maximum likelihood evaluation revealed that the predictability of a choice is quite high, irrespective of whether the subject is consistent with the linear model or not.

Korhonen et al. (2012) also found out that subjects were quite likely to make inconsistent choices in the control questions, but this had very little impact on the linear consistency of the subject. Despite the fact that 49,3 % of subjects made at least one inconsistent choice regarding the control questions, only one subject changed status from linear to nonlinear, when the control questions were used as a replacement for the originals. Additionally, the rate of change from linear to nonlinear was not significantly different from the rate of change from nonlinear to linear.

The linear model used in Korhonen et al. (2013, 2012) proved to be quite robust in the bi-criteria situation. With the epsilon formulation, only 38,9 % of subjects were consistent with the linear function. Allowing for 5 % of inconsistencies in choices, 64,6 % of subjects were linear consistent. Allowing 10 % inconsistencies raised the consistency ratio to 83,3 %. The model predicted choices better than weights provided by the participants. The lack of prediction power of directly provided weights could be due to a multitude of factors. It could be that subjects

- do not know initially what they prefer
- preferences could evolve
- could fail to choose according to their preferences, due to
  - lack of attention
  - unable to be fully consistent
  - making errors

Different answers to the question of why the provided weights are not the best prediction scheme demand different inferences regarding the constructivity or computational limits of participants. Inconsistencies with control questions can be explained by the subjects' tiredness towards the end of the experiment. However, tiredness is of no import when trying to explain the bad performance of provided weights. The classic position of revealed preferences would demand that subjects' are fully aware of their preferences, which ought to mean that they are able to provide optimal weights. Of course, one could also argue that the subjects can provide them, but perhaps they do not want to. However, I see no good reason why the subjects would be reluctant to provide the experimenters with that information. One could also argue that perhaps, even though the subjects were aware of their preferences, they are unable to fully choose according to them. Perhaps they lacked sufficient attention, or did not have the computational capability to succesfully distinguish, which of the options actually satisfies their preferences best.

2.3.3 Objections to linear models

The use of linear models instead of human judgments is not always regarded as a favorable development. Some of the most common arguments against predicting choices with the linear models have been categorized as technical, psychological, and ethical objections (Dawes, 1979).

The most common technical objection revolves around the idea that the worse performance of human DMs results from the fact that either the sampled experts were the wrong ones or that human decision making is worse than a linear equation on short term, but outperforms the linear model on a longer term. Dawes dismisses these both objections with the simple counterargument that these claims are purely speculative and no supporting data exists (Dawes, 1979).

A common psychological objection is that surely the reality must be quite predictable, and if the linear model only has a correlation coefficient of .4, then surely another type of predicting will be better. Dawes says that this is clearly just wishful thinking, and no better models or methods have been found - and we must remember that the human DMs do even worse. Dawes speculates that some of the psychological resistance and the feeling that *surely humans must be better than models at this* arises from our distorted memory, causing us to remember good predictions of experts rather than predictions that were completely off the mark (Dawes, 1979). This is explained by the representativeness heuristic (Tversky and Kahneman, 1974).

Lastly, the ethical objections arise due to the uncomfortable fact that humans' behavior is being reduced to numbers. This is somehow seen to be demeaning or violating the humanity of those being approximated by the linear model. Dawes comments that surely the most ethical treatment of other people is to use the best predictive model, no matter what it is. In the case of a grad school application, if a linear model does better than interviewing, surely the ethical thing to do is to use the model. (Dawes, 1979)

2.4 Behavioral issues

The study of decision making has historically had a heavy emphasis on normative models and has often compared the behavior of subjects with these models. It has sometimes been said that if a subject behaves in a different way than the model would suggest, well, so much worse for the individual. The fact that there is an optimal model that is better than the subject's model is essentially what Einhorn & Hogarth claim as a major principle of a lot of decision making research(Einhorn and Hogarth, 1981). According to them, the abstract way of representing rational decisions by their instrumental rationality - ie. effectiveness in reaching given goals - is not in all cases sensible, as some goals are more sensible than others. Einhorn & Hogarth (1981) claim that future research in decision making needs to adopt a broader perspective. This means, for example, treating preferences as contextual and not just as some abstract structure that the decisions of a subject reflect. This is similar to what Goldstein says about the local and global context of relative importance (Goldstein, 1990).

The challenge of the behavioral stream is that perhaps humans do not have any objectively attainable value function at all. If this is so, then a linear model can only ever be a predictive model and does not in any way represent the decision making of the subject. This, in combination with Einhorn & Hogarth's (1981) claim that rational decisions are those we post hoc deem to attain "good results", would imply that even a good predictive model implies nothing about the underlying rationality of the subject.

This issue has been recognized previously by Korhonen et al. (2012), who find that people choose inconsistently with respect to a linear or any function by, for example, choosing a dominated alternative. (Korhonen et al., 2012)

A major challenge arising from the behavioral stream is the question regarding the nature of our preferences. Do we already have preferences ex ante, or are these formed only along the way as we are prompted by the practical situation or the experiment? This debate has been a very contested issue, although psychology has amassed evidence for the constructive perspective (Weber and Johnson, 2009; Tversky and Simonson, 1993). Next I will review some of the discussion regarding the possibility of constructive preferences.

2.4.1 Constructive preferences

One of the foundations of decision theory is the concept of preference. The standard analysis of choice assumes that preferences are stable and invariant (Tversky and Simonson, 1993). According to this account, experiments in decision analysis merely reveal these underlying preferences, which existed already before the experiment. This account, therefore, requires that preferences are independent of the method of measurement and the context of the experiment. The utility or value of a choice option therefore depends only on the option itself. The standard account treats the DM as a fully rational agent that has the necessary computational abilities to perfectly calculate which option maximizes the value that he receives from the decision.

In contrast, the idea of constructive preferences claims that human beings might not be these ideal actors capable of precise maximization. Based on the concept of bounded rationality (Simon, 1955), it is argued that human decision making is constrained by limitations and that hence the concept of stable and invariant preferences is implausible. If our preferences are constructive, ie. generated in response to the task or experiment at hand (Payne et al., 1992), there is possibly no stability outside the domain of the experiment. Descriptive decision making research has demonstrated that DMs are highly susceptible to a variety of task and context factors. Task factors are *"general characteristics of the decision problem [–], which do not depend on the particular values of the alternatives"* (Payne et al., 1992), whereas context factors are dependent on the particular alternatives and their values. Research shows that influential task factors are, for example, response mode (Einhorn and Hogarth, 1981), similarity of alternatives (Tversky and Simonson, 1993). Furthermore, some of the salient characteristics of the alternatives that DMs focus on may be normatively of little weight or relevance, such as salience (Tversky

and Kahneman, 1974), or representativeness (Kahneman, 2003). It is also know that responses in joint evaluation differ from those made in separate evaluations (Hsee, 1998). This means that subjects evaluate two options differently when they see them both at the same time, compared to when they see them separately. This aspect has been explained with loss aversion (Hsee, 1998).

If preferences are highly constructive, it creates a problem for interpreting the results of any decision experiment, which compares DMs actions to the normative theory of expected utility. This is exactly the critique made by Tversky et al. (1988). They have empirically identified two principles related to the divegence in results of choice and matching tasks: the prominence effect and the combatibility principle.

The prominence effect states that the more important dimension of a decision problem looms larger in choice than in matching. The prominence effect is explained as an instance of the combatibility principle. This principle states that dimensions that are compatible with the response mode are weighted more heavily. Quite clearly the prominence effect violates the normative theory, and more specifically, the principle of invariance. The principle of invariance states that choices should only depend on the situation, not on how it is presented. (Kahneman and Tversky, 1984).

It is important to note that the theory of constructive preferences does not argue that *all* of our preferences would be constructive. As a matter of fact it is clear that we do have at least some quite stable preferences. Were it not so, the rational choice theory would have never emerged. As Bettman et al. (1998) note:

> *People are most likely to have well-articulated preferences when they are familiar and experienced with the preference object, and rational choice theory may be most applicable in such situations.*

However, even if the underlying preference structure is stable, a decision can still be influenced by external factors. This idea of contextual, situational, and processual factors infiltrating an otherwise stable decision situation suggests that there are three sources of variance in a decision situation. First, we have the stable underlying preference. Second, we have the contextual factors mentioned before. And third, we have error in the measurement itself. Decision making, then, is actually a product of these two kinds of systematic variance, and not of just stable preferences, as described by rational choice theory.

A constructive view of preferences suggests three sources of why invariance might fail (Payne et al., 1992):

- Decisions involve conflicting values
- Invariance results from simplifying complex tasks
- Invariance may result from uncertainty of values

Conflicting values mean that our underlying value system has internal conflicts, and the decision task forces us to think how much we are willing to give up from some aspiration to facilitate reaching another goal. As this happens in different situations, we may exploit different strategies and hence cause invariance. Simplification of

tasks means that since problems are often complex, our heuristics and other methods for simplification might result in invariance. Thirdly, we might be uncertain about our future values. For example, even though I know how a new apartment looks like I might be uncertain how much I will enjoy the spacious design after finally having moved in there.

A further complication in the assessment of utility and preferences is the fact that our view of the problem is coloured by the way information is presented to us. Järvenpää (1990) found that information processing was influenced by how graphic displays are organized. More exactly, Järvenpää found that the most attention was focused on the attribute with the highest difference in the values of the alternatives. In our experiment, this would mean that subjects' attention focuses on that attribute with the largest difference in the height of the columns. For a screenshot of how the information was presented graphically, see Appendix B.

A reasonable hypothesis made by Payne et al. (1992) is that the more uncertainty there is related to one's preferences, the more influence task and context effects will have. This would mean that when we make a decision, of which we are quite certain and that is related to clear values, the less subject we are to confounding effects. This is an important argument for choosing a decision problem that the subjects are at least somewhat familiar with for the experiment.

Next, let us look at the dual system processing theory, which has been very influential in the behavioral field and forms a backbone that a lot of research builds on. Especially, the dual processing theory tries to explain the duality of heuristic and analytic processing.

2.4.2 The two systems of cognitive processing

A research tradition that is by now already well over 30 years old has conclusively shown that humans regularly and quite predictably differ from the rational homo economicus in several kinds of settings. Hundreds of empirical studies have reported systematic irrationalitites and biases that plague the decision making in several contexts (Stanovich and West, 2008; Kahneman, 2003).

These deviations from the rational norm have fuelled a debate regarding their meaning, a debate that has also lasted for decades now. Several possible interpretations have been suggested, but one of the most prominent ones is the idea of two different systems of cognitive processing. Even though there are minor differences, the idea of a fast and frugal heuristic system and a slower and more analytic one, has been the kernel in many theories (e.g. Sloman, 1996; Evans and Over, 1996; Hammond, 1996; Epstein, 1994). Stanovich and West (2008) label these two different systems as System 1 and System 2.

System 1 is an automatic and holistic process, whereas System 2 is largely rule-based and controlled. Another very vital difference is that System 1 is thought to be less demanding in terms of cognitive resources. In contrast, System 2 demands attentional and computational resources that we might not always have at the moment of need.

As Kahneman (2003) it succintly expresses:

> *The operations of System 1 are typically fast, automatic, effortless, associative, implicit (not available to introspection), and often emotionally charged; they are also governed by habit and are therefore difficult to control or modify. The operations of System 2 are slower, serial, effortful, more likely to be consciously monitored and deliberately controlled; they are also relatively flexible and potentially rule governed.*

Another key element in the dual view is the monitoring function of System 2. Essentially, System 2 is thought to monitor the quality of mental operations and necessitating an override when the quality of System 1 response is deemed unfit for the task. This monitoring function and the process of overriding the heuristic response is displayed below in Figure 1.

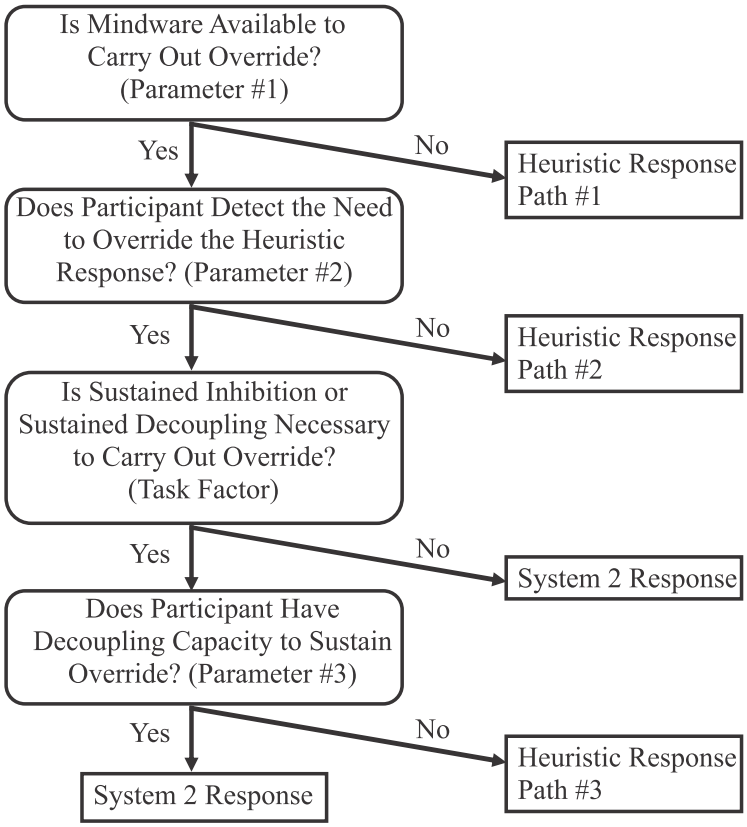


Figure 1: Heuristic override process (Stanovich and West, 2008)

Stanovich (2008) defines mindware as

> *the rules, procedures, and strategies that can be retrieved by the analytic system and used to substitute for the heuristic response.*

This is the first necessary component for overriding a heuristic response. Secondly, the participant needs to notice that an override is useful and necessary in this situation. Finally, if the situation requires constant cognitive decoupling, ie. sustained

inhibition of the heuristic response, the subject needs to have the necessary resources and capacity to continue this inhibition.

In the context of this study, Stanovich's & West's (2008) framework is sufficient to highlight the fact that there are necessary conditions for making a rational System 2 decision. In this case, the subject is faced with a repeating decision task. Success in this task requires attention to all four criteria and a reflection of how the different criteria contribute to the utility of the given options. Additionally, an optimal decision maker would be aware of his or her previous choices and could use these to reflect on the value of different criteria.

The idea of two cognitive systems immediately gives rise to the question that which system is more prominent in our decision making. Here, as mostly in science, the answer is a qualified "it depends". But it depends on what exactly? This is one of the contributions of this thesis: to highlight the connection of decision making to that of the cognitive system. More exactly, my proposition is that those subjects, who tend to use System 2 more, would use a more analytic decision making strategy, whereas those relying on System 1 rely on a more heuristic strategy. This could mean that System 2 thinkers' choices might be better predicted by a more computationally demanding but more comprehensive choice strategy, such as a linear function. Conversely, System 1 thinkers might make use of a more heuristic strategy such as a lexicographic model. Importantly, Payne et al. (1992) predict that in a two-alternative choice situation people prefer generally compensatory strategies, such as the weighted additive model, whereas in situations with more than two alternatives, they prefer noncompensatory strategies.

This supports my hypothesis in the sense that, perhaps for thinkers with a more heuristic style, the limit for switching to a heuristic style might be already achieved with two alternatives and four criteria. In this study, we measure the cognitive style of the subject with Frederick's (2005) Cognitive Reflection Test.

2.5 Relative importance of criteria

What does it mean to say that one criterion is more important than some other criterion? When buying a car, what do I mean when I say that price is more important than fuel efficiency? The question regarding the meaning of relative importance has puzzled researchers for a long time. However, it has rarely been specifically addressed. Korhonen et al. (2013) tried to find out if the relative importance of two criteria have implications regarding the linear weights of those criteria, finding that for a third of the participants, the relationship of importance and linear weight was in fact inverse (Korhonen et al., 2013).

Intuitively, one might expect that when comparing alternative a and alternative b, the alternative with the "more important" criteria on its side would be deemed as the preferred one. These kinds of conflicts, combined with the logic of aggregation in MAUT, form the basis for understanding the relative importance of criteria (RIC) in our context. (Roy and Mousseau, 1996)

One methodology that has been suggested for reconciling the judgments of importance with the elicitation of weights is the Analytic Hierarchy Process by Saaty

(1980), which will be introduced next.

2.5.1 The Analytic Hierarchy Process

The Analytic Hierarchy Process, designed by Thomas L. Saaty (1980), is a methodology designed to utilize the fact that we humans are adept at recognizing relations and ratios among different objects in reality. The philosophy behind this approach is that if we form a matrix consisting of the ratios of different alternatives regarding some criteria, this matrix can then be used to hierarchically rank the different alternatives in an order. In this study, I use the AHP methodology to obtain subjects' rankings of the different criteria and then transform these into a vector of weights. These vectors then represent the ex ante judgments of importance that subjects have provided for different criteria. Essentially, this method can be considered to be a variation of obtaining the weights from subjects directly. The exact method is presented below.

The subjects' are asked to rank the criteria by doing pairwise comparisons. The AHP methodology defines importance of criteria on a relative scale so that each criterion is compared to every other criterion, and the relative importance value is picked from the table below. The verbal values in Table 1 refer to the wordings we have used in the form, whereas the explanations refer to those given in (Saaty, 2008). I have used the term attribute instead of activity here, as in this study the AHP values refer to the attributes of the different options.

Relative importance	Verbal value	Explanation
1	Equally important	Two attributes contribute equally to the objective
3	Moderately more important	Experience and judgement slightly favour one attribute over another
5	Strongly more important	Experience and judgement strongly favour one attribute over another
7	Very strongly more important	An attribute is favoured very strongly over another; its dominance demonstrated in practice
9	Absolutely more important	The evidence favouring one attribute over another is of the highest possible order of affirmation

Table 1: Saaty (2008) criteria of importance

The users are also given the opportunity to define a criterion as less important than another one. In this case, the reciprocals of the above values are used, with "more" replaced by "less" in the verbal values. With this methodology, the whole comparison matrix can be defined with six pairwise criterion comparisons, as we assume that if price is strongly more important than size, then size is strongly less

important than price. Mathematically, this means that if in the comparison matrix $a_{ij} = b$ then $a_{ji} = \frac{1}{b}$.

With the AHP method, we obtain a criteria importance matrix as follows. Let us assume that in our experimental problem a subject regards size strongly more important than price, very strongly more important than distance to University and absolutely more important than distance to leisure activity location. Moreover, let us assume that price is moderately more important than distance to University and strongly more important than distance to leisure location. Finally, distance to University is moderately more important than distance to leisure location. Using the AHP categories and their numerical analogues, the following matrix can be constructed as in Table 2.

	size	price	distance to Uni	distance to leisure
size	1	5	7	9
price	1/5	1	3	5
distance to Uni	1/7	1/3	1	3
distance to leisure	1/9	1/5	1/3	1

Table 2: Example of the AHP methodology in our experimental setting context

According to the AHP method, the weights of the different criteria can be obtained by computing the principal eigenvector of the matrix. For simplification, I have used an approximation recommended by Saaty (1980), which is to multiply the elements of each row and take the 4th root of the result. Finally, the resulting weights are normalized so they sum to one. This method gives us the following weight vector: $w = (0,654467; 0,204451; 0,095507; 0,045575)$ The vector we would get with the exact principal eigenvector computation would be $w = (0,658152; 0,202696; 0,093612; 0,04554)$. As can be seen, the computationally less intensive approximation is suitable for our purposes, as in this study we are mostly concerned with the rank ordering of the weights, for which the exact last decimals have little importance.

2.6 Lexicographic ordering

Let us now look at a decision strategy that is noncompensatory and does not depend on a linear value function assumption. For an example of this class of models I have picked the lexicographic model. In the context of this study, it is meant to act as a comparison standard for the linear models and also to check, whether nonlinear subjects might be better approximated by some heuristic model, such as the lexicographic model.

A lexicographic decision maker looks at each option with respect to the most important criterion first. If there is a single option with the highest value on this criterion, that option is chosen. If there are several options with the best value,

then they are compared according to the second most important criterion. This algorithm continues until the DM reaches the point where only a single option is the best on some criterion. Formally, the algorithm is as follows:

1. Set the criteria ordering so that x_1 is the most important, x_2 the second most important etc.
2. Set $j = 1$.
3. Choose the next most important criterion x_j
4. $\forall b \in A$: if $x_j(a) > x_j(b) \implies$ remove b from A.
5. If $|A| = 1 \implies$ choose the remaining element a.
6. Increase j by 1.
7. Go to step 3.

Lexicographic ordering is of interest here simply for checking if the fact that the problem has four attributes causes some DMs to use a lexicographic method, which is much easier and cognitively less demanding compared to a compensatory model (Hastie and Dawes, 2009). In effect, this hypothesis is based on the consideration that when faced with a more cognitively demanding problem DMs have three alternatives at their disposal: 1) they can expend more effort than with a simpler problem, 2) they can choose a simpler strategy that demands less cognitive resources, or 3) they can give up and make their decision at random.

2.7 Weights

A linear model implies that when the range of an attribute is exogenously expanded, that attribute should have a higher weight (Fischer, 1995; Weber and Borcherding, 1993). This is clear when one considers the weights as representing tradeoffs between units of different criteria. These tradeoffs naturally depend on the range of the attribute (von Nitzsch and Weber, 1993). When comparing jobs, for example, the weight placed on salary should increase if the range of salary increases from [1500, 2000] to [1300, 3500]. This is intuitively clear, as expanding the range means that the attribute has a greater role in the differences between the attributes.

In his article, Goldstein (1990) has analyzed the pattern according to which subjects' decisions are related to the subjects' concept of relative importance. There are basically three possibilities:

- subjective weights vary in an unrelated manner
- subjective weights remain fixed
- subjective weights vary correspondingly

Based on an experiment he has conducted, Goldstein argues that the two latter options are what he calls *global* and *local* interpretations of relative importance. A global interpretation means that subjective weights are thought to refer to the

subject's underlying values. A local interpretation means that subjective weights vary according to the stimulus set. This would mean that the weights should then be independent of the stimulus set.

The experiment that Goldstein conducted supports a local interpretation of weights. It should be noted that a local interpretation of the weights would seem to be against the notion that the weights have anything to do with the importance of a criterion. This is motivated by the claim that importance is really a product of the linear weight and the variance of the criterion. If a criterion has low variance (ie. only small differences between options), even a large linear weight does not ensure that the criterion has a large impact on the aggregate utility. (Goldstein, 1990)

Fischer (1995) investigates the relation of changes in the range of an attribute and its effects on the weight of that attribute in the linear function. It is clear that the weight of an attribute should reflect its impact on the overall index of value. This means, in effect, that a larger range for an attribute means the attribute explains more of the variation of the options - thereby increasing the weight of that attribute.

Formally, Fischer develops the argument as follows: Let us denote the globally best attribute values as X_i^*, X_j^* and the globally worst attribute values as X_i^0, X_i^0. Let us similarily mark the locally best and worst values with x_i^*, x_i^* and x_i^0, x_i^0. Now, let us assume a *globally defined value function* $V(X_i, X_j)$ so that $V(x_i.x_j) = \gamma_i V_i(x_i) + \gamma_j V_j(x_j)$ where

$$\begin{aligned} 0 < \gamma_i = V(X_i^*, X_j^0) < 1 \\ 0 < \gamma_j = V(X_i^0, X_j^*) < 1 \\ \gamma_i + \gamma_j = 1 \end{aligned}$$

The *locally defined value function* will be denoted by $v(x_i, x_j) = \lambda_i v_i(x_i) + \lambda_j v_j(x_j)$, where

$$\begin{aligned} 0 < \lambda_i = V(x_i^*, x_j^0) < 1 \\ 0 < \lambda_j = V(x_i^0, x_j^*) < 1 \\ \lambda_i + \lambda_j = 1 \end{aligned}$$

Assuming the decision maker is consistent in his preferences across local and global contexts (independence assumption), then the relationship of the local and global functions is:

$$v(x_i, x_j) = \frac{V(x_i, x_j) - V(x_i^0, x_j^0)}{V(x_i^*, x_j^*) - V(x_i^0, x_j^0)} = \lambda_i \frac{V_i(x_i) - V_i(x_i^0)}{V_i(x_i^*) - V_i(x_i^0)} + \lambda_j \frac{V_j(x_j) - V_j(x_j^0)}{V_j(x_j^*) - V_j(x_j^0)}$$

Note that the variables x_i, x_j refer to the local range here. This gives us the derivation of the weights as

$$\lambda_i = \frac{\gamma_i [V_i(x_i^*) - V_i(x_i^0)]}{\gamma_i [V_i(x_i^*) - V_i(x_i^0)] + \gamma_j [V_j(x_j^*) - V_j(x_j^0)]}$$

$$\lambda_j = \frac{\gamma_j [V_j(x_j^*) - V_j(x_j^0)]}{\gamma_i [V_i(x_i^*) - V_i(x_i^0)] + \gamma_j [V_j(x_j^*) - V_j(x_j^0)]}$$

From this format it can be clearly seen that the weights depend on the maximum and minimum values in the local contex, and also on the global context weights. Assuming the global weights are the same in any local situation for a given individual, and also assuming that the maximum and minimum levels of x_j stay the same, it is easy to see that the weight of x_i is going to increase, if the range of x_i increases. The increase of range can either be so that we have a new, lower x_i^0 or a new, higher x_i^*, or both. In previous research (Fischer, 1995) it has been shown that subjects are usually less range-sensitive than the normative theory would imply.

As Weber & Borcherding (1993) have noted, the exogenous increase in the range of a criteria should be reflected in a higher weight in the linear function. In our case the travel time is bounded from below so that the lowest possible value (1 minute) reflects the best possible option. So, in this case a larger range determined as feasible by the subject reflects the fact that the subject is willing to accept worse values, ie. live further away. Now, if a student is willing to live far away, it seems intuitive that the student does not regard distance to campus as a very important criterion. Therefore, contrary to Weber's theoretical stance, I predict that an increase in the range of the distance criteria correlates with a *decrease* in the weight of the linear function weight for those criteria.

The concept of weight can only be defined with regard to a specific theory of preferences (Weber and Borcherding, 1993). However, from a prescriptive point of view the following should hold for all weights: description invariance and procedure invariance. Descriptive invariance means that weights should not be affected by the description of the attribute or the structural components of the value tree. Procedure invariance means the procedure to elicit the weights should be unimportant, as long as the elicitation method assumes the same definition of weights. Furthermore, weights ought to be sensitive to the attribute value range. (Weber and Borcherding, 1993) Weber and Borcherding (1993) claim that MAUT model implicitly assumes that a decision maker will have well defined global preferences, which he or she uses to answer questions posed for weight elicitation.

Procedural invariance can be further divided into three variables. For a method to have procedural invariance (Weber and Borcherding, 1993):

- the decision maker is able to understand the method and use it consistently
- different weight elicitation methods yield similar weights
- context effects have no influence

Weber and Borcherding (1993) state that inconsistency is quite common with different methods. It is also known that on an individual level, consistency has only a low correlation with judged difficulty of using the given method (Borcherding et al., 1991). On that note, Weber and Borcherding conclude that some inconsistency is to be expected, but that inconsistency as such cannot be yet related to the validity of the method in any clear way.

Convergence between methods is one way to analyze the validity between methods. The problem is that we have no unambiguous way to compare the weights

generated by different methods. Weber and Borcerding conclude that the effect of elicitation methods on the weights is stable, and remains even when the subject has a chance to reconcile weights from different methods.

Despite these behavioral shortcomings, Weber and Borcherding conclude that MAUT is the only axiomatically founded theory, and that other methods are beset by the same biases, or more (Weber and Borcherding, 1993). As this study uses only a single setting there is no need for a large concern regarding descriptive invariance. However, one should keep in mind that, considering the inconsistencies of methods in the previous literature, perhaps it is not so surprising if the AHP weights fail to conform to the linear model weights, for example.

3 Model and methods

3.1 Basic model

Considering that this study is concerned with the prediction power of a linear value function, the problem can be set up by assuming a linear representation of the subject's preferences. This is easily done by using a linear programming formulation. The LP problem is constructed so that the choices made by the subject define the possible solution space. By using a linear function in constructing the constraints, the problem can be set up so that there is a solution only if the subject is consistent with a linear value function. Let us define the problem in the following way.

It is assumed that the subject's preferences are represented by a linear value function $f(X) = \sum_{j=1}^{p} \lambda_j x_j$, where

- p is the number of criteria, in our case $p = 4$,
- x_j refers to the attribute number j so that the attributes are from the set {price, size, distance to University, distance to leisure location}, as previously defined, and
- λ_j refers to the weight of the attribute x_j in the value function of the subject.

For each pair $(X_r, X_s) \in P$, X_r is the preferred alternative to X_s. Both alternatives are defined according to the aforementioned four criteria so that $X_r = [x_{r1}, x_{r2}, x_{r3}, x_{r4}]$. This means we get the following LP model.

$$\max \epsilon \text{ subject to:}$$

$$\sum_{j=1}^{p} \lambda_j x_{rj} - \epsilon \geq \sum_{j=1}^{p} \lambda_j x_{sj}, \text{for all } (X_r, X_s) \in P$$

$$\sum_{j=1}^{p} \lambda_j = 1$$

$$\lambda_j \geq \theta, j = 1, 2 \ldots, p, \text{where } \theta > 0 \text{ is non-Archimedean.}$$

As $\theta > 0$ is non-Archimedean so clearly $\lambda_j > 0 \forall j = 1, 2 \ldots p$.

Using the DMs responses, constraints for the LP-problem are constructed. For each "I prefer A over B" response the following inequality restriction is constructed:

$$\sum_{j=1}^{p} \lambda_j x_{Aj} - \epsilon \leq \sum_{j=1}^{p} \lambda_j x_{Bj}$$

Conversely, for each "I prefer B over A" response, the following inequality is created:

$$\sum_{j=1}^{p} \lambda_j x_{Bj} - \epsilon \leq \sum_{j=1}^{p} \lambda_j x_{Aj}$$

For each indifference response, no extra constraints are set, as setting the utility values of the options to be equal would place too strict constraints on the weights.

The LP formulation then gives us a model in which, for every strict preference the subject has stated, there is a constraint as defined above. This model is then run as a linear programming model. If a positive value for ϵ is obtained, this means that there is a positive difference between the preferred and the nonpreferred option for each of the subject's preferences. If $\epsilon < 0$ there is at least one preference, which is inconsistent with the linear model and hence the subject's choices are not consistent with a linear value function.

If this is the case, we can see how many such choices there are by removing constraints until the subject's choices are consistent with a linear value function, ie. until we obtain $\epsilon > 0$. Constraints are removed in the order of their shadow prices, ie. according to the order of how much their removal contributes to the increase of ϵ. The number of removed constraints then pragmatically represents how closely the subject's choices can be represented by a linear value function. This is exactly the same method as used in Korhonen et al. (2013). As a side note, it should be noted that as the constraints are removed one by one, it is not necessarily the case that they are removed so that the final model is the optimal one. However, this is not a problem for the experiment, as the number of removed constraints is meant to be an approximation of how close the fit to a linear function is.

3.2 The Cognitive Reflection Test

Frederick's (2005) Cognitive Reflection Test consists of three questions, which are so designed as to entice the fast and intuitive System 1 response.

1. *A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?*

2. *If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?*

3. *In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?*

The first question immediately invites the answer 10 cents that, upon some reflection, is clearly incorrect. The correct answer would obviously be 5 cents, as some easy math confirms. The two other questions in the test are similar in the sense that they also feature a question, for which the intuitive response is wrong.

Answering these questions correctly requires the respondent to suppress the heuristic response of System 1 and instead use a slower and more effortful System 2 process to respond. This is precisely what the test is aiming for: measuring the respondent's ability to suppress the heuristic system and replace it with the use of a slower analytic system. The general model of the suppression of a heuristic

response was presented previously in Figure 1, based on the two-system model of Stanovich and West (2008).

The CRT correlates with several other cognitive measures, such as SAT, but it is not just another measure of intelligence (Toplak et al., 2011; Stanovich and West, 2008; Frederick, 2005). The CRT is a good predictor of rational thinking that retains a significant portion of the variance it explains even when analyzed independently of intelligence, or executive functioning (Toplak et al., 2011). As Toplak et al. (2011) argue the CRT seems to reflect especially well the fact that some subjects engage more in miserly, ie. heuristic and superficial, processing than others. Miserly processing is characterized by using heuristics to provide a quick and dirty solution, which is computationally less costly but more inaccurate. The CRT measures this especially, as in contrast to other types insight and intelligence tasks, it primes the respondent with the wrong answer, thus requiring more controlled processing to override the primed wrong answer and reach the correct one.

Weber and Johnson (2009) have the following to say about the CRT:

> *The cognitive reflection test (CRT) is a three-item math-puzzle test designed to elicit an incorrect "intuitive" answer (generated by System 1) that needs to be overridden by System 2 intervention .*

As far as the uses of the CRT are concerned, Weber and Johnson (2009) say that

> *[CRT] suggests that* ***normative choice models may turn out to be descriptive for at least a subset of the general population****, those who have a greater ability or inclination to use rational/analytic processing in their decisions. CRT scores correlate moderately with conventional IQ measures, some of which show higher correlations than the CRT with normative choices in specific domains. However, the CRT is the most consistent predictor across choice measures and by far the easiest test to administer.*[emphasis added]

3.3 Hypotheses

The purpose of this study can be succintly stated in the following three questions:

1. Is the linear model a good model for predicting decisions in the context of this problem?
2. Does the prediction power of the linear model differ when comparing groups with different CRT scores?
3. What , if any, is the connection between judgments of importance and linear weights?

Hypothesis 1: The median consistency differs when comparing low-CRT and high-CRT groups

One could hypothesise that differences in the use of cognitive systems are connected either to the strategy that the subject uses or the consistency in using that

strategy. Either of these situations would result in differing levels of consistency between the groups. For example, if those relying more on System 1 use a strategy that picks the option with the largest proportional advantage on some criterion (price, for example) there would be situations in which the linear model makes a different choice. Or, if they use a linear model but are worse in executing it, for example, by focusing attention on irrelevant things, then the performance of the two groups would differ.

Hypothesis 2: The linear model predicts mens' choices better in comparison to womens' choices

This result arose in the study of Korhonen et al. (2012), but there was no explanation for it. My hypothesis is that the result was not a simple artifact and the same thing will happen again. This hypothesis is also reinforced by the fact that as in Frederick (2005), men tend to have a higher average CRT score.

Hypothesis 3: The linear model is the best predicting method in the group of methods consisting of estimated weights, equal weights, AHP weights and a lexicographic model.

What Korhonen et al. (2013) found was that the linear method was better at predicting choices than a linear model with AHP weights. It also tied with equal weights on inconsistent subjects, when the linear weights were forced to be inconsistent to the importance order of the criteria. Korhonen et al. (2013) hypothesize that inconsistent weights might even outperform equal weights with another data set. In this study, I have also included a lexicographic model to see whether it performs better than a linear model. After all, it could be that linearly inconsistent subjects follow such a heuristic strategy rather than use a linear value function.

3.4 Design of the criteria set

The problem in this study is a MCDM problem, in which each choice option is defined by four criteria: size, price, distance to University and distance to leisure location. For each criterion less is better, with the exception of size, for which more is better. This criteria set fulfills to a satisfactory level the requirements set by Keeney and Raiffa (1993) introduced in chapter 2.3.1.

Completeness. The criteria set is quite close to being complete. Preliminary discussions before the study with undergraduate and graduate students showed that size and price were the two criteria that subjects first thought of. Most mentioned location, transport connections or proximity to Helsinki, the University or their hobbies. Many also mentioned the layout or the condition of the apartment, but these were dropped due to the difficulty of operationalizing them.

Operational. As noted before, all of the chosen criteria were mentioned in preliminary discussions. Additionally, the questionnaire software and the information presentation format was tested on a few colleagues and classmates of the author. These tests proved the criteria to be meaningful and understandably presented.

Decomposable. As we only have four criteria it was not necessary to decompose the set into smaller pieces.

Nonredundancy. The criteria clearly involve different things. In the real world

price, size and transport connections are interconnected due to market effects. This was not simulated in the study, and we treated the criteria independently. An exception to this are the distance criteria. As they have a geometric relation that depends on the leisure location the subject was thinking of, this relation was included in the questionnaire software.

Minimum size. As I wanted to decompose the transport connection criterion into two separate criteria, this four-criteria format is the minimum size possible.

3.5 The experiment

3.5.1 Design of the experiment

Students completed the task with a web-based questionnaire tool that the author had developed. First the program asks for background information such as study year and gender of the participant. Then the subject defines his or her realistic ranges for the criteria. "Realistic" in this context means a range that the subject considers realistic for him or herself. For example, the subject could define his apartment size range as 15-40 square meters, price as 250-350 euros/month, distance to University as 1-35 minutes and distance to leisure location as 1-45 minutes. Defining the ranges meant that students would be faced with only realistic options. Defining ranges was necessary, as the preferences of students and their monetary resources differ considerably, so presenting the same options to everyone would not prove fruitful.

We also asked participants to provide the importance of different criteria to them. This was done by the methodology of the Analytic Hierarchy Process (Saaty, 1980), as explained previously in section 2.5.1.

Each student also answered the Cognitive Reflection Test, as formulated by Frederick (2005). Criterion values for size, price, distance to University and distance to leisure location are then generated by drawing them randomly from the intervals (using a uniform distribution). With these values the student is presented with 30 pairwise comparisons. The student can always answer that he prefers option A, option B or that he is indifferent. If the student does not answer at all and just moves on, the answer is recorded as indifferent.

Screenshots of the questionnaire program are provided in the appendix B.

The following types of questions were included in the study:

1-20: Normal
: Normal questions were pairwise comparisons, in which none of the options dominated the other one, and the criterion values were always inside the ranges defined by the participant. The criterion values always differed between alternatives, ie. the problem was always strictly a four-criteria problem.

21-22: Dominance
: Dominance questions had values inside the defined bounds, but so that one of the options dominated the other one.

23-26: Linear transformed
: Linear transformed questions were similar to normal questions, except that

each transformed question was obtained from a normal question by multiplying each criterion value by a scaling factor. The multiplication was done so that one criterion value was exactly on the defined range bound and all other values were inside the bounds.

27-28: Linear transformed + 10 %
These were similar to linear transformed questions, except that the scaling was done so that one criterion value was outside the defined bounds by 10 % .

29-30: Control
These were exact replicas of questions 9 and 14.

Linear transofrmed questions were used to see whether linear consistency is impacted by choices outside the subject's defined criteria range. Dominance and control questions were used to see how well subjects can retain their attention on the task. Normatively, a subject ought to always select the dominant option, and in control questions to select the same option he had selected before.

3.5.2 Participants

Participants were students at the Aalto School of Science. The questionnaire was marketed to three different Industrial Management undergraduate courses. The choice of student housing was chosen as the decision task because it was deemed to be familiar to the students.

As shown in Table 3 below, out of a total of 147 subjects, only 20 subjects, or 13,6 % did not complete the full questionnaire, which suggests that there were hardly any technical issues and that the students were motivated to complete the task. Students were motivated by rewarding 100 randomly determined participants with a movie ticket. To have a chance of obtaining a movie ticket, the student had to complete the full questionnaire.

Study year	# of subjects
All subjects	147
Incomplete answers	20
Total subjects	127

Table 3: Complete responses

The subjects had the following demographic distributions:

As shown in Table 4 below, the gender distribution is heavily skewed in favor of males. This is not surprising, as the Aalto University School of Science has a student population with a majority of males. Moreover, the participants were from courses that also have large amounts of students minoring in Industrial Management, and that population is even more male-dominant than Industrial Management majors.

Gender	# of subjects
male	87
female	39
unknown	1
sum	127

Table 4: Subject gender distribution

Surprisingly, even though the study was conducted with students from courses that are often taken during the first two undergraduate years, we also got a lot of responses from more advanced students, as can be seen in Table 5. This quite likely reflects the fact that there are more advanced students on the courses than what we originally expected. However, the study was not designed in any way just for undergraduates, so this does not pose a threat to the validity of the study.

Study year	# of subjects
1	24
2	27
3	37
4	20
5 or more	19
sum	127

Table 5: Distribution of study years of subjects

As comparing two options with four criteria each is already cognitively quite challenging, attribute information was presented in two different forms, both as a table with numbers and also as a column chart that displays the same information graphically (see appendix B3). This was thought to increase the chance of the student noticing a dominating alternative and therefore decreasing the chance that they choose a dominating option just because they haven't noticed the dominance.

3.5.3 CRT scores

The distribution of the scores of the Cognitive Reflection Test in the study are displayed below in Figure 2 and Table 6.

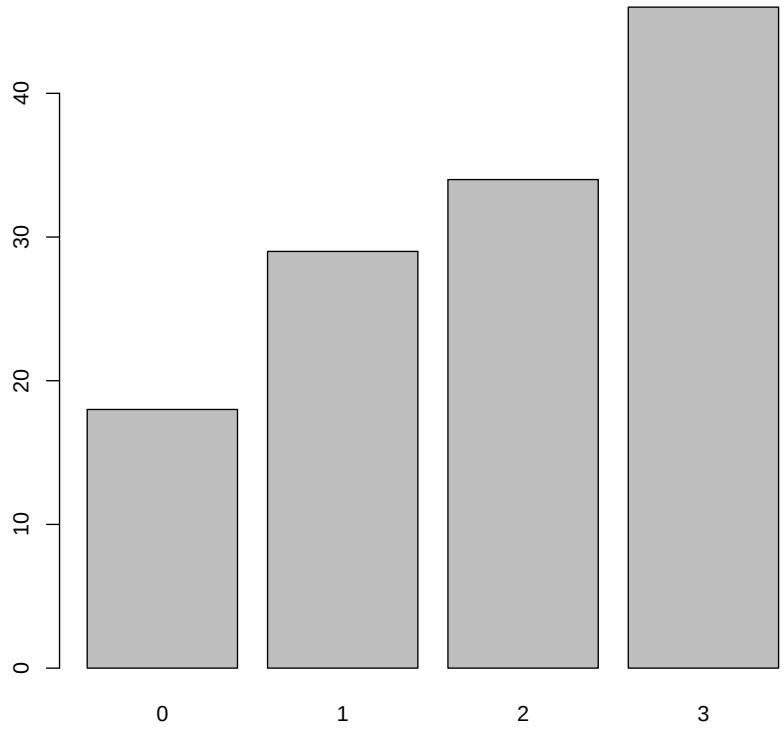


Figure 2: CRT score distribution of the subjects

CRT score	# of subjects
0	18
1	29
2	34
3	46
sum	127

Table 6: CRT score distribution of the subjects

As can be seen in Figure 2, the CRT scores are heavily skewed towards higher scores. Quite intriguingly, the CRT scores of the subjects are actually quite close to what Frederick (2005) obtained in his sample from the MIT. The reason for the skewness remains unclear and it will not be further explained in this study.

The CRT as a test regularly shows a difference between males and females, with males receiving higher scores than females (Oechssler et al., 2009; Frederick, 2005). It is so far unclear why this effect persists, although Frederick (2005) has suggested numerical ability, but this was not supported by Welsh et al. (2013).

3.6 Data Analysis

Consistency of preferences is analyzed by first checking for violations of dominance. Secondly, I check whether the subjects answer consistently to questions where the options are a linear transformation of another, previously encountered option pair. If they are linearly consistent, they should choose the same options in a linear-transformed option pair.

The fit with the linear function is also analyzed by looking at how much allowing 5% or 10% of errors (meaning 1 or 2 erroneous choices, respectively) changes the percentage of consistent subjects. If the impact is substantial, it can be hypothesized that the main driver of nonlinearity might be that subjects simply are incapable of choosing consistently to a linear function, due to tiredness or lack of attention, for example. On the other hand, if allowing errors has only a limited impact on the proportion of consistent subjects, it must be concluded that there is a substantial group of subjects, whose preferences cannot be well modelled by a linear function. It could be that such subjects use some other choice strategy that cannot easily be approximated by a linear value function.

The probability of a subject making consistent choices is analyzed in the same way as in Korhonen et al. (2012), ie. by using the maximum likelihood method. The binomial likelihood function provides an estimate of how likely an occurence of an event, such as a correct prediction or the choice of a dominant option are, given the observations in the study. These events are denoted by the probability p. Getting the maximum likelihood provides us with the most likely value of p, given the data.

In the function, k is the total number of choices, i represents the choices of interest and f_i stands for the number of subjects making those choices of interest.

The binomial likelihood function is the following:

$$L(p) = \prod_{i=0}^{k} (p^i (1-p)^{k-i})^{f_i}$$

This can then be transformed to log form:

$$L(p) = \prod_{i=0}^{k} [p^{if_i}(1-p)^{(k-i)f_i}] = p^{\sum_{i=0}^{k} if_i}(1-p)^{\sum_{i=0}^{k}(k-i)f_i}$$

$$\Rightarrow \log L = \sum_{i=0}^{k} if_i \log p + \sum_{i=0}^{k} (k-i) f_i log(i-p)$$

Since we are trying to maximize the likelihood function, the solution can be found

when the derivative $dL = 0$.

$$\begin{aligned}
\frac{dL}{dp} &= \frac{\sum_{i=0}^{k}}{p} - \frac{\sum_{i=0}^{k}(k-i)f_i}{1-p} = 0 \\
&\Rightarrow (1-p)\sum_{i=0}^{k} if_i = \sum_{i=0}^{k}(k-i)f_i p \\
&\Rightarrow p = \frac{\sum_{i=0}^{k} if_i}{\sum_{i=0}^{k}(k-i)f_i + \sum_{i=0}^{k} if_i} \\
&= \frac{\sum_{i=0}^{k} if_i}{\sum_{i=0}^{k} kf_i} = \frac{\sum_{i=0}^{k} if_i}{kn}
\end{aligned}$$

Additionally, I analyze the consistency of AHP weights and linear weights with the Kendall rank-order correlation (Kendall, 1948). This is a correlation coefficient that represents the conformity of the rank order of AHP weights and the rank order of linear weights.The rank-order correlation of Kendall is defined as

$$\tau_A = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}, \text{where}$$

$n_c, n-d$ are the number of concordant and discordant pairs, respectively.

The values of the rank-order correlation range from -1 to 1 so that 1 means perfect positive correlation, -1 means perfect negative correlation, ie. if AHP criteria order is, for example, {price, distance to University, distance to hobby, size}, then the linear weights order would be {size, distance to hobby, distance to University, price}. A rank-order correlation of 0 means that the rank orders are (approximately) independent.

The maximum likelihood estimation method is also used to compare the predictive power of different methods. In the prediction case, the maximum likelihood estimation gives us the probability that the given method predicts a single choice correctly. Confidence limits for the value of p can be computed by using Wilson's confidence interval approximation for the binomial distribution (Brown et al., 2001):

$$p_W = \frac{1}{1+\frac{z^2}{n}}\left[p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{(n-1)} + \frac{z^2}{4n^2}}\right]$$

where $z = z(1-\alpha/2)$. In this study, I routinely use a 5 % risk level. The Wilson interval should be used instead of the regular Clopper-Pearson (CP) interval, as it features a coverage probability that is more closer to the original value (Brown et al., 2001). Often the CP interval is used for its simplicity with the following qualifications:

1. $np \geq 5, n(1-p) \geq 5$
2. $np(1-p) \geq 5$
3. n quite large

The above list is not conclusive any way, but that is not of importance for two reasons. First, in this study, some of the above properties fail to be satisfied due to the small size of the CRT groups. Sencond, and more importantly, as Brown et al. (2001) convincingly demonstrate, the Wilson interval is preferable to the CP interval, especially for small samples.

Additionally, I compare the different prediction methods pairwise to each other. This difference measure is defined as

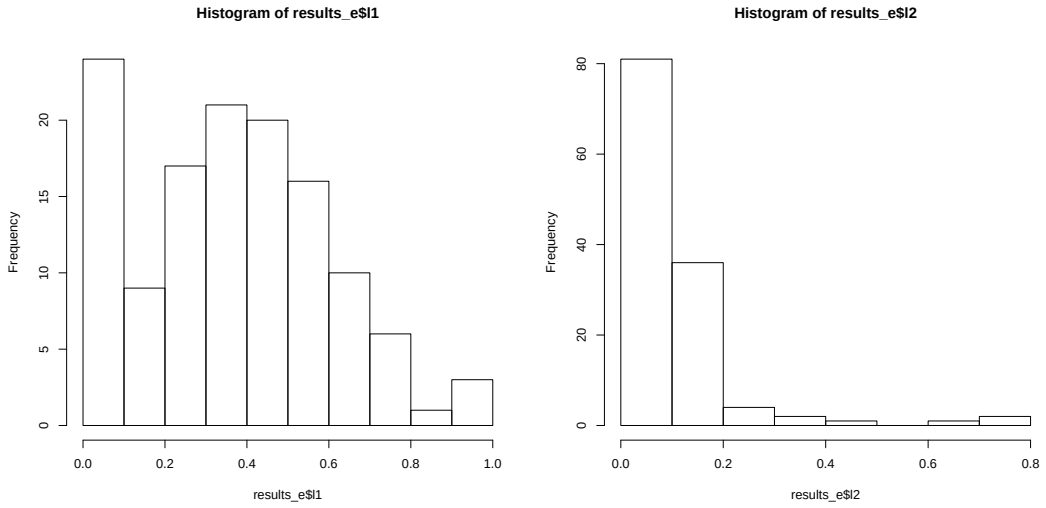
$$x_d = \overline{x_a} - \overline{x_b} = \frac{\sum_{j=1}^{10} \sum_{i=1}^{10} (i-j) f_{ij}}{n}$$

where $a, b \in (1, 2, 3, 4, 5)$, correspond to the different prediction methods that I have compared. This gives us the distribution of prediction power difference between the two methods and then a one-tailed t-test can be used do determine if the means of the number of succesful predictions differ in favor of one of the methods.

4 Results

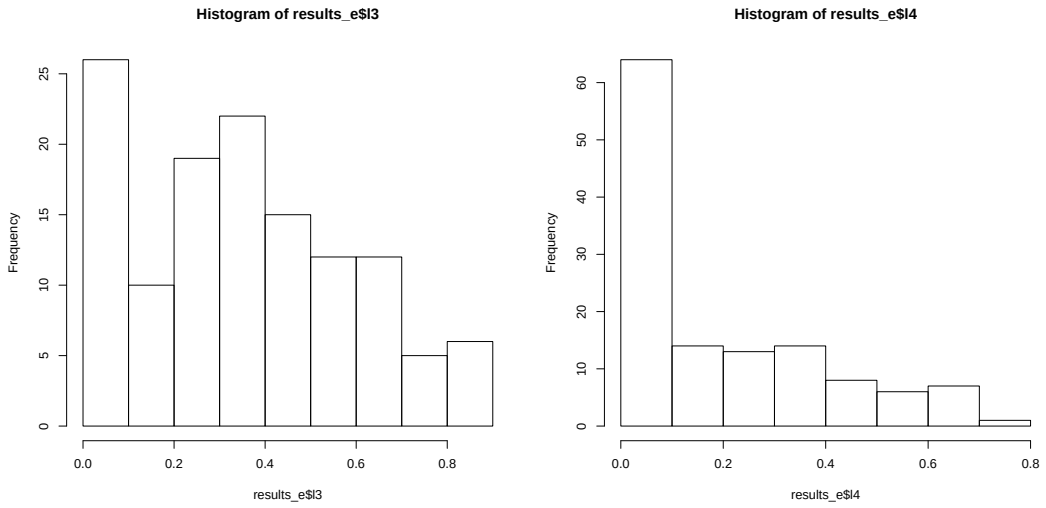
4.1 Results using original values

This section reports the results obtained with original criteria, ie. criteria values as they were defined for the participant in the task. Let us first look at the weights of the linear function. The following histograms for the linear function weights can be obtained:



(a) Size

(b) Price



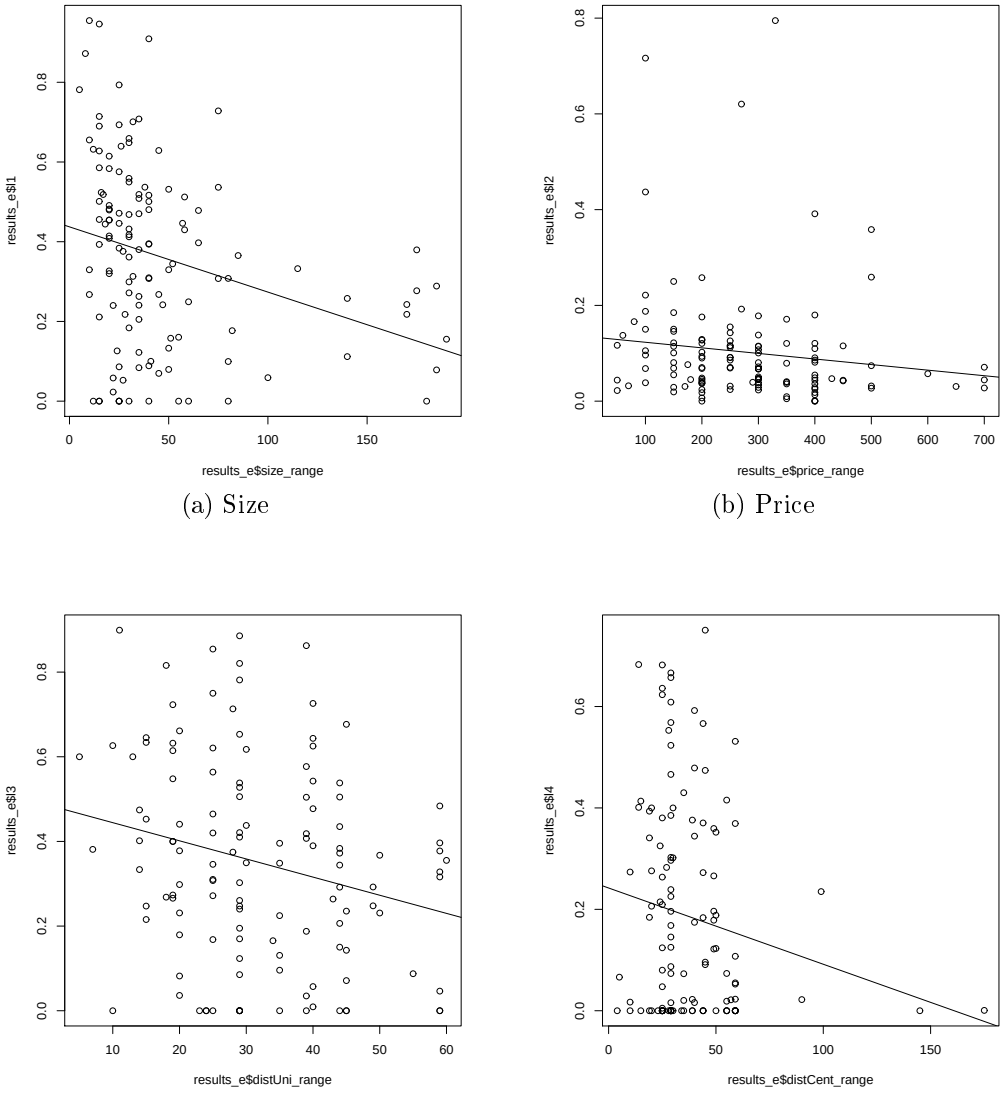
(c) Distance to University

(d) Distance to leisure location

Figure 3: Histograms of linear value function weights, original criteria

Surprisingly, only the weights for the apartment size and distance to University is quite evenly spaced. Especially surprising is the fact that the weight for the apart-

ment price is very small. Considering the fact that the participants are students, I would have expected them to weigh the price very heavily. However, once one realizes that we are looking at weights with unscaled criteria values, the explanation becomes obvious. The ranges for the price criterion were typically something around 250 to 600 euros, whereas size ranged from 15 to 60 square meters. Therefore, the value of price was typically approximately 10 times as large as the value of size. So, to have the same impact on the utility value of the option, the weight of size would need to be ten times as large as the weight of price. This explains the extremely low weights for the price variable.



(a) Size

(b) Price

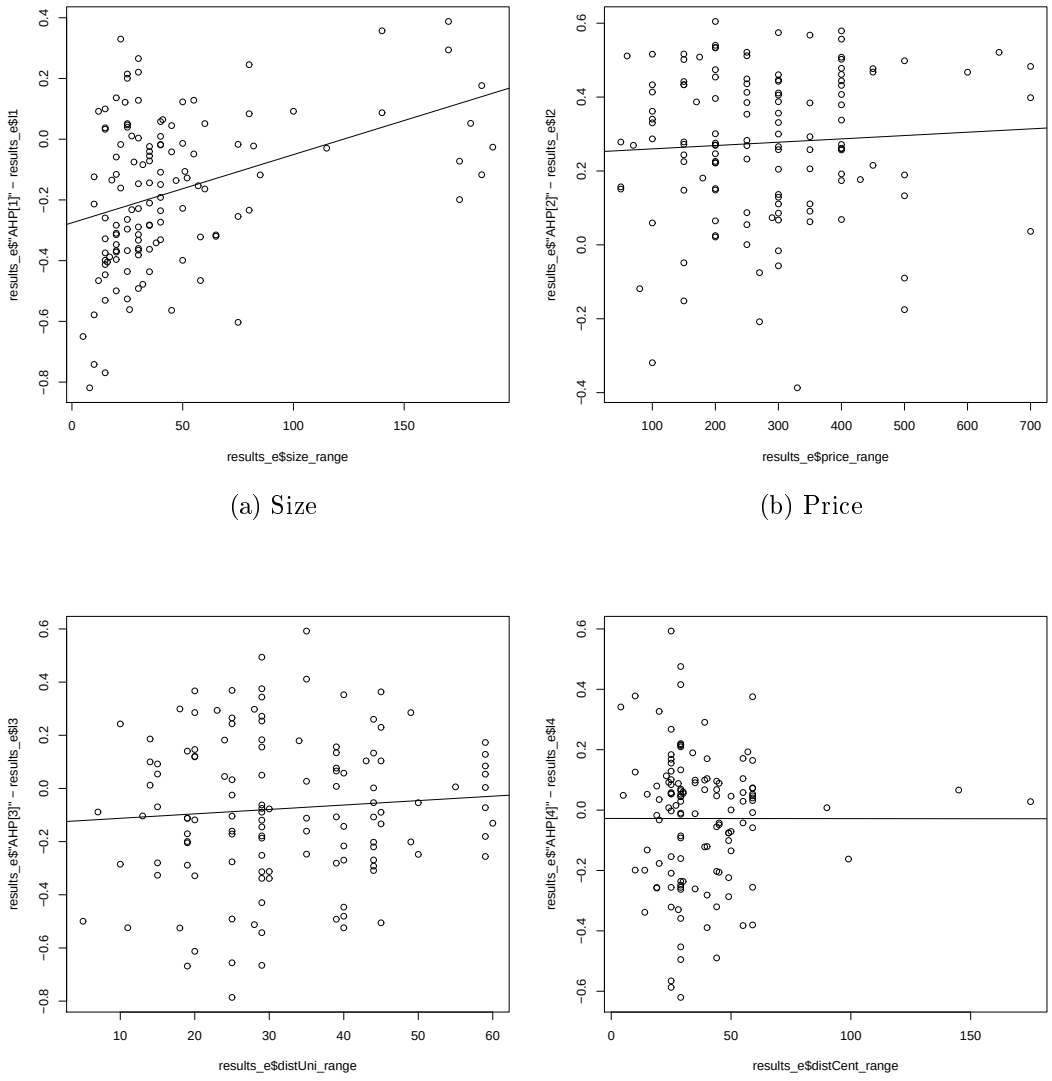
(c) Distance to University

(d) Distance to leisure location

Figure 4: Plots of linear value function weights, original criteria

As can be seen from the figures, even when checking for the range of the attributes, the weight of the apartment price is almost equally distributed across categories. The linear regression drawn in the picture clearly shows that there is almost no relationship between the range of the price and the weight. This supports the conclusion above, that the low weight of price depends on the comparatively large value of it.

Next, I have plotted the attribute weight differences between the AHP weights and the linear weights by the range of the attribute as can be seen in Figure 5. In the figure we can see that a majority of the subjects has weighed size more than they did with the AHP weights, and that they have also weighed price less than with the AHP. The differences regarding the other two criteria are negligible.



(a) Size

(b) Price

(c) Distance to University

(d) Distance to leisure location

Figure 5: Plots of AHP and linear weight differences, original criteria

If we look at the differences in weights, it is clear that for both distance attributes the linear weights are very close to the AHP weights. This can be interpreted to mean that subjects were quite capable of estimating how much weight they will put on the distance parameters regarding the apartment. However, looking at the figures we can also see that there is quite a bit of variation in terms of the y-axis. Essentially, what this means is that for a single individual the difference of the AHP and linear weights can be anything between -0,7 and 0,6. The purported difference of nearly zero is only an aggregate value. What this tells us is that for a single subject there is in fact considerable uncertainty regarding the weights of the linear function. For a large group of subjects, on the other hand, we can safely say that on average their estimates would be quite accurate.

For size, the subjects with a small range have largely underestimated the weight they will place on the attribute, whereas for subjects with a large range the result is exactly the converse. For price, subjects across ranges have on average overestimated the weight they will place on the price. Now, this is of course dependent on the fact that we have not scaled the values of the criteria, and hence the weight of price in the linear function is extremely small. However, the purpose of the figures is to point out exactly that. In the absence of scaling, it cannot be concluded that judgments of importance would be related to the linear weights of criteria in any meaningful way.

Finally, let us also look at the best lexicographic predictor for the subject. The best lexicographic predictor is the criterion, which most commonly has the better value for the chosen option. For example, if in 17 out of 20 choices the preferred option has the bigger apartment size (and no other criteria has a larger proportion) then size is the best lexicographic predictor. As Figure 6 shows, for a clear majority of subjects the best lexicographic predictor is the price. This is to be expected, as students usually have limited finances and focus heavily on the price of the apartment. However, using the lexicographic prediction scheme does not produce very good results. This will be analyzed more carefully in the subsection 4.2 Prediction.

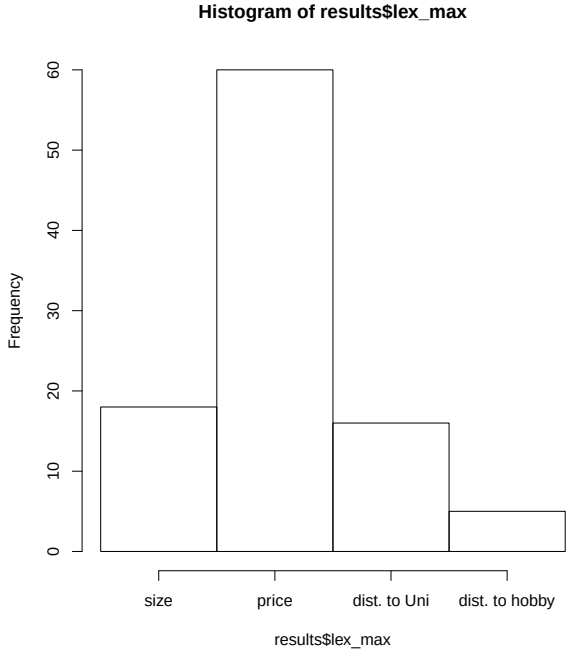


Figure 6: Histogram of the best lexicographic predictor

4.1.1 Consistency

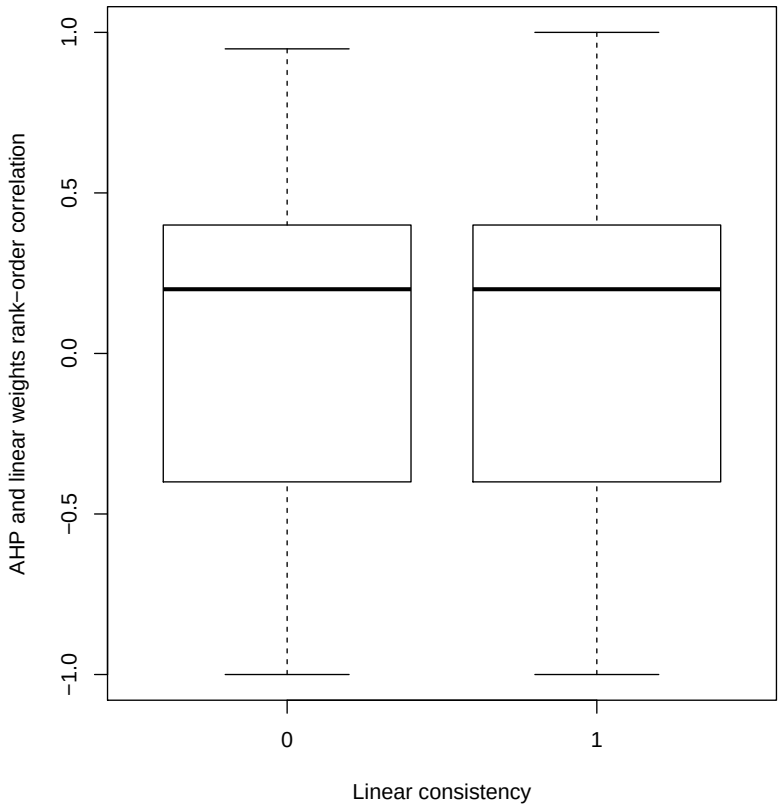


Figure 7: Rank-order correlation of AHP and linear weights by subject's linear consistency, original criteria

In the Figure 7 above is further displayed as a boxplot the relationship of the rank-order correlation of AHP weights with linear weights. On the x-axis is displayed whether the subject was consistent with our linear model. As the figure clearly shows, the medians of both groups are very much equal and even the variance of the rank-order correlations is very similar. We can therefore infer that the linear consistency of the subject is not related to the correlation of AHP weights and linear function weights. Therefore, the predictive value of AHP weights is not dependent on the linear consistency of the subject.

The independence of the AHP weights and linear weights can be checked by a t-test, which gives us $t(117, 875) = -1,0082$ with a p-value of $p = 0,1577$. Therefore we can conclude that the AHP and linear weights are independent.

Looking at the rank-order correlation of AHP weights and linear weights it is obvious that the rank order correlation is independent of the linear consistency of the subject. Therefore, we cannot infer anything from the rank order correlation of a subject regarding the linear consistency, or vice versa. Admittedly, this is

quite surprising, as I would have expected the linear consistent subjects to be at either ends of the correlation measure. Apparently, even linearly consistent subjects are equally incapable of providing ex ante weights that would reflect their choices. However, one must not forget that this holds only in aggregate: as the figure shows, there are subjects in both groups with a perfect rank-order correlation. This means that there were some subjects, whose AHP weights were exactly in the same order as their linear weights.

	Linear model consistency			
Gender	No	Yes	All	% consistent
female	21	18	39	46,2%
male	38	49	87	56,3%
unknown	1	0	1	100,0%
sum	60	67	127	

Table 7: Linear consistency by gender

In Table 7, I display the linear consistency cross-tabulated by gender. In the article by Korhonen et al. (2012), the linear function explained 50% of mens' choices but only 28,4% of women's. In this study, I was unable to replicate the difference between men and women. A χ^2 test of independence gives a result of $\chi^2 = 2,2456$, which has $p = 0,3259$. This means that linear consistency and gender are in this study independent. Hence, it can be concluded that the hypothesis about gender differences in linear consistency was not supported.

I analyze consistency regarding linear transformed questions, control questions and dominance question by two different measures: "strict" and "loose". For strict consistency, any answer besides the correct one is wrong. For loose consistency, indifference is accepted as well. The correct answer is defined in dominance as the dominant alternative, in control questions and linear transformed questions as being the same answer as in the original question.

	Linear consistency			
Control wrong	No	Yes	sum	% consistent
0	19	27	46	58,7 %
1	31	38	69	55,1 %
2	10	2	12	16,7 %
sum	60	67	127	

Table 8: Control consistency by linear consistency

In Table 8 we can see that many subjects have answered one control question wrong. However, as the controls were the last questions in the test, this can plausibly be attributed to the subjects getting tired. The fact that of those who got both of the controls wrong were less consistent with the linear function seems to suggest

that those subjects may have either changed their preferences during the test or used some very unpredictable method of making choices. For control questions there were no differences between the strict and the loose rule.

The maximum likelihood estimation for the control questions gives us $p = (46 * 2 + 69 * 1 + 12 * 0)/127 * 2 = 0,634$. This means that there was a $36,6\%$ chance that a subject chose a different option than in the original question. It has to be noted that the control questions were the last questions to be presented, and hence it is quite likely that this had an adverse effect on the results. Should the control questions have been presented sooner, likely a larger proportion of subjects would have chosen consistently.

	Linear model consistency			
Consistency w/ controls	No	Yes	All	% consistent
No	53	6	59	10%
Yes	7	61	61	90%
sum	60	67	127	

Table 9: Linear consistency by control replacement consistency

As can be seen in Table 9, replacing questions 9 and 14 with the respective control questions does not have a large impact on the linear consistency of the subjects. For only 6 subjects changing the questions changes their status from nonlinear to linear consistent. 7 subjects had the reverse effect, changing from linear to nonlinear. From this can be deduced that the inconsistencies with control questions seem not to be related to the overall consistency of the subject in any meaningful way. A possible explanation for these inconsistencies could be that the subjects simply grow tired over time.

If the inconsistency in the control question answers does not change the status of the nonlinearity, it means that the particular question is not part of the binding constraints of the subject's linear model. In this case, a change in the preference of the subject has no impact on the functional form of the linear model representing his decisions.

Based on the above reasoning it can be assumed that a subject's choices can be represented by a linear function, but that the subjects choose inconsistently on some occasions. Therefore, I have investigated how many responses need to be eliminated from the model to make the subject's choices consistent with a linear value function. This was done by purging iteratively the constraint with the highest shadow cost, until consistency with the linear model was reached. The shadow cost represents how much a given constraint contributes to the value of ϵ, ie. how much effect the constraint has to the linear consistency of the subject. The results of this analysis can be seen below in Table 10.

Strictly preferred pairs	# of removed constraints										sum	Error rate		
	0	1	2	3	4	5	6	7	8	9		0%	5%	10%
13	0	0	0	0	0	0	0	0	1	0	1	0,0%	0,0%	0,0%
14	1	0	0	0	0	0	0	0	0	0	1	100,0%	100,0%	100,0%
15	0	0	0	0	0	0	1	1	0	0	2	0,0%	0,0%	0,0%
16	2	0	0	0	1	0	0	0	0	0	3	66,7%	66,7%	66,7%
17	2	1	0	2	0	0	0	0	0	0	5	40,0%	60,0%	60,0%
18	5	0	0	0	1	0	0	0	0	0	6	83,3%	83,3%	83,3%
19	17	1	1	0	1	1	0	1	0	0	22	77,3%	81,8%	86,4%
20	40	4	3	8	8	9	7	4	3	1	87	46,0%	50,6%	60,6%
sum	67	6	4	10	11	10	8	6	4	1	127	52,8%	57,5%	60,6%

Table 10: Removed constraints to reach linear consistency

As can be seen in Table 10, a clear majority of the subjects are either completely consistent with the linear model or very close to consistency. Allowing for no errors, 52,8% of subjects are consistent with the model. Allowing 5% or 10% error rate, ie. 1 or 2 questions wrong, raises the consistency ratio to 57,5% and 60,6%, respectively. These numbers are in line with the results of (Korhonen et al., 2012), in which consistency ratios of 38,9%, 64,6% and 83,3% were reached in a two-criteria problem setting. Interestingly, in this study the results imply that more subjects were exactly consistent with the linear model, but also that more subjects further away from the linear model. This is seen in the fact that in this setting, allowing errors had only a very minor impact for the proportion of consistent subjects. Therefore, I hypopthesize that some subjects have used a choice strategy, which cannot easily be approximated by a linear value function. Later in the case of prediction power we will see how a common noncompensatory strategy, the lexicographic choice strategy, fares in the prediction task against linear models. Next, I look at the other measures of consistency, meaning dominance, and linear transformed questions.

Dominance wrong	Linear consistency			
	No	Yes	sum	% consistent
0	55	65	120	54,2%
1	3	2	5	40,0%
2	2	0	2	0,0%
sum	60	67	127	

Table 11: Dominance consistency by linear consistency

Table 11 shows that most of the subjects got all of the dominance questions correctly. This implies that they were focusing and paying attention to the task at least enough to notice the dominance relation. As the dominance questions were

just after the questions used for prediction, it can be concluded that at least up to this point in the questionnaire, the responses of the students can be trusted.

The maximum likelihood for dominance gives us $p = (120*2+5*1+2*0)/127*2 = 0,965$ meaning that there was a $3,5\%$ chance that a subject did not notice a dominance relation and chose the dominated option. As dominance is normatively very compelling, this can be considered as a baseline measure of the level of awareness of the subjects after answering the first 20 questions (the dominance questions were numbers 21-22).

In Table 12 we can see the consistency in the linear transformations on the bound. A majority of the respondents answered all of the linear tranformations in the same way as the original questions. Using the loose form of consistency does not change results significantly. However, due to a software bug in the questionnaire, six respondents did not get to answer all the four linear transformed questions. This could cause a bias in the results.

	Linear consistency			
Onbound strict	No	Yes	sum	% consistent
0	45	53	98	54,1%
1	15	14	29	48,3%
sum	60	67	127	

Table 12: Onbound strict consistency by linear consistency

	Linear consistency			
Overbound strict	No	Yes	sum	% consistent
0	40	58	98	59,2%
1	20	9	29	31,0%
sum	60	67	127	

Table 13: Overbound strict consistency by linear consistency

Table 13 displays the consistency of the linear transformations over the bounds set by subjects. Essentially, the distribution is the same as it was for transformations on the bound. Using loose form of consistency had almost no effect to the results. However, these questions suffered from the same bug as the previous ones. Hence, these results cannot be used to conclude much about the performance of subjects outside the bounds they had defined.

The maximum likelihood principle for the overbound questions gives us $p = (99 * 2 + 28 * 1)/127 * 2 = 0,890$ and $p = (98 * 2 + 29 * 1)/127 * 2 = 0,886$ for the loose and strict conditions, respectively. This means that there was a $11,0\%$ chance in the loose condition and a $11,4\%$ chance in the strict condition that the subject chose inconsistently when the question was scaled so that on criterion went over the bounds defined by the subject. Here, the loose condition accepts indifference as the same answer, whereas strict defines indifference as inconsistent.

4.1.2 Prediction

I have compared the predictive efficacy of the different methods below in Tables 14 and 15. Table 14 compares the estimated weights against equal weights, whereas Table 15 compares estimated weights to AHP weights.

Estimated weights	# correct predictions	Equal weights								
		3	4	5	6	7	8	9	10	sum
	4	0	0	0	0	1	0	0	0	1
	5	0	0	0	0	1	1	1	0	3
	6	0	0	0	1	1	0	2	0	4
	7	1	0	0	3	5	8	4	2	23
	8	0	1	3	2	5	7	4	0	22
	9	0	0	0	3	6	5	12	5	31
	10	0	0	0	2	1	2	6	4	15
	sum	1	1	3	11	20	23	29	11	99

Table 14: Prediction power of estimated weights vs. equal weights, original criteria

Estimated weights	# correct predictions	AHP weights								
		3	4	5	6	7	8	9	10	sum
	4	0	0	0	0	0	1	0	0	1
	5	0	0	0	0	0	1	2	0	3
	6	0	0	0	2	1	1	0	0	4
	7	1	0	1	2	9	6	2	2	23
	8	0	0	3	5	8	5	1	0	22
	9	0	0	1	3	5	6	10	6	31
	10	0	0	1	1	1	4	2	6	15
	sum	1	1	3	11	20	23	29	11	99

Table 15: Prediction power of estimated weights vs. AHP weights, original criteria

Figure 8 illustrates the distribution of $x_{i1} - x_{i2}$, where $x_{ik}, k = 1, 2$ is the number of correct predictions with $k = 1$ referring to estimated weights and $k = 2$ referring to the equal weights.

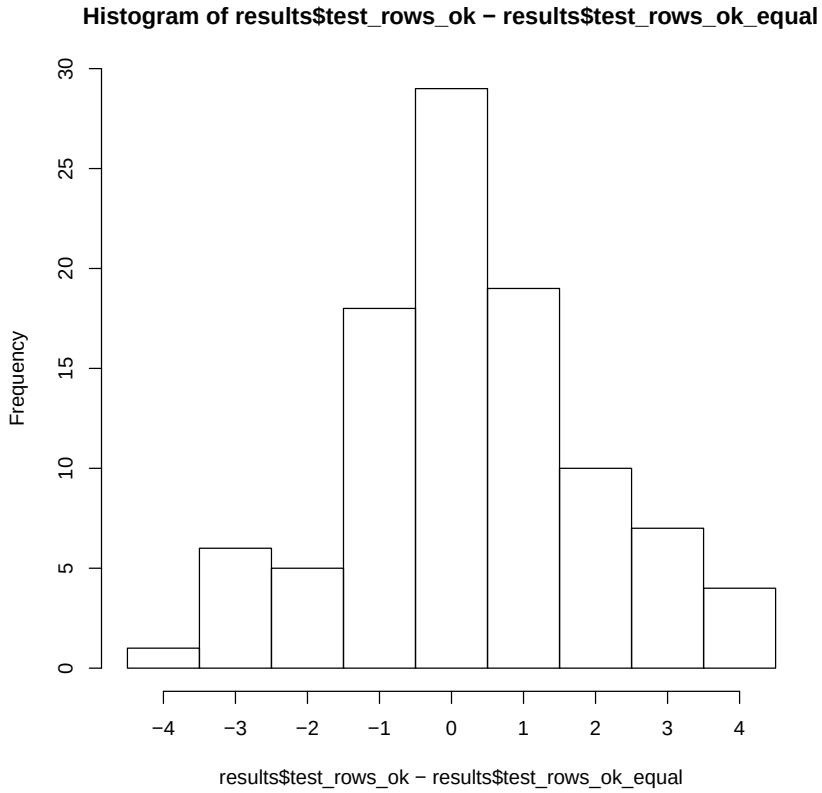


Figure 8: Histogram of estimated weights vs. equal weights, original criteria

As the figure shows, for 29 subjects (out of 99, ie. 29,3 %) the predictive power of equal weights was the same as the estimated weights. For 40 subjects (40,4 %) estimated weights proved better. For 30 subjects (30,3 %) the equal weights fared better.

I use a matched-sample t-test to investigate whether the prediction power is better when using estimated weights from the linear model than the equal weights provided by the participants. We have the following hypotheses: $H_0 : \mu_1 - \mu_2 = 0$ $H_1 : \mu_1 - \mu_2 > 0$

Here μ_1 is the number of correct predictions in the population when using estimated weights, and μ_2 is the number of correct predictions using AHP weights. As we have 99 respondents with all the answers in the population, the degrees of freedom in the t-test is $99 - 1 = 98$. The value for the t-test is $t(98) = 1,5265$ with a p-value of $p = 0,06505 > 0,01$. Hence, it must be concluded that equal weights fare equally well with estimated weights. The mean of $\mu_1 - \mu_2 = 0,2626$, so estimated weights are slightly better, even though not statistically significantly.

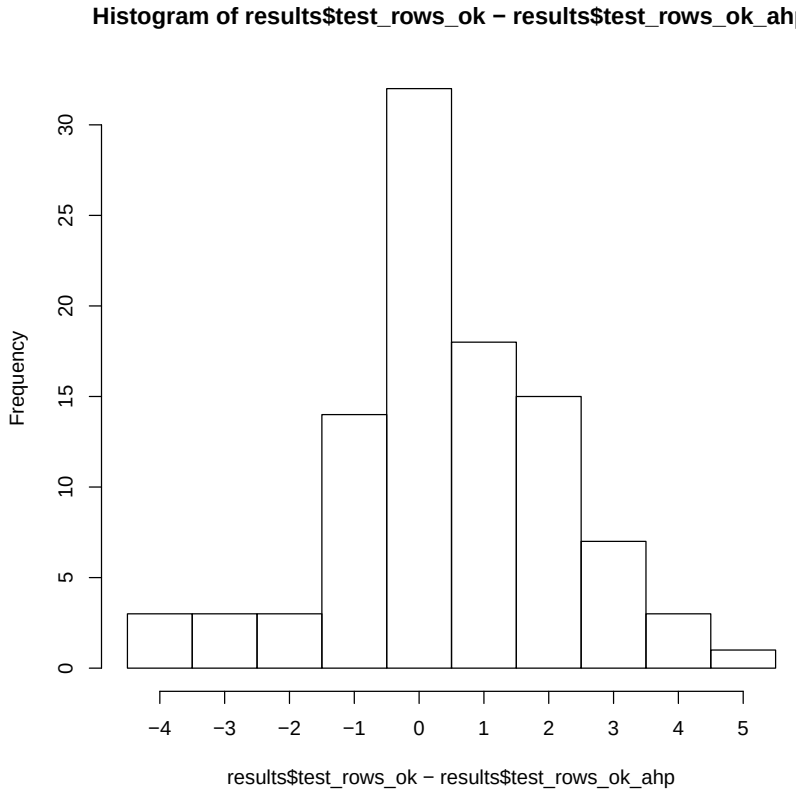


Figure 9: Histogram of estimated weights vs. AHP weights, original criteria

Figure 9 illustrates the distribution of $x_{i1} - x_{i3}$, where $x_{ik}, k = 1, 3$ is the number of correct predictions with $k = 1$ referring to estimated weights and $k = 3$ referring to the AHP weights. Again, I used a matched pairs t-test to compare the prediction power of estimated weights vs. AHP weights. I have the following hypotheses: $H_0 : \mu_1 - \mu_3 = 0$ $H_1 : \mu_1 - \mu_3 > 0$

Here, as before, μ_1 refers to the number of correct predictions in the population when using estimated weights, and μ_3 refers to the number of correct predictions when using equal weights. Again, as we have the same population, the degrees of freedom is 98. The value for the t-test is $t(98) = 2,6087$ with a p-value of $p = 0,005255 < 0,01$. Hence, it can be concluded that the estimated weights are better. The mean of $\mu_1 - \mu_3 = 0,4545$.

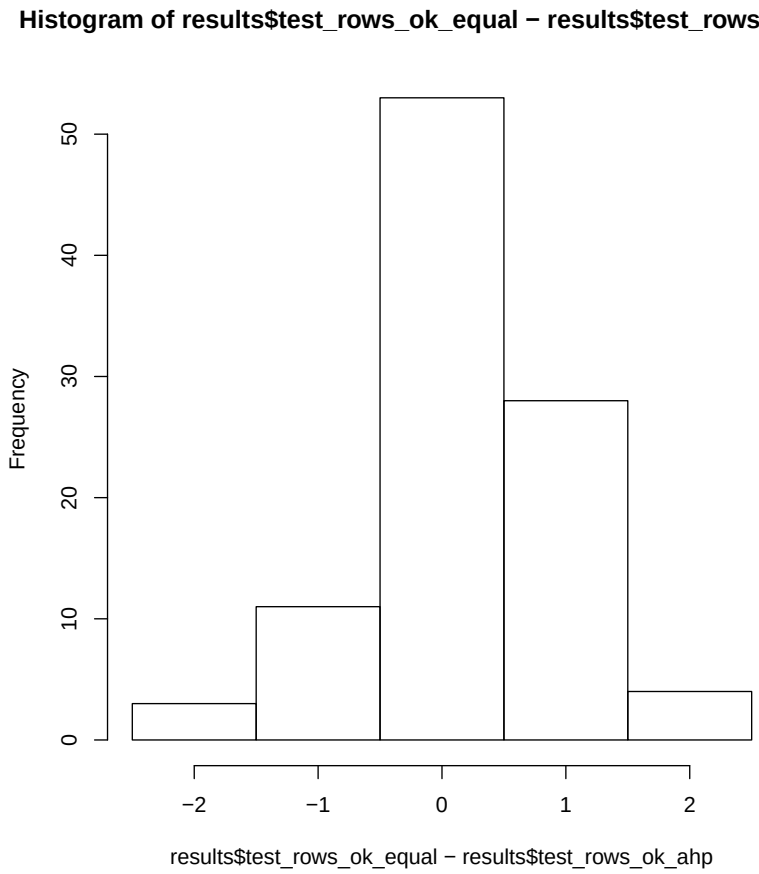


Figure 10: Histogram of equal weights vs. AHP weights, original criteria

Then, I compare equal weights to AHP weights. Figure 10 illustrates the distribution of $x_{i2} - x_{i3}$where $x_{ik}, k = 2, 3$ is the number of correct predictions with $k = 2$ referring to AHP weights and $k = 3$ referring to the AHP weights.. We have the hypotheses $H_0 : \mu_2 - \mu_3 = 0$ $H_1 : \mu_2 - \mu_3 > 0$ Once again, the matched pairs t-test is done with 98 degrees of freedom. The value of the test is $t(98) = 2,375$ with a p-value of $p = 0,009748 < 0,01$. Hence, it can be concluded that equal weights are better for prediction than AHP weights. The mean of $\mu_2 - \mu_3 = 0,1919$. However, it should be noted that the difference in the means is quite small.

Estimated weights	# correct predictions	LEX model									
		2	3	4	5	6	7	8	9	10	sum
	4	0	0	0	0	0	0	1	0	0	1
	5	0	0	0	1	1	0	0	1	0	3
	6	0	0	1	0	1	2	0	0	0	4
	7	1	1	1	4	1	9	4	1	1	23
	8	0	1	0	1	3	11	5	1	0	22
	9	0	1	1	5	2	4	5	10	3	31
	10	0	0	0	0	2	2	3	5	3	15
	sum	1	3	3	11	10	28	18	18	7	99

Table 16: Prediction power of estimated weights vs. LEX model, original criteria

Table 16 shows the differences in predictive power between estimated weights and the lexicographic model. Figure 11 illustrates the distribution of $x_{i1} - x_{i4}$, where $x_{ik}, k = 1, 4$ is the number of correct predictions with $k = 1$ referring to estimated weights and $k = 4$ referring to the lexicographic model (LEX).

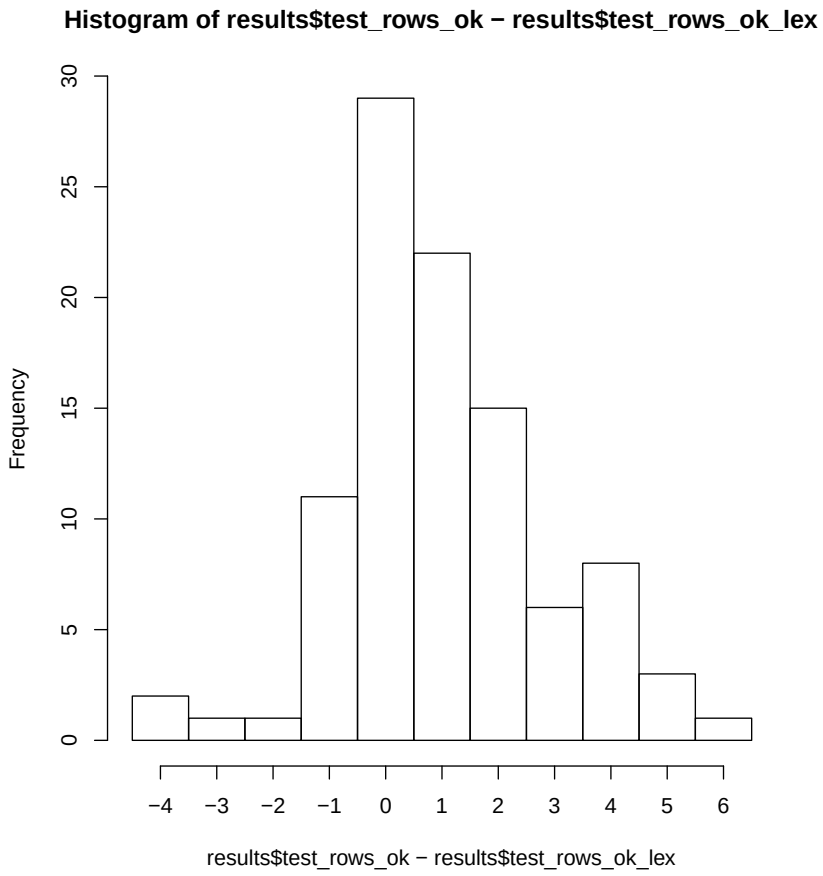


Figure 11: Histogram of estimated weights vs. LEX model, original criteria

As the Figure 11 shows, for 29 subjects (out of 99, ie. 29,3 %) the predictive power of the LEX model was the same as the estimated weights. For 55 subjects (55,6 %) estimated weights proved better. For 15 subjects (15,1 %) the LEX model fared better.

I used a matched-sample t-test to investigate whether the prediction power is better when using estimated weights from the linear model than the LEX model. I have the following hypotheses: $H_0 : \mu_1 - \mu_4 = 0$ $H_1 : \mu_1 - \mu_4 > 0$

Here μ_1 is the number of correct predictions in the population when using estimated weights, and μ_4 is the number of correct predictions using LEX model. As we have 99 respondents with all the answers in the population, the degrees of freedom in the t-test is $99 - 1 = 98$. The value for the t-test is $t(98) = 5,4058$ with a p-value of $p = 2,27 * 10^-7 < 0,0001$. Hence, we conclude that estimated weights outperform the lexicographic model by a large margin. The mean of $\mu_1 - \mu_4 = 1$, so for each subject, the estimated weights on average predict one answer more than the lexicographic model.

Next, I compare the lexicographic model to the predictions made by the AHP model. As so far the AHP weights have been the worst predictor, by using them as a comparison point we can see whether the LEX model will be the worst. If the AHP model proves better, then the LEX model will be the worst performer. If LEX wins out, then we still need to compare it to equal weight predictions to finalize the ranking of the different methods.

We get the following Table 17 for the comparison.

	# correct predictions	LEX model									
		2	3	4	5	6	7	8	9	10	sum
AHP weights	3	0	1	0	0	0	0	0	0	0	1
	4	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	2	0	1	2	1	0	6
	6	0	1	1	3	5	3	0	0	0	13
	7	0	0	2	2	3	16	1	0	0	24
	8	1	1	0	2	1	6	10	3	0	24
	9	0	0	0	0	1	2	2	12	0	17
	10	0	0	0	2	0	0	3	2	7	14
	sum	1	3	3	11	10	28	18	18	7	99

Table 17: Prediction power of AHP weights vs. LEX model, original criteria

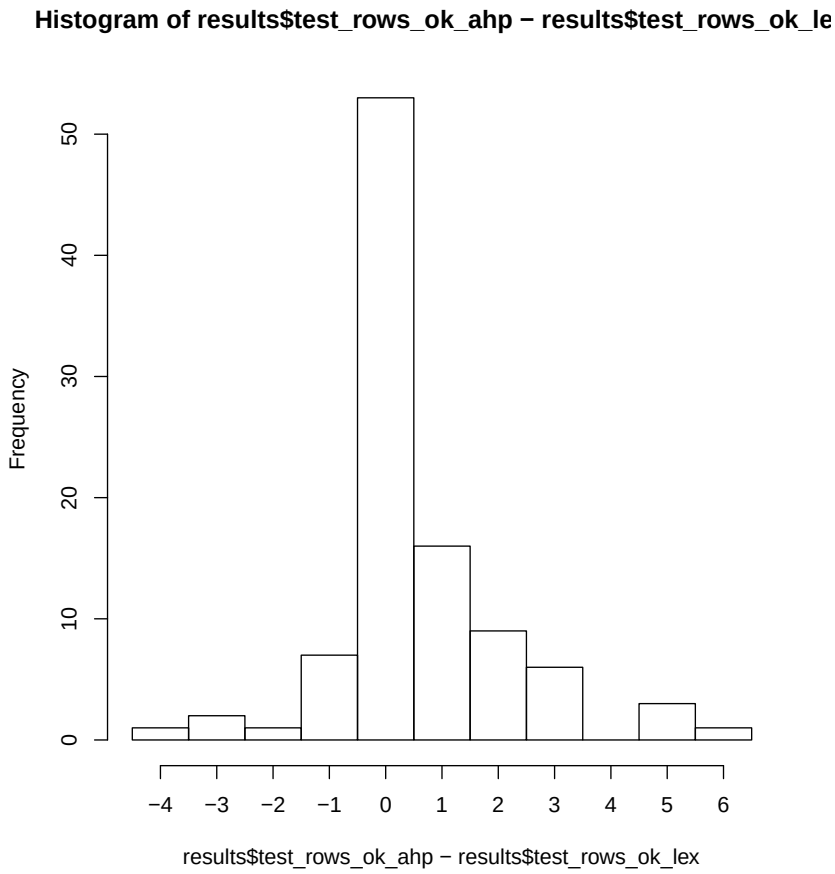


Figure 12: Histogram of AHP weights vs. LEX model, original criteria

As Figure 12 shows, for 53 subjects (53,5 %) the prediction power of the LEX model is equal to the power using AHP weights. For 11 subjects (11,1 %) the LEX model is better, and for 35 subjects (35,5 %) AHP weights are better.

I use a matched-sample t-test to investigate whether the prediction power is better when using estimated weights from the linear model than the equal weights provided by the participants. We have the following hypotheses: $H_0 : \mu_3 - \mu_4 = 0$ $H_1 : \mu_3 - \mu_4 > 0$

Here μ_3 is the number of correct predictions in the population when using AHP weights, and μ_4 is the number of correct predictions using the LEX model. As we have 99 respondents with all the answers in the population, the degrees of freedom in the t-test is $99 - 1 = 98$. The value for the t-test is $t(98) = 3,5232$ with a p-value of $p = 0,0003247 < 0,001$. Hence, it can be concluded that AHP weights outperform the lexicographic model clearly. The mean of $\mu_3 - \mu_4 = 0,545$. In Figure 12 we can clearly see that for just four subjects the LEX model is clearly better (predicting at least two more choices than the AHP weights), whereas the converse is true for 19 subjects (AHP weights predicting at least two more than the LEX model). The histogram is clearly skewed to the right.

Finally, I have tested whether forcing the linear weights to be inconsistent with the AHP weights is better than using equal weights. This comparison was performed

on those subjects who had a negative rank-order correlation coefficient between linear and AHP weights. There were 40 such subjects. The comparison of the methods gives us the following table:

Incons. weights	# correct predictions	Equal weights								
		3	4	5	6	7	8	9	10	sum
	6	1	0	0	1	1	1	1	0	5
	7	0	0	1	1	4	2	1	0	9
	8	0	0	0	0	1	3	2	0	6
	9	0	0	0	1	2	1	5	2	11
	10	0	0	0	1	0	1	6	1	9
	sum	1	0	1	4	8	8	15	3	40

Table 18: Prediction power of equal weights vs. inconsistent weights

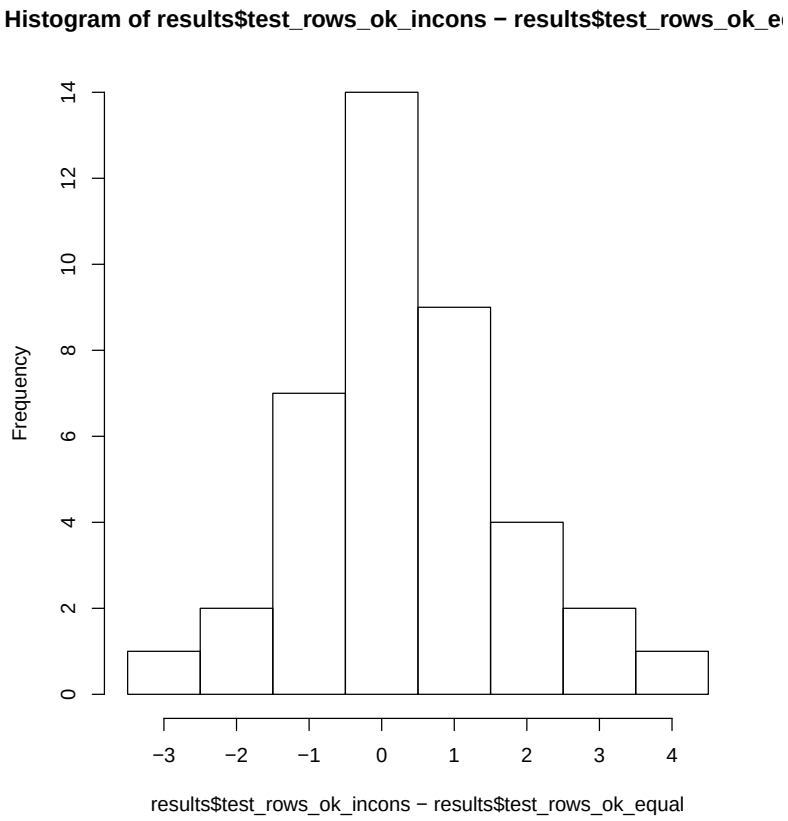


Figure 13: Histogram of equal weights vs. inconsistent weights, original criteria

Seen in Figure 13 , for 14 subjects (35,9 %) the prediction power of the inconsistent weights model is equal to the power using equal weights. For 16 subjects

(40,0 %) the inconsistent weights model is better, and for 10 subjects (27,0 %) equal weights are better.

I use a matched-sample t-test to investigate whether the prediction power is better when using estimated weights from the linear model than the equal weights provided by the participants. We have the following hypotheses: $H_0 : \mu_5 - \mu_2 = 0$ $H_1 : \mu_5 - \mu_2 > 0$

Here μ_2 is the number of correct predictions in the population when using equal weights, and μ_5 is the number of correct predictions using the inconsistent weights model. As we have 40 respondents with all the answers in the population and a negative AHP-linear rank-order correlation, the degrees of freedom in the t-test is $40-1 = 39$. The value for the t-test is $t(39) = 1,4463$ with a p-value of $p = 0,07804$. The mean of $\mu_5 - \mu_2 = 0,325$. This means that we do not have enough information to conclude the inconsistent weights to be better, but they fare comparably to equal weights. Changing the cut-off value for the rank order correlation did not affect the conclusion.

Including only the 99 respondents who answered to all the 20 questions we obtain the following table:

	Epsilon ok			
Correct predictions	No	Yes	sum	cumulat. predicted subj. %
10	4	11	15	15,2
9	12	19	31	46,5
8	14	8	22	68,7
7	13	10	23	91,9
6	4	0	4	96,0
5	2	1	3	99,0
4	1	0	1	100,0
sum	50	49	99	

Table 19: Predictability of the last 10 choices by linear consistency, 99 full respondents, original criteria

As can be seen in Table 19, the model does a pretty good job of predicting the choices of the respondents.

	Gender			
Correct predictions	female	male	unknown	sum
10	4	11	0	15
9	10	20	1	31
8	6	16	0	22
7	12	11	0	23
6	1	3	0	4
5	2	1	0	3
4	0	1	0	1
sum	35	63	1	99

Table 20: Predictability of the last 10 choices by gender, 99 full respondents, original criteria

As Table 20 shows, the distributions of the predictability of the function for men and women seem very much the same. When using the maximum likelihood principle, we obtain the following probabilities for a correct prediction: $p_{men} = 0,8286$ and $p_{women} = 0,7943$. As their 95 % confidence intervals include each other, on a 5 % risk level I cannot conclude any difference.

I use the maximum likelihood principle to investigate the prediction success for both the linear consistent and nonconsistent populations separately. For the linear consistent case we obtain $p_{est1} = 0,857$ and $p_{est2} = 0,778$. The probability based on the whole sum distribution is $p = 0,817$. Here the subscript *est* refers to using estimated weights, 1 refers to the subject being linear consistent and 2 to the subject not being linear consistent.

The same maximum likelihood principle is used to analyze the predictive power of the equal weights and AHP weights. For them, we obtain $p_{eq1} = 0,837$, $p_{eq2} = 0,746$, $p_{ahp1} = 0,816$ and $p_{ahp2} = 0,728$. From these results it can be seen that the estimated weights are narrowly better than equal weights, and that AHP weights lose out to both of the aforementioned methods. Below in Table 21 we have displayed the previous maximum likelihood values along with their 95% confidence intervals. The same information is presented below graphically in Figure 14.

	Interval min	ML estimate	Interval max
Nonlinear subjects			
Estimated	0,645	0,778	0,871
Equal	0,611	0,746	0,846
AHP	0,592	0,728	0,832
LEX	0,524	0,662	0,777
Linear consistent subjects			
Estimated	0,733	0,857	0,929
Equal	0,710	0,837	0,915
AHP	0,686	0,816	0,900
LEX	0,639	0,773	0,868
All subjects			
Estimated	0,730	0,817	0,881
Equal	0,701	0,791	0,859
AHP	0,680	0,772	0,843
LEX	0,622	0,717	0,796

Table 21: 95% confidence intervals of the ML estimates of a successful prediction, Wilson estimation, original criteria

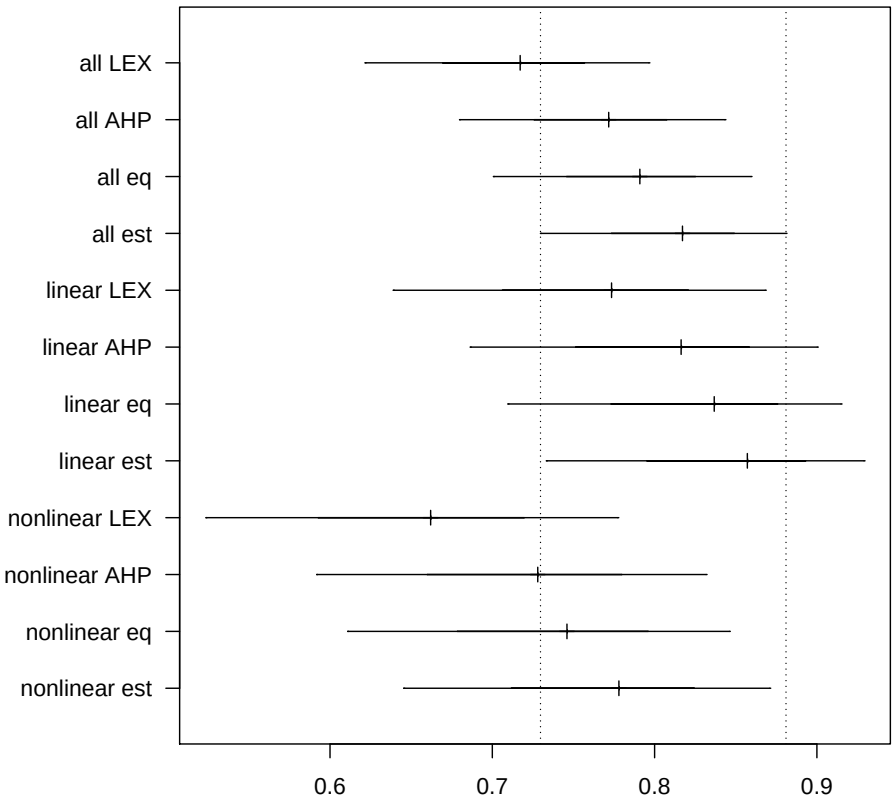


Figure 14: ML estimates with confidence regions by different methods and populations, Wilson CI estimation, original criteria

Figure 14 above shows the maximum likelihood estimates of the different prediction methods with their 95% confidence intervals, computed with the Wilson method. The vertical dotted lines represent the confidence interval minimum and maximum value for the maximum likelihood of a correct prediction when using the estimated weights from the linear model over all subjects. This could be named as the baseline, in that it uses our proposed method and the data from all of the subjects.

The figure shows clearly that the estimated weights outperform the other methods for all three categories: the estimated weights' confidence interval is towards the higher end of the likelihood range in comparison to the other methods. The equal weights also outperforms AHP weights, which is the worst of the three methods, although not by a significant margin. Additionally, it ought to be noted that all three methods are much better in predicting choices of those participants who are consistent with the linear model. This is to be expected, as a nonlinear subject can use any nonlinear strategy, even a random one. A linear consistent subject, in contrast, has to consistently use the linear function or a function which produces the same results. On a final note, the figure also displays that all of the three methods are better than chance even when predicting the choices of nonlinear subjects. Interestingly, the differences in the prediction power of the methods seems to be irrespective of the linear consistency of the subject.

On the other hand, all of the methods still use the linear model, only with different weights so they should have the same kind of model bias. Therefore, I have also examined whether a lexicographic account fares better in predicting some of the choices. This examination is based on the idea that perhaps nonlinear subjects might use a lexicographic account, which is easy to check.

Of course, the possibilities of nonlinear choice strategies is infinite, so I cannot check for all of them. The lexicographic strategy has been picked here due to its simplicity and ease of use. It is meant to be a comparison standard for our linear model, not as a serious account to explain the choice strategy of nonlinear subjects.

It turns out, as Figure 14 clearly displays, that the LEX model loses out to all the other methods, both regarding linear and nonlinear subjects. This strengthens our conclusion that the nonlinear subjects are not using a lexicographic strategy. The lexicographic method has a maximum likelihood estimate of $p_{lex1} = 0,773$ and $p_{lex2} = 0,662$ for linear and nonlinear subjects, respectively. Considering that chance alone ought to give us $p = 0,500$, the method's predictive power is not at all that impressive.

In Table 22 I have displayed the prediction success with a lexicographic model.

	Epsilon ok			
Correct predictions	No	Yes	sum	cumulat. predicted subj. %
10	1	6	7	7,1
9	5	13	18	25,3
8	8	10	18	43,4
7	17	11	28	71,7
6	7	3	10	81,8
5	7	4	11	92,9
4	2	1	3	96,0
3	2	1	3	99,0
2	1	0	1	100,0
sum	50	49	99	

Table 22: Predictability of the last 10 choices by linear consistency with lex model, 99 full respondents

4.1.3 CRT as predictor

In Table 23 below, we display the linear consistency of the respondents over classes of subjects defined by the Cognitive Reflection Test. As before, the classes are defined so that the low class consists of subjects with 0 correct answers, the medium class has 1-2 correct answers and the high class has all three answers correct.

	Linear model consistency			
CRT class	Yes	No	All	% consistent
low	10	8	18	44%
medium	30	33	63	52%
high	20	26	46	57%
sum	60	67	127	

Table 23: Linear consistency by CRT class

As can be seen in the table, the different groups have quite small differences in terms of consistency. To look at the significance of these differences, I look at the χ^2 test of independence. The test gives us a result of $\chi^2 = 0,7642$, which gives us $p = 0,6824$. Therefore, the null hypothesis cannot be rejected and the categories are independent. Dividing the CRT scale into two groups with results 0-1 and 2-3 does not change the result ($\chi^2 = 3,1162$ with $p = 0,2105$). Hence, it can be concluded that the hypothesis about the CRT being the link between linear consistency and gender was not supported.

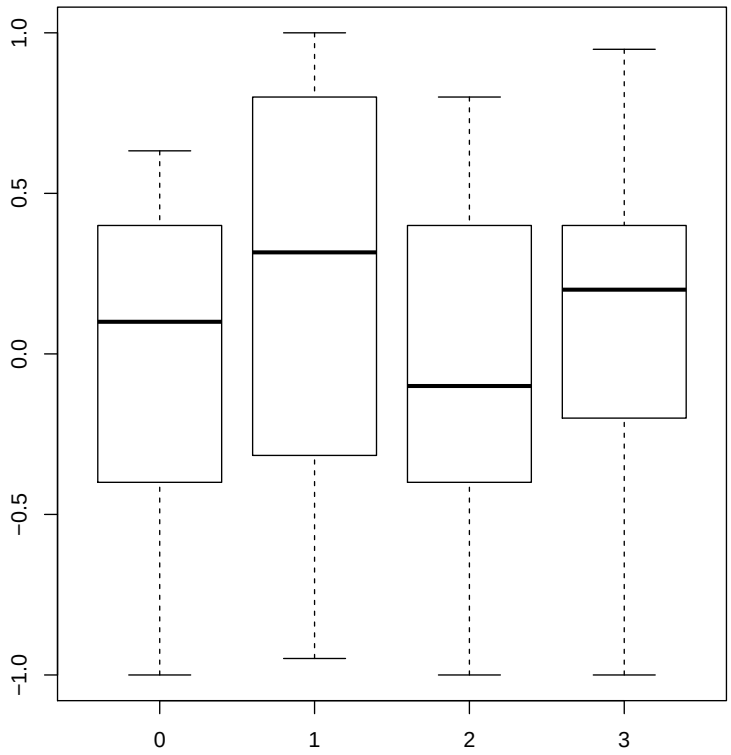


Figure 15: Rank-order correlation of AHP and linear weights by subject's CRT score, original criteria

As we have seen in Figure 15, the rank order correlation is also independent of the CRT score of the subject. It seems that the cognitive style of the subject also does not hold any relation to how informative the AHP weights of the subject are. As a conjecture I entertained that the CRT would predict either linear consistency, or higher AHP-linear correlation, or both. The motivation for this claim was that it might be that more reflective, analytical subjects would be better at predicting their own preferences. However, as figure 15 shows this was clearly not the case. This suggests that, going back to the heuristic override process in figure 1 by Stanovich and West (2008), the low correlation of the AHP and linear weights does not seem to reflect a failure to provide a System 2 response. Had there been a difference between the different CRT groups, it could have been hypothesized that the discrepancy in the weights arises due to an inability to override a heuristic response, for example. Now, as there are not differences between CRT groups, obviously such a hypothesis cannot be supported. As Table 24 shows, there are no differences in predictability between CRT classes.

	CRT class			
Correct predictions	low	medium	high	sum
10	2	6	7	15
9	4	15	12	31
8	3	10	9	22
7	3	14	6	23
6	0	3	1	4
5	0	2	1	3
4	0	1	0	1
sum	12	51	36	99

Table 24: Predictability of the last 10 choices by CRT class, 99 full respondents, original criteria

Finally, in Figure 16 is displayed the ML estimation of the predictive accuracy for different methods.

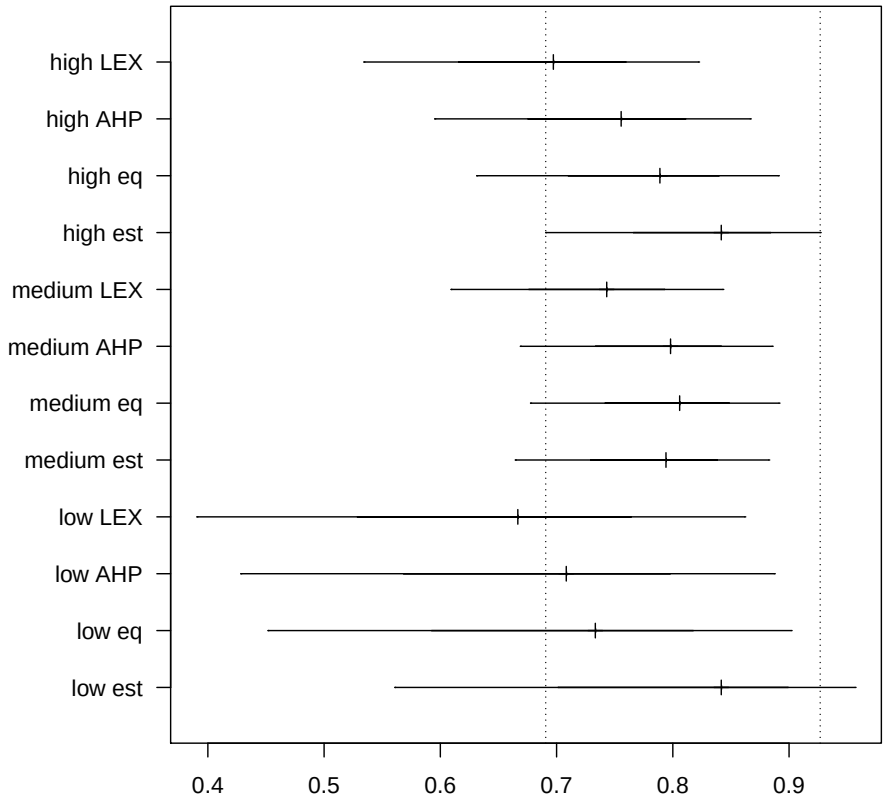


Figure 16: ML estimates with confidence regions by different methods and CRT groups, Wilson CI estimation, original criteria

As can be seen in Figure 16, the maximum likelihood of different methods does not differ very much between different CRT groups. The likelihood confidence interval for the low group is very wide, as the low group consists of only 12 individuals. This is a serious drawback in the data. Due to the wide confidence interval, it cannot be concluded from the data that there would be significant differences between the CRT groups.

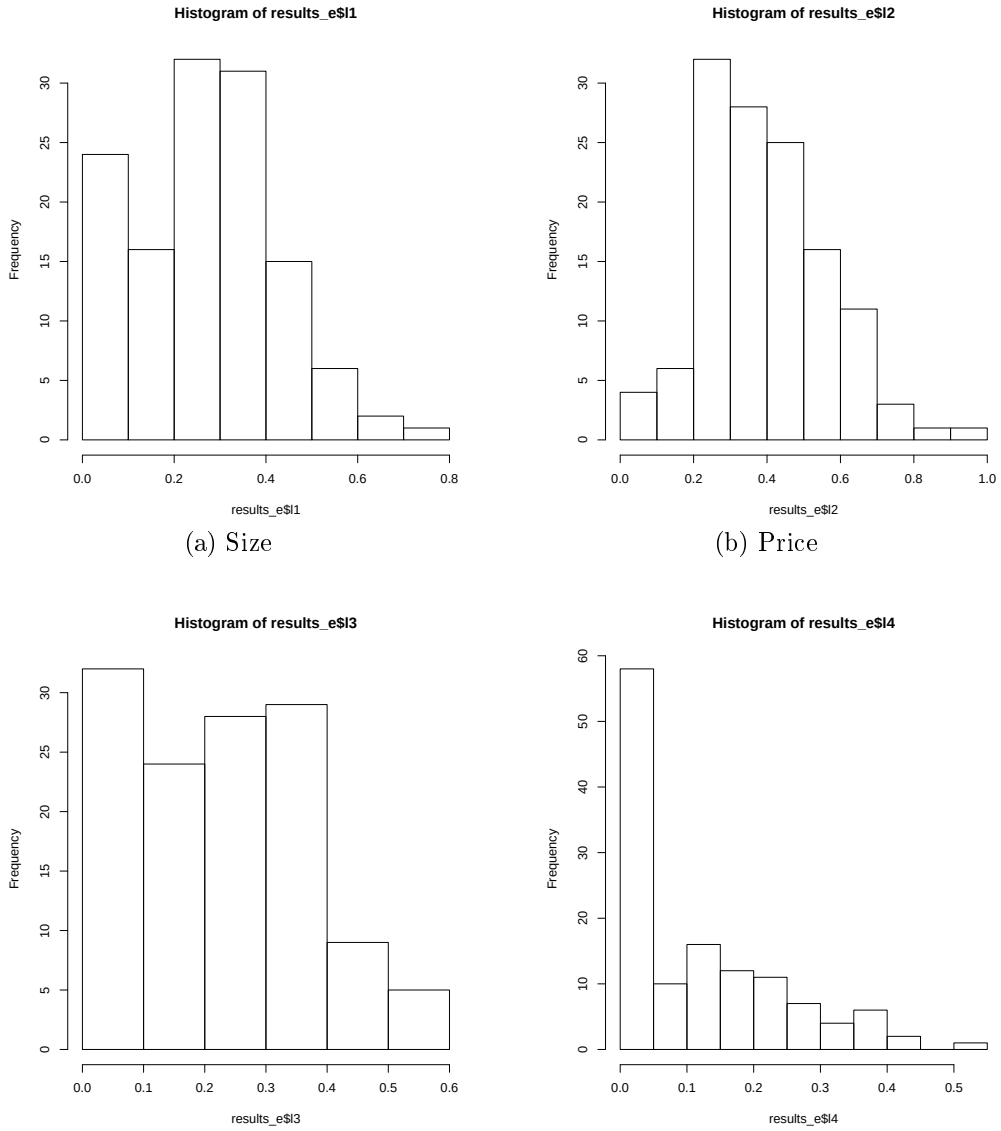
However, we can see that the estimated weights prove to be the best method with both high and low CRT scoring subjects. With medium scorers, the AHP and equal weights methods emerge as the front-runners. However, the differences between methods are extremely small.

All in all, the lack of differences between groups means that it must be concluded that the CRT score of the subject is not to be related to the predictive efficacy of the methods. Even though the medians of high and low scoring CRT groups differ, due to the very wide confidence interval of the low group this difference cannot be considered to be very informative.

4.2 Results using scaled values

Now, let us turn to the normatively more appealing option of using scaled criteria values. In this study, the criteria were scaled to the 0-1 interval by dividing the criterion value by the maximum value in the choice set.

The following histograms for the linear function weights can be obtained when using scaled criteria:



(a) Size (b) Price

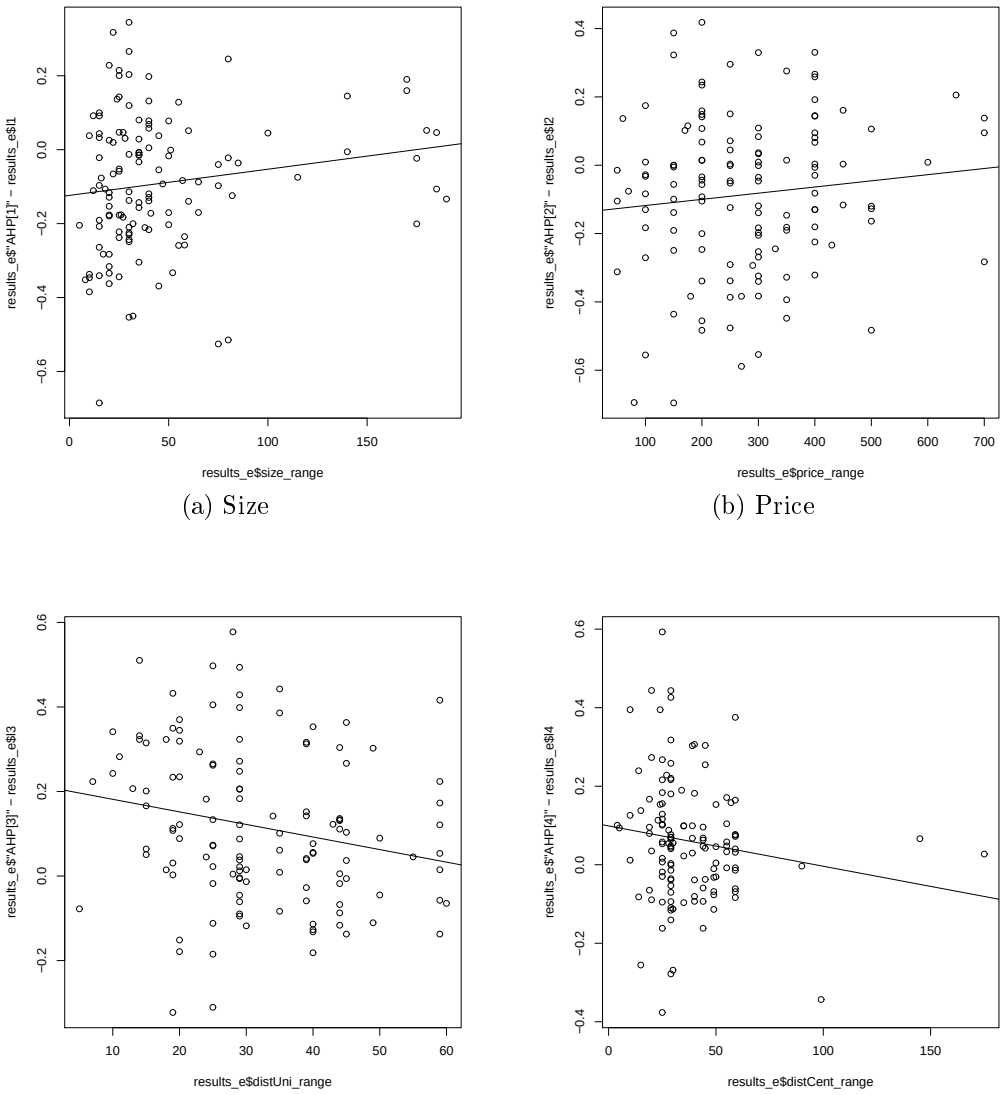
(c) Distance to University (d) Distance to leisure location

Figure 17: Histograms of linear value function weights with scaled criteria

Figure 17 displays histograms of the weights of different criteria when using criteria values scaled to the 0-1 interval. As the criteria are normalized, the weights then reflect how much each of them has been emphasized in the subject's choices. We can see in the Figure that price and size have both relatively high values, in comparison to the distance criteria. This means that the distance criteria have been less important for making a decision. We can also compare Figure 17 to Figure 3, which displays the weights for original valued criteria. With original values the weight of price was extremely low, because using original values means that price is going to be much larger in absolute terms than the other criteria. Price ranges often from 200 to 500 euros, whereas an apartment size might be from 20 to 50 square

meters. So, to have the same impact, the weight of the size criterion would need to be ten times as large as the weight for price.

Next, I have plotted the attribute weight differences between the AHP weights and the linear weights by the range of the attribute as can be seen in Figure 5. In the figure we can see that a majority of the subjects has weighed size a little bit more than they did with the AHP weight. Distance was weighed less than with the importance judgments.



(a) Size

(b) Price

(c) Distance to University

(d) Distance to leisure location

Figure 18: Plots of AHP and linear weight differences with scaled criteria

As can be seen in Figure 18, subjects have weighed size on average quite as they did with the AHP criteria. However, there is a lot of variation between subjects,

and some have weighed some criteria with tens of percents more or less than they initially thought. This causes some concern; subjects seem to be relatively bad at anticipating the weights of the linear value function. Of course, this is not necessarily a problem for prediction, if the model still works even with the ex ante defined criteria values. This will be discussed later in the section about prediction. Finally, it seems that there are no significant effects due to range. As can be seen in Figure 18, the regression lines for range effects are quite flat and there is a lot of variation between subjects, meaning that the difference of AHP and linear weights is independent of the range defined by the subject.

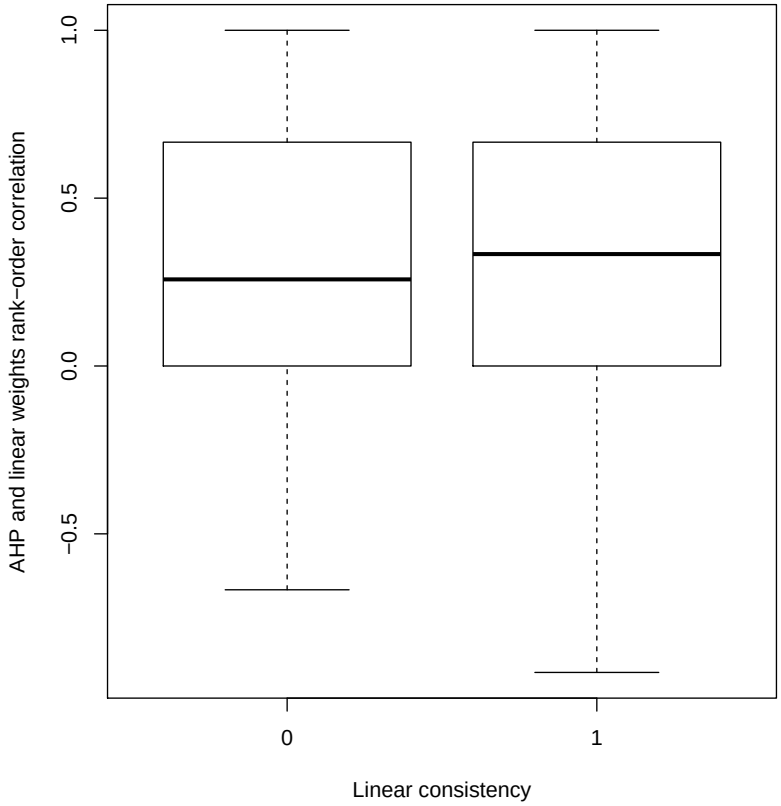


Figure 19: Rank-order correlation of AHP and linear weights by subject's linear consistency, scaled criteria

In Figure 19 above I have displayed the relationship of the rank-order correlation of AHP weights with linear weights. In the boxplot the x-axis shows whether the subject was consistent with our linear model, while the y-axis displays the Kendall tau value. As the figure clearly shows, the medians of both groups are very much equal. However, there is a clear difference in the shape of the distribution: it can be seen that linearly consistent subjects have a distribution skewed towards the positive end. However, there is no statistically significant difference in the means of Kendall correlations between the groups ($t(124,905) = -0,8506, p = 0,1983$). They tend

to be more consistent with the AHP weights they have given than nonconsistent subjects. In the linear consistent group there are only some subjects, who have a negative correlation between the linear weights and AHP weights. This quite intriguing, and suggests that the AHP weights are more representative of linear consistent subjects. We will see below in the section regarding prediction, whether this difference has an effect for the predictive power of the AHP and linear weights across subject groups differing in their linear consistency.

4.2.1 Prediction with scaled criteria

Figure 8 and Table 25 illustrate the distribution of $x_{i1} - x_{i2}$, where $x_{ik}, k = 1, 2$ is the number of correct predictions with $k = 1$ referring to estimated weights and $k = 2$ referring to the equal weights.

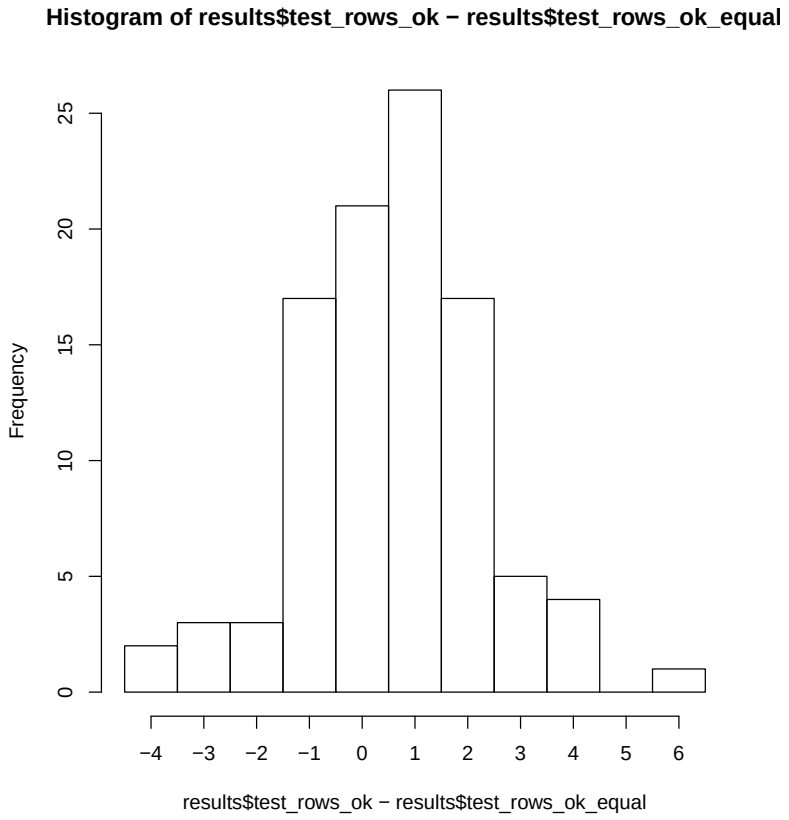


Figure 20: Histogram of estimated weights vs. equal weights, scaled criteria

Estimated weights	# correct predictions	Equal weights							
		4	5	6	7	8	9	10	sum
	4	0	0	0	0	1	0	0	1
	5	0	1	0	0	1	1	0	3
	6	0	2	0	2	0	0	0	4
	7	1	2	5	4	7	2	2	23
	8	0	0	6	4	6	5	1	22
	9	0	2	2	6	11	7	3	31
	10	1	0	2	2	3	4	3	15
	sum	2	7	15	18	29	19	9	99

Table 25: Prediction power of estimated weights vs. equal weights, scaled criteria

As the figure shows, for 21 subjects (out of 99, ie. 21,2 %) the predictive power of equal weights was the same as the estimated weights. For 53 subjects (53,5 %) estimated weights proved better. For 25 subjects (25,3 %) the equal weights fared better.

I use a matched-sample t-test to investigate whether the prediction power is better when using estimated weights from the linear model than the equal weights provided by the participants. We have the following hypotheses: $H_0 : \mu_1 - \mu_2 = 0$ $H_1 : \mu_1 - \mu_2 > 0$

Here μ_1 is the number of correct predictions in the population when using estimated weights, and μ_2 is the number of correct predictions using AHP weights. As we have 99 respondents with all the answers in the population, the degrees of freedom in the t-test is $99 - 1 = 98$. The value for the t-test is $t(98) = 3,3065$ with a p-value of $p = 0,00066 < 0,01$. Hence, it can be concluded that equal weights fare worse than estimated weights. The mean of $\mu_1 - \mu_2 = 0,5758$.

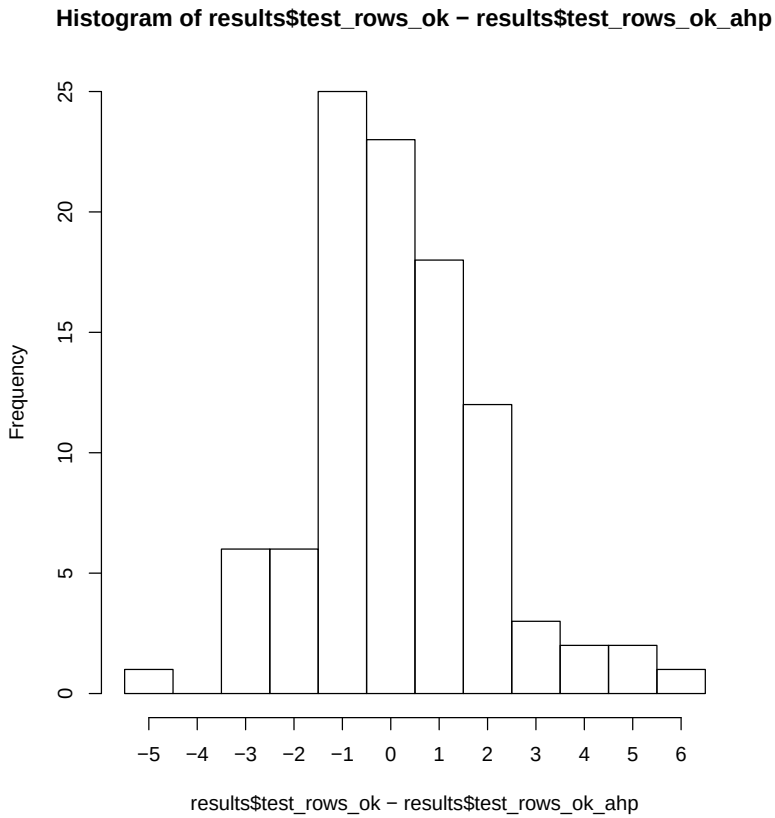


Figure 21: Histogram of estimated weights vs. AHP weights, scaled criteria

	# correct predictions	AHP weights							
		4	5	6	7	8	9	10	sum
Estimated weights	4	0	1	0	0	0	0	0	1
	5	0	0	0	0	2	0	1	3
	6	1	0	1	0	0	2	0	4
	7	0	1	3	3	9	5	2	23
	8	0	1	5	2	7	6	1	22
	9	1	1	2	2	9	7	9	31
	10	1	1	1	0	3	4	5	15
	sum	3	5	12	7	30	24	18	99

Table 26: Prediction power of estimated weights vs. AHP weights, scaled criteria

Figure 21 and Table 26 illustrate the distribution of $x_{i1} - x_{i3}$, where $x_{ik}, k = 1, 3$ is the number of correct predictions with $k = 1$ referring to estimated weights and $k = 3$ referring to the AHP weights. Again, I used a matched pairs t-test to compare the prediction power of estimated weights vs. AHP weights. I have the following hypotheses: $H_0 : \mu_1 - \mu_3 = 0$ $H_1 : \mu_1 - \mu_3 > 0$

Here, as before, μ_1 refers to the number of correct predictions in the population when using estimated weights, and μ_3 refers to the number of correct predictions when using equal weights. Again, as we have the same population, the degrees of freedom is 98. The value for the t-test is $t(98) = 0,8133$ with a p-value of $p = 0,209 > 0,1$. Hence, it can be concluded that the estimated weights and AHP weights are equally good. The mean of $\mu_1 - \mu_3 = 0,1515$.

The used scaling method does not influence the comparison of estimated weights and the lexicographic model. So, to reiterate the results from the original values section, estimated weights proved much better at prediction than the lexicographic model. The p-value for the t-test for the difference of means was $p = 2,27 * 10^-7 < 0,001$.

Then, I compare equal weights to AHP weights. Figure 22 and Table illustrates the distribution of $\mu_2 - \mu_3$. We have the hypotheses $H_0 : \mu_2 - \mu_3 = 0$ $H_1 : \mu_2 - \mu_3 < 0$ Once again, the matched pairs t-test is done with 98 degrees of freedom. The value of the test is $t(98) = -2,5422$ with a p-value of $p = 0,0063 < 0,1$. Hence, it can be concluded that AHP weights outperform equal weights. The mean of $\mu_2 - \mu_3 = -0,4242$.

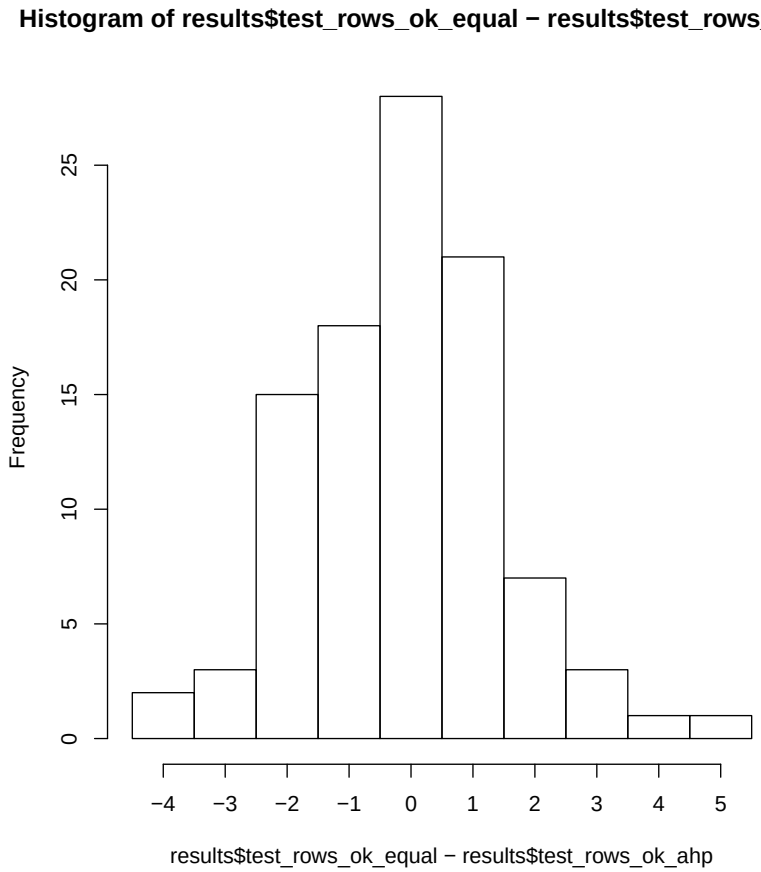


Figure 22: Histogram of equal weights vs. AHP weights, scaled criteria

AHP weights	# correct predictions	LEX model									
		2	3	4	5	6	7	8	9	10	sum
	4	0	0	0	0	0	1	0	2	0	3
	5	0	0	0	0	0	3	1	1	0	5
	6	1	2	0	0	2	2	3	0	2	12
	7	0	0	0	1	2	3	1	0	0	7
	8	0	0	1	5	3	9	5	6	1	30
	9	0	1	2	3	0	9	5	3	1	24
	10	0	0	0	2	3	1	3	6	3	18
	sum	1	3	3	11	10	28	18	18	7	99

Table 27: Prediction power of AHP weights vs. LEX model, scaled criteria

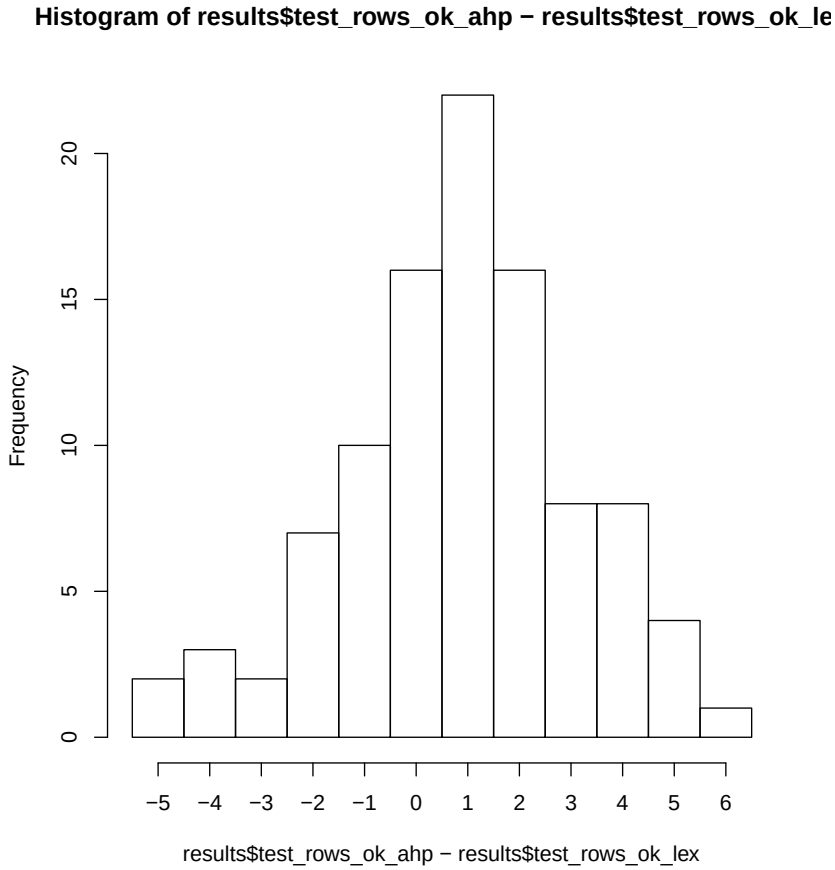


Figure 23: Histogram of AHP weights vs. LEX model, scaled criteria

As Figure 12 shows, for 16 subjects (16,1 %) the prediction power of the LEX model is equal to the power using AHP weights. For 24 subjects (24,2 %) the LEX model is better, and for 59 subjects (59,6 %) AHP weights are better.

I use a matched-sample t-test to investigate whether the prediction power is better when using estimated weights from the linear model than the equal weights

provided by the participants. We have the following hypotheses: $H_0 : \mu_3 - \mu_4 = 0$ $H_1 : \mu_3 - \mu_4 > 0$

Here μ_3 is the number of correct predictions in the population when using AHP weights, and μ_4 is the number of correct predictions using the LEX model. As we have 99 respondents with all the answers in the population, the degrees of freedom in the t-test is $99 - 1 = 98$. The value for the t-test is $t(98) = 3,72$ with a p-value of $p = 0,0001661 < 0,001$. Hence, it can be concluded that AHP weights outperform the lexicographic model clearly. The mean of $\mu_3 - \mu_4 = 0,848$.

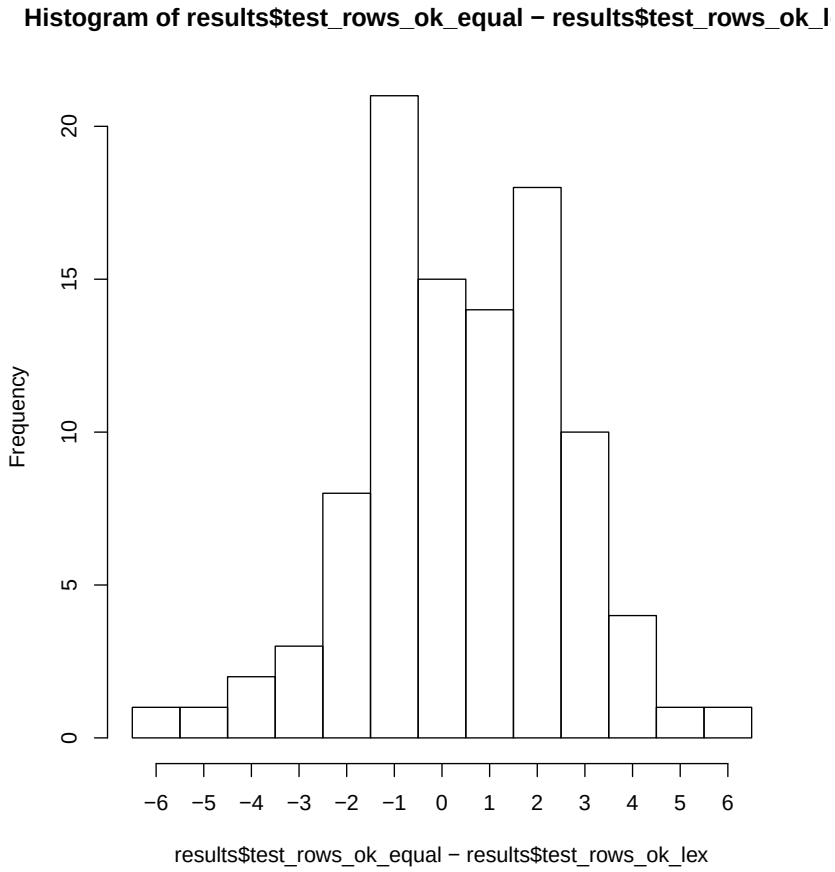


Figure 24: Histogram of equal weights vs. LEX model, scaled criteria

As the Figure 24 shows, for 15 subjects (out of 99, ie. 15,2 %) the predictive power of the LEX model was the same as the equal weights. For 48 subjects (48,5 %) equal weights proved better. For 36 subjects (36,4 %) the LEX model fared better.

I used a matched-sample t-test to investigate whether the prediction power is better when using equal weights from the linear model than the LEX model. I have the following hypotheses: $H_0 : \mu_2 - \mu_4 = 0$ $H_1 : \mu_2 - \mu_4 > 0$

Here μ_2 is the number of correct predictions in the population when using equal weights, and μ_4 is the number of correct predictions using LEX model. As we have 99 respondents with all the answers in the population, the degrees of freedom in

the t-test is $99 - 1 = 98$. The value for the t-test is $t(98) = 1,9565$ with a p-value of $p = 0,02663 < 0,05$. Hence, we conclude that equal weights outperform the lexicographic model by a slight margin. The mean of $\mu_2 - \mu_4 = 0,4242$.

Next, the results of the maximum likelihood calculations are reported in Table 28.

	Interval min	ML estimate	Interval max
Nonlinear subjects			
Estimated	0,645	0,778	0,871
Equal	0,579	0,716	0,822
AHP	0,630	0,764	0,860
LEX	0,524	0,662	0,777
Linear consistent subjects			
Estimated	0,733	0,857	0,929
Equal	0,672	0,804	0,891
AHP	0,714	0,841	0,918
LEX	0,639	0,773	0,868
All subjects			
Estimated	0,730	0,817	0,881
Equal	0,667	0,760	0,833
AHP	0,713	0,802	0,869
LEX	0,622	0,717	0,796

Table 28: 95% confidence intervals of the ML estimates of a successful prediction, Wilson estimation, scaled criteria

Let us display the same information graphically below in Figure 25.

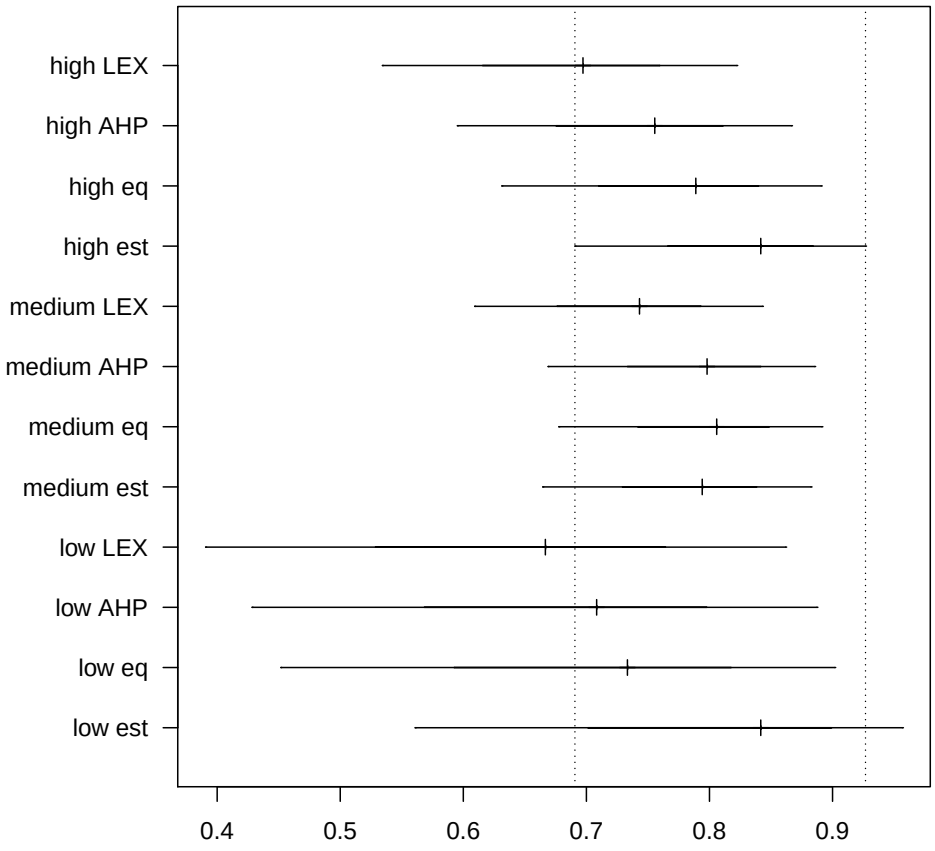


Figure 25: ML estimates with confidence regions by different methods and populations, Wilson CI estimation, scaled criteria

Figure 25 above shows the maximum likelihood estimates of the different prediction methods with their 95% confidence intervals, computed with the Wilson method. The vertical dotted lines represent the confidence interval minimum and maximum value for the maximum likelihood of a correct prediction when using the estimated weights from the linear model over all subjects. This could be named as the baseline, in that it uses our proposed method and the data from all of the subjects.

What can be seen in Figure 25 is essentially the previous pairwise method comparisons all together. Estimated weights outperform AHP weights only by a very narrow margin. Equal weights fare clearly worse than either of these two models. Finally, the lexicographic methods loses to all of the previous methods in terms of the likelihood of a successful prediction. It is important to note that the same order of methods pertains across all subject groups. That is, for both linear and nonlinear groups AHP and estimated weights are the best prediction methods. Even with nonlinear subjects equal weights cannot match the performance of the best methods.

4.2.2 CRT as predictor

One hypothesis considered whether the CRT score could be used to explain the performance of a prediction model. Another, but perhaps related question would be to ask, whether the CRT groups differ in terms of how well the subjects can predict their own choices. It could be, for example, that subjects with a higher CRT score would be better at making a reflective decision, and therefore be better at predicting their own choices.

This can be analyzed by looking at the Kendall rank-order correlation of AHP and linear weights across categories of subjects with a certain CRT score. That is displayed below in Figure 26.

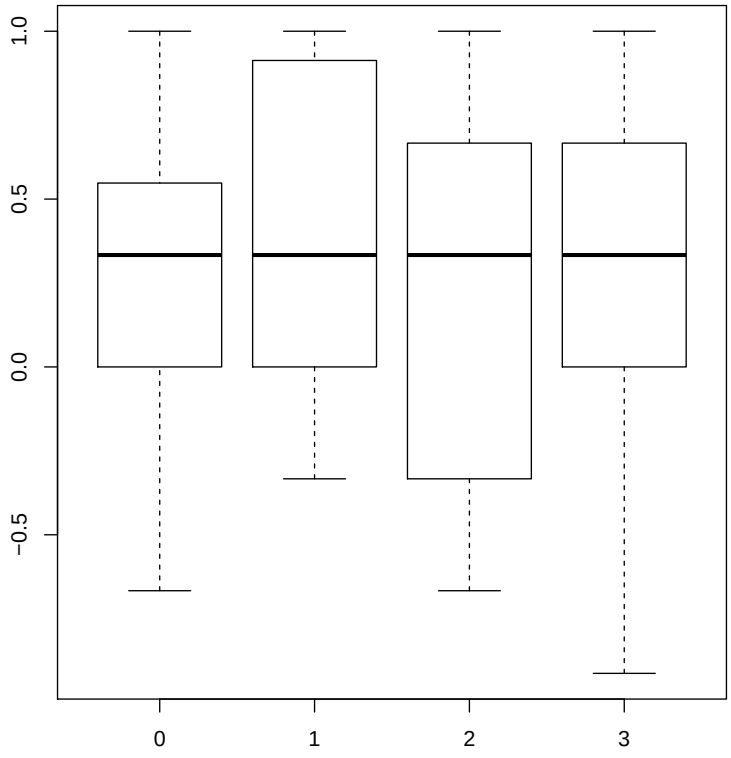


Figure 26: Rank-order correlation of AHP and linear weights by subject's CRT score, scaled criteria

As can be seen in Figure 26, the rank order correlation is independent of the CRT score of the subject. It seems that the cognitive style of the subject also does not hold any relation to how informative the AHP weights of the subject are. As a conjecture I entertained that the CRT would predict either linear consistency, or higher AHP-linear correlation, or both. The motivation for this claim was that it might be that more reflective, analytical subjects would be better at predicting their

own preferences. However, as Figure 26 shows this was clearly not the case. This suggests that, going back to the heuristic override process in Figure 1 by Stanovich and West (2008), the low correlation of the AHP and linear weights does not seem to reflect a failure to provide a System 2 response. Had there been a difference between the different CRT groups, it could have been hypothesized that the discrepancy in the weights arises due to an inability to override a heuristic response, for example. Now, as there are not differences between CRT groups, obviously such a hypothesis cannot be supported. Finally, in Figure 27 is displayed the ML estimation of the predictive accuracy for different methods.

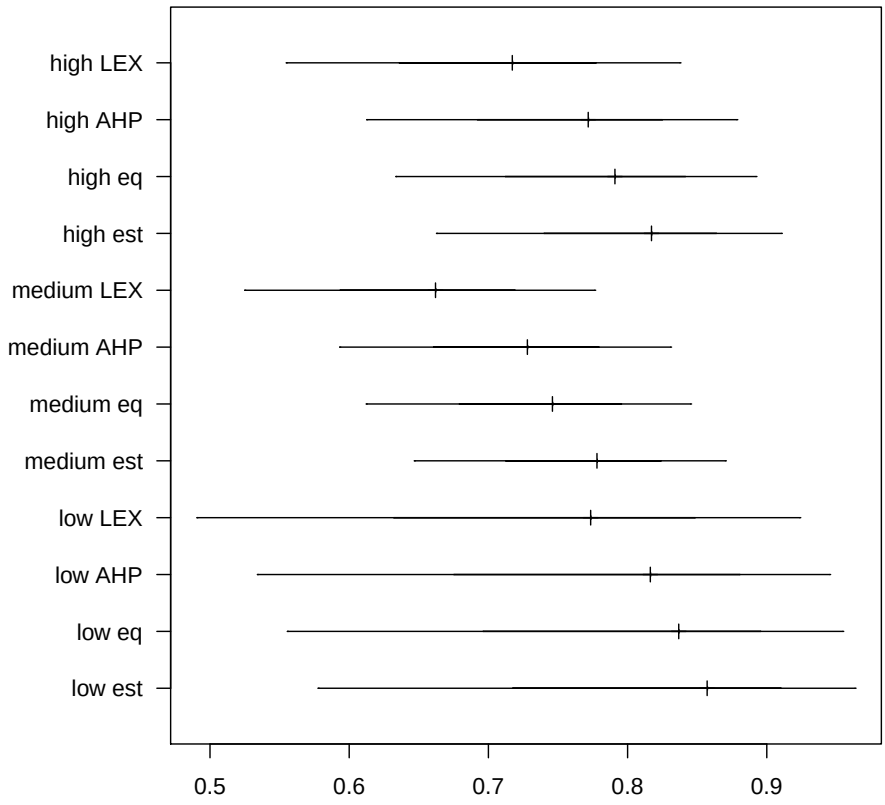


Figure 27: ML estimates with confidence regions by different methods and CRT groups, Wilson CI estimation, scaled criteria

As can be seen in the Figure 27, the maximum likelihood of different methods does differ very much between different CRT groups, but the order of different methods remains the same. The likelihood confidence range for the low group is very large, as the group consists of only 12 individuals. Due to the wide confidence intervals, it cannot be concluded from the data that there would be significant differences between the CRT groups.

5 Discussion

5.1 Consistency

The model seems to divide subjects into two categories: linear consistent and inconsistent ones. This is apparent in the result that allowing errors or removing constraints does not make a very large difference in the consistency ratio of subjects. If 10 % errors are allowed, that still only increases consistency from 52,8 % to 60,6 %, a result that pales in comparison to the result in Korhonen et al. (2012): a change from 38,9 % to 83,3 %. In this study, there is a substantial number of subjects that need to have five or more constraints removed to make them consistent with the linear model. In this study, we have 29 such subjects (22,8 %) whereas in Korhonen et al. (2012) the respective amount was just three subjects (2,1 %).

The reason for this result does not seem to lie in the lack of attention of the subjects. Only seven subjects (5,5 %) made any errors at all in the dominance questions, whereas in Korhonen et al. (2012) the number was 11 subjects (7,6 %). The same pattern repeats itself with the linear transformed questions: with transformations on the bound, here we had 29 subjects making errors (22,8 %) whereas in Korhonen et al. (2012) they had 113 subjects (78,5 %). With tranformations over the bound, the respective amounts were 29 (22,8 %) and 115 (79,9 %). With control questions we had 81 subjects (63,8 %) making at least one error, whereas in Korhonen et al. (2012) the number was 71 subjects (49,3 %).

It seems that in this study subjects were actually even slightly better at choosing consistently with respect to linear transformed questions. Though they were less apt to choose the same option in the control questions, this had almost no effect on the linear consistency of the subjects of our study (see Table 8). Therefore, it can be concluded that in this study, subjects stayed more consistent with respect to their previous choices, but were still less consistent with the linear model than in Korhonen et al. (2012). It could be that a four-criteria setting induces some subjects to use some other strategy, which is not so amenable to approximation by a linear model. Indeed, this was one of the motives for the hypothesis that the CRT score of the subject might be related to the predictive power. However, as we have seen, the CRT score was not related to linear consistency nor to the predictive power of the subject. Now, of course, this does not exclude the possibility that a change in strategy still might have occurred. But, as seen in the low power of the lexicographic strategy, that was at least not the alternative strategy used by the subjects.

5.2 Prediction

The main finding regarding prediction is that scaling the criteria makes a big difference to the success of different methods. The relative success of different methods has been summarized in Tables 29 and 30. As can be seen in the Tables, when using original values, estimated weights beat all of the other methods except equal weights. Equal weights, on the other hand, fared better than AHP and the lexicographic method. Finally, the AHP weights proved better than the LEX method.

It can also be inferred from the results that the equal weights would most likely outperform the LEX method, even though this was not tested directly. Due to the lack of direct testing, this relationship is marked in Table 29 in parentheses.

In contrast, when using criteria scaled to the range 0-1, estimated weights and AHP weights fared equally well. In third place was equal weights, and the LEX model came last.

Intriguingly, even though only 52,8 % of subjects were consistent with the linear model - and allowing 10 % errors raised this to 60,6 % - the model is very good at predicting the choices of subjects. An average prediction rate of 81,7 % is quite good. Even more surprisingly, the prediction rate is not influenced very much by the linear consistency of the subject. Even though the model seems to divide subjects into two classes: consistent and inconsistent, and there is very little chance due to errors, the model is almost equally successful at predicting the choices of both groups. When using scaled criteria, the AHP weights also fare well for both linear and nonlinear subjects.

	Estimated	Equal	AHP	LEX
Estimated	-	No difference	Estimated	Estimated
Equal		-	Equal	(Equal)
AHP			-	AHP
LEX				-

Table 29: Relative performance of different prediction methods, original criteria

	Estimated	Equal	AHP	LEX
Estimated	-	Estimated	No difference	Estimated
Equal		-	AHP	Equal
AHP			-	AHP
LEX				-

Table 30: Relative performance of different prediction methods, scaled criteria

The fact that equal weights performed comparably to the epsilon-formulation with original criteria does not corroborate the sentiment expressed by Dawes & Corrigan already in the mid-1970s. They noted : *"The whole trick is to decide what variables to look at and then know how to add."* (Dawes and Corrigan, 1974). Although the equal weights fare comparably to the epsilon-formulation with original weights, the equal weights fared clearly worse when scaled values were used. Scaling the values is normatively appealing, as otherwise a single criterion value can easily dominate in the utility calculation. In fact, that is the reason why equal weights fared so well with original valued criteria: the price is the biggest factor for the choice, and as it was originally approximately 10 times larger than the other criteria, using an equal weights scheme with original criteria meant that the prediction was mostly dependent on the price.

5.3 Importance judgments

As seen in Figure 7, the rank-order correlation regarding the AHP weights and the linear weights varied very much between subjects when using original values. Indeed, they varied from -1 to 1, which is the maximum variation for the Kendall correlation coefficient. Moreover, the rank-order correlation was independent of the linear consistency of the subject. This seriously questions the interpretation, on which the AHP weights would reflect the importance of different criteria. Although for some subjects the rank-order correlation was almost perfect, these were a very small minority of the subjects: 20 or 15,5 % with a correlation of 0.75 or more. On average, as the median rank-order correlation was zero, we cannot infer anything from the AHP weights to the linear weights created by the model. When using scaled criteria values, 29 subjects (22,8 %) achieved a AHP-linear weight correlation of over 0.75. However, this time the median correlation was 0.33. This shows that linear weights are not related to importance judgments in the absence of scaling. To interpret the connection of importance and weights, we need to know the scaling factor.

A similar pattern emerged when using scaled criteria. However, this time the average Kendall correlation coefficient was not zero, but instead approximately 0.25. Additionally, the Kendall coefficient was higher for linear consistent subjects than nonlinear subjects. It is to be noted that using a scaling factor for criteria improved the consistency of importance judgments seems to suggest that the scaling factor matters for importance statements. It could be, for example, that subjects are thinking of the importance on some scale that, being purely subjective, we cannot know. This could be examined by looking at how different scales influence the Kendall correlation coefficient. Perhaps the coefficient can be maximized with some scaling rule. However, at this stage this is just tentative speculation, and more research would be needed.

All in all, this study echoes the conclusion of Korhonen et al. (2013) that the importance judgments are not explained simply by the weights of the linear function. This holds true both in this four-criteria setting and their two-criteria experiment. As the correlation depends on the scaling factor, it cannot be said that importance is reflected only in the weights. Rather, to make sense of importance judgments, we need to know the weights and the scale that the subject is using.

5.4 Gender

The fact that the gender effect in Korhonen et al. (2012) was not replicated in this study makes it reasonable to ask for the explanation of these differences. It could be that the gender effect in their study was a statistical anomaly and happened by chance. However, the effect was deemed statistically significant. If it was not an anomaly, the lack of the effect in this study should be explained, as the same method was used here. There are, of course, several candidates for the explanation:

- Our population differed from the population in Korhonen et al (2012)
- The effect does not materialize in a four-criteria problem

- The effect does not materialize in this particular context (apartment choice)

Understandably, the results of this study do not provide enough information to settle this issue. All of the above are possible reasons for the lack of the effect, and none of them can be concluded directly from this study. Hence, more research is needed to find out the reason. However, this study provides at least evidence that the gender effect is not pervasive in MCDM problems, so at least a priori claims of gender effects ought to be treated somewhat sceptically. The fact that the gender effect disappeared in this study could also be because of the problem context. In this case, more research might show what is the defining factor in the context that made the effect disappear. On the other hand, it might also be that the gender effect disappears as the number of criteria increases. In this case, the lack of the gender effect in this study would not be surprising. However, to my knowledge there is no evidence supporting this kind of claim to date.

5.5 CRT effects

As seen in Figures 16 and 27, the differences of predictive efficacy of the different methods did not differ considerably between CRT groups. Therefore, it must be concluded that the CRT is not related to the predictive power of the models, either. It seems that the CRT does not explain the linear consistency or the predictive power of the subjects. This suggests that cognitive style is not the explanation why some subjects are inconsistent, or why some subjects' choices are hard to predict.

6 Summary

This study has shown us that the epsilon-formulation developed in Korhonen et al. (2012) is well suited for use in a four-criteria setting as well. In a four criteria setting, subjects make errors quite commonly, but these do not affect the linear consistency of the subjects markedly. In contrast, subjects seem to be divided into two categories: those who are linearly consistent already and those who cannot be made linearly consistent even by allowing a few errors.

However, despite the fact that not all of the subjects are linearly consistent, the model succeeds well in predicting choices. On average, the model predicts 85,7 % of the choices of linear subjects and 77,8 % of the choices of nonlinear subjects. Prediction success with comparison models was heavily influenced by whether original criteria or scaled criteria values were used. The scaled criteria situation is more typical and normatively more appealing. In both of these domains the predictive accuracy of the model is greater than that of three out of four comparison models. When using scaled criteria, a pairwise comparison of prediction methods across all subjects revealed that the epsilon formulation outperformed equal weights and the lexicographic model. It tied with AHP weights, with a nonstatistically significant difference in favour of the epsilon method.

It can be concluded from the results that the judgments of importance that subjects made in the beginning are quite uninformative if original criteria values are used. In fact, the informativeness of the importance judgments depends on the scaling factor of the criteria values. With 0-1 scaling, AHP weights perform as well as linear weights. However, knowing the AHP weights in the absence of a scaling factor does not enable any stable inferences regarding the choices or their consistency.

Finally, it was shown that the cognitive style, ie. the CRT score, of subjects is not related to the linear consistency or predictive accuracy of any of the methods. Even for nonlinear subjects, the epsilon method proved to be the best predicting method of those methods in our comparison set. Of course, this does not mean that there could not be a better prediction method for nonlinear subjects. All in all, the results from the prediction comparison highlight the fact that the scaling factor has a lot of influence for some methods, such as AHP and equal weights.

References

Bettman, J. R., Luce, M. F., and Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research*, 25(3):187–217.

Borcherding, K., Eppel, T., and Winterfeldt, D. v. (1991). Comparison of weighting judgements in multiattribute utility measurement. *Management Science*, 37(12):1603–1619.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–117.

Choo, E. U., Schoner, B., and Wedley, W. C. (1999). Interpretation of criteria weights in multicriteria decision making. *Computers & Industrial Engineering*, 37(3):527–541.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American psychologist*, 34(7):571–582.

Dawes, R. M. and Corrigan, B. (1974). Linear models in decision making. *Psychological bulletin*, 81(2):95–106.

Einhorn, H. J. and Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Journal of Accounting Research*, 19(1):1–31.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49(8):709–724.

Evans, J. S. B. T. and Over, D. E. (1996). *Rationality and Reasoning*. Psychology Press, Hove.

Fischer, G. W. (1995). Range sensitivity of attribute weights in multiattribute value models. *Organizational Behavior and Human Decision Processes*, 62(3):252–266.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19(4):25–42.

Goldstein, W. M. (1990). Judgments of relative importance in decision making: Global vs local interpretations of subjective weight. *Organizational Behavior and Human Decision Processes*, 47(2):313–336.

Hammond, K. R. (1996). *Human judgement and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. Oxford University Press, New York, NY, US.

Hastie, R. and Dawes, R. M. (2009). *Rational choice in an uncertain world: an introduction to judgement and decision making*. SAGE, London.

Hsee, C. K. (1998). Less is better: when low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, 11(2):107–121.

Järvenpää, S. L. (1990). Graphic displays in decision making - the visual salience effect. *Journal of Behavioral Decision Making*, 3(4):247–262.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9):697–720.

Kahneman, D. and Tversky, A. (1984). Choices, values, and frames. *American psychologist*, 39(4):341–350.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, Cambridge, UK.

Kendall, M. (1948). *Rank correlation methods*. Griffin, Oxford, England.

Korhonen, P., Moskowitz, H., and Wallenius, J. (1992). Multiple criteria decision support - a review. *European Journal of Operational Research*, 63(3):361–375.

Korhonen, P. J., Silvennoinen, K., Wallenius, J., and Öörni, A. (2012). Can a linear value function explain choices? an experimental study. *European Journal of Operational Research*, 219(2):360–367.

Korhonen, P. J., Silvennoinen, K., Wallenius, J., and Öörni, A. (2013). A careful look at the importance of criteria and weights. *Annals of Operations Research*, 211(1):1–14.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of measurement: Vol. 1: Additive and Polynomial Representations*. Academic Press, New York.

Oechssler, J., Roider, A., and Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, 72(1):147–152.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, 43(1):87–131.

Roy, B. and Mousseau, V. (1996). A theoretical framework for analysing the notion of relative importance of criteria. *Journal of Multi-Criteria Decision Analysis*, 5(2):145–159.

Saaty, T. L. (1980). *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1):83–98.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1):3–22.

Stanovich, K. E. and West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94(4):672–695.

Toplak, M. E., West, R. F., and Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7):1275–1289.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Tversky, A., Sattath, S., and Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review July 1988*, 95(3):371–384.

Tversky, A. and Simonson, I. (1993). Context-dependent preferences. *Management Science*, 39(10):1179–1189.

von Nitzsch, R. and Weber, M. (1993). The effect of attribute ranges on weights in multiattribute utility measurements. *Management Science*, 39(8):937–943.

Weber, E. U. and Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, 60(1):53–85.

Weber, M. and Borcherding, K. (1993). Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operational Research*, 67(1):1–12.

Welsh, M. B., Burns, N. R., and Delfabbro, P. H. (2013). The cognitive reflection test: how much more than numerical ability? *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1587–1592.

A Proofs of Lemmas 1 and 2

Lemma 1. *The consistency property of a linear value function is invariant under a linear transformation of the criteria.* $x_{ij} \to \alpha x_{ij} + \beta_j, i \in N, j = 1, 2 \ldots 4$, *and* $\alpha > 0$.

Proof: Korhonen et al. (2012), reproduced in Appendix 1.

$$\sum_{j=1}^{4} \lambda_j x_{rj} > \sum_{j=1}^{4} \lambda_j x_{sj} \iff \alpha \sum_{j=1}^{4} \lambda_j x_{rj} > \alpha \sum_{j=1}^{4} \lambda_j x_{sj} \iff$$

$$\sum_{j=1}^{4} \lambda_j \alpha x_{rj} + \sum_{j=1}^{4} \lambda_j \beta_j > \sum_{j=1}^{4} \lambda_j \alpha x_s rj + \sum_{j=1}^{4} \lambda_j \beta_j \iff$$

$$\sum_{j=1}^{4} \lambda_j \alpha x_{rj} + \beta_j > \sum_{j=1}^{4} \lambda_j \alpha x_{sj} + \beta_j \text{ for all} (X_r, X_s) \in P. \square$$

Lemma 2. *If a linear value function* $\sum_{j=1}^{4} \lambda_j x_{ij}$ *with vector* $\lambda \geq 0$, *is consistent with the DM's preferences, then the linear value function* $\sum_{j=1}^{4} \mu(\lambda_j \alpha x_{ij} + \beta_j), i \in N$ *and* $j = 1, \ldots 4$, *where* $a_j > 0, \mu_j = \frac{\lambda_j}{\alpha_j}$, *is consistent with all preferences* $(X_r, X_s) \in P$.

Proof: Korhonen et al. (2012), reproduced in Appendix 1.

$$\sum_{j=1}^{4} \mu(\lambda_j \alpha x_{rj} + \beta_j) > \sum_{j=1}^{4} \mu(\lambda_j \alpha x_{sj} + \beta_j) \iff$$

$$\sum_{j=1}^{4} \mu \lambda_j \alpha x_{rj} > \sum_{j=1}^{4} \mu \lambda_j \alpha x_{sj} \iff$$

$$\sum_{j=1}^{4} \lambda_j x_{rj} \geq \sum_{j=1}^{4} \lambda_j x_{rj},$$

as $\alpha_j > 0$. Define $\lambda_j = \mu_j \alpha_j \Rightarrow \mu_j = \frac{\lambda_j}{\alpha_j}$. Then $\sum_{j=1}^{4} \mu(\lambda_j \alpha x_{ij} + \beta_j)$ is consistent with all preferences $(X_r, X_s) \in P$. $\square$

B The questionnaire

This appendix presens screenshots from the original questionnaire used in the study.

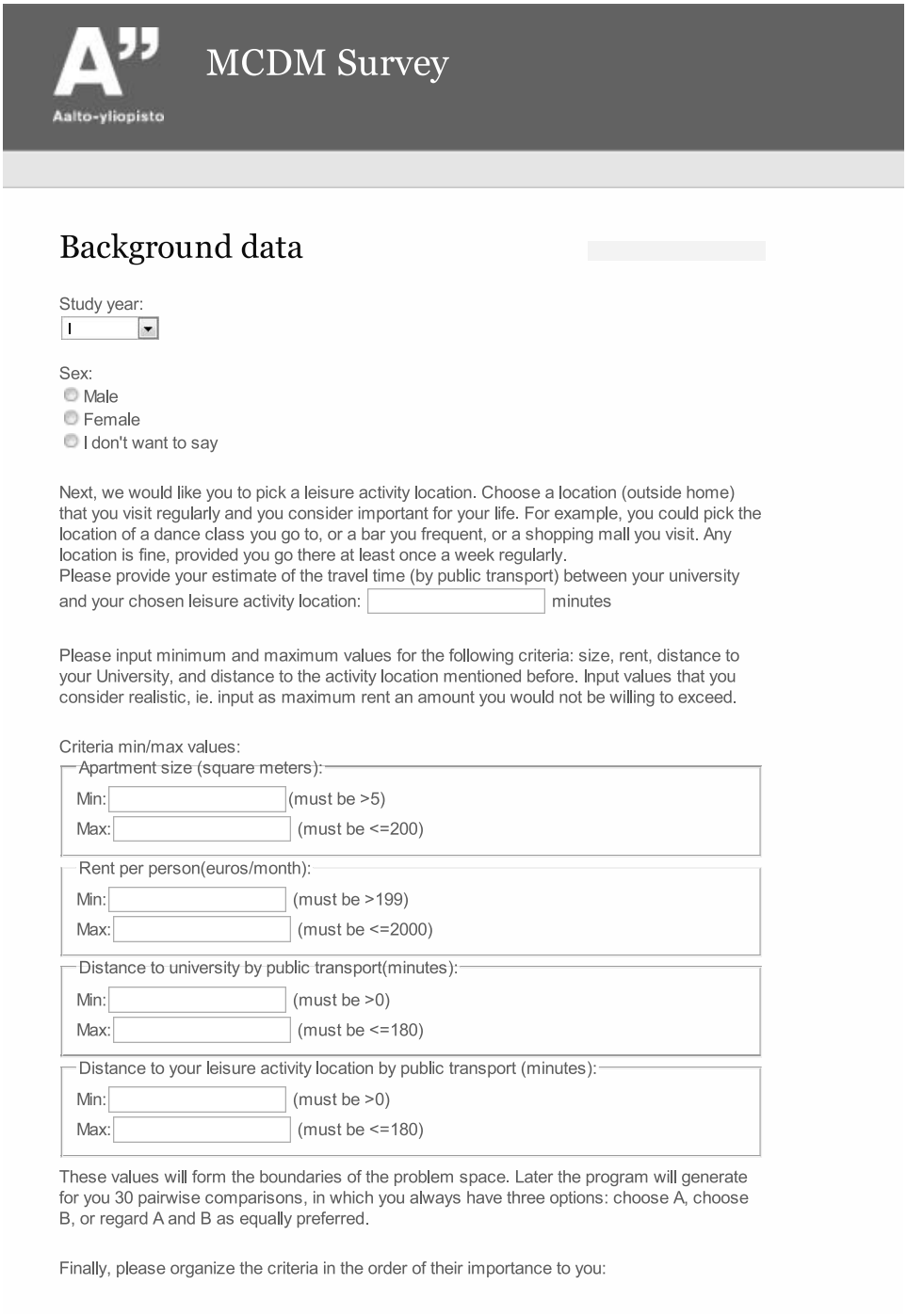


Figure B1: First page of the questionnaire

Figure B2: Second page of the questionnaire

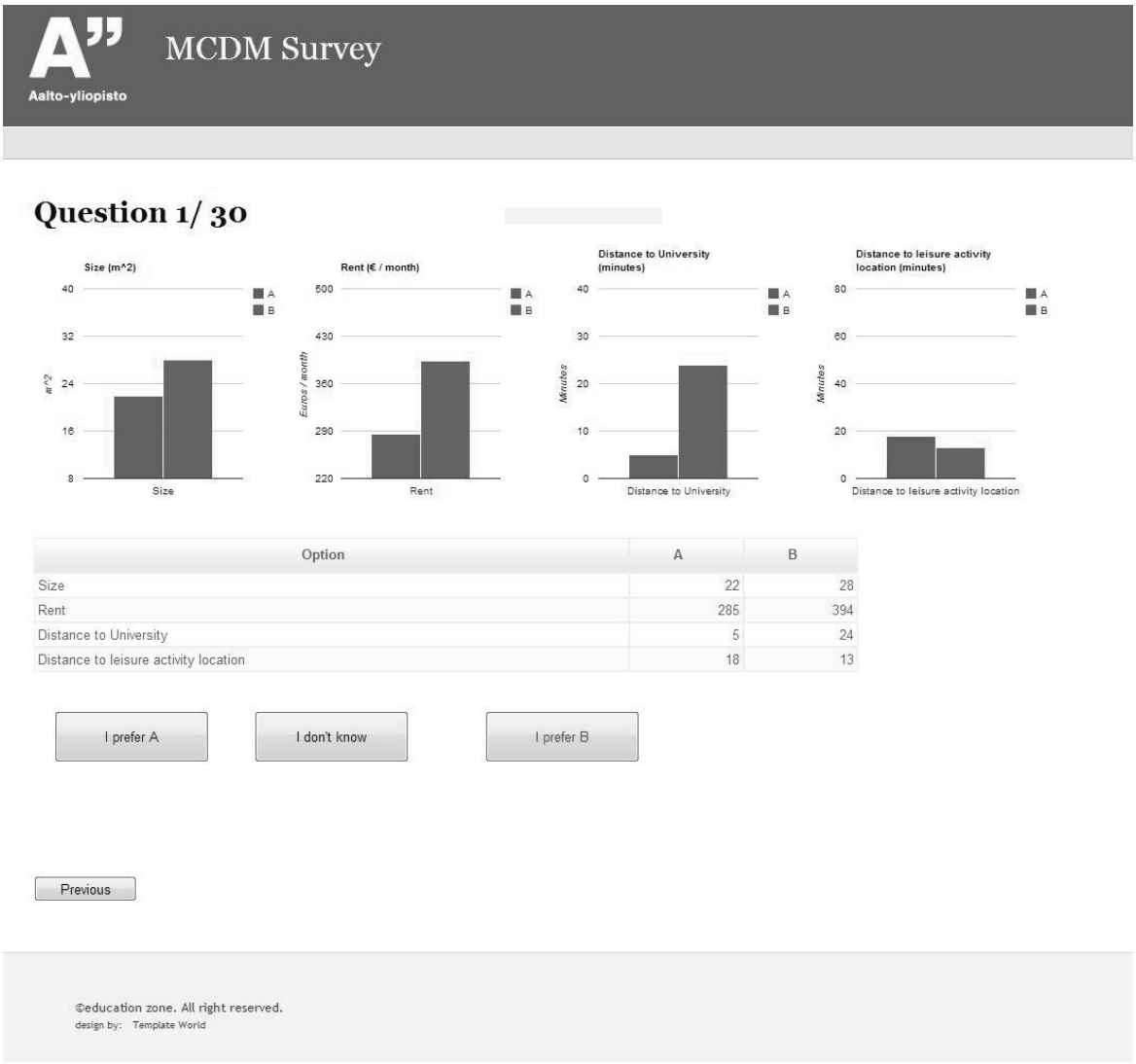


Figure B3: Third page of the questionnaire