Essays on the measurement of economic concepts in surveys

Inaugural-Dissertation

zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2012

vorgelegt von

Susanne Hoffmann

Erstgutachter:	Prof. Dr. Joachim Winter
Zweitgutachter:	Prof. Dr. Gebhard Flaig
Promotionsabschlussberatung:	15. Mai 2013

Datum der mündlichen Prüfung: 29. April 2013

Namen der Berichterstatter: Joachim Winter, Gebhard Flaig, Alexander Danzer *to Erwin Alois Hoffmann*

Acknowledgements

Writing this thesis has been an inspiring and challenging time. Its completion would not have been possible without the support of numerous people to whom I wish to express my gratitude at this point.

First and foremost I want to thank my thesis supervisor Joachim Winter. His enthusiasm about econometrics and his teaching were a great inspiration and motivated me to start this dissertation in the field of empirical economics. I benefited enormously from his comments and his discussions on the survey experiments that make up this thesis as well as on drafts of the resulting essays at all stages. I also want to thank him for always being welcome at his chair during the time of my doctorate. I am also deeply indebted to Gebhard Flaig for agreeing to serve as second supervisor on my thesis committee and Alexander Danzer for completing my committee as third examiner.

I am particularly grateful to my co-authors Frank Caro, Alison Gottlieb and Iris Kesternich for working as a team in the survey experiment that is the major part of this thesis. Their encouragement, patience and support are outstanding.

During the course of writing this dissertation, I benefited greatly from inspiring discussions with my colleagues and friends and from their support. I particularly would like to thank all current and former members of the Seminar für empirische Wirtschaftsforschung for their professional help and insightful feedback. Furthermore, I am indebted to my colleagues Tanja Greiner, Wolfgang Habla, Ines Pelger, Arno Schmöller, Michael Specht, Lisa Stadler, Till Stowasser and Elisabeth Wieland for their patience, encouragement and feedback in countless discussions. I am very grateful to also count them to my closest friends.

It has been an honor to be a teaching assistant at this faculty and I gratefully acknowledge the support of the department of economics in administrative as well as in personal matters, most notably from Silke Englmaier.

Most importantly, I am deeply indebted to my parents Ingrid Platt-Hoffmann and Roger Hoffmann for their support in my educational choices and my entire family for their love and understanding. I would like to thank my dear friends Barbara Freudenstein, Marlene Müller, Dorota Stolarek, Sidy Soumare and Julia Tygel for being a part of my life for such a long time and supporting me in countless ways. Thank you.

Susanne Hoffmann
Munich, May 2013

Contents

Chapter 1

Introduction

The field of empirical economics is concerned with the analysis of data to examine the pathways in which economies as a whole or single units of these economies, for example individuals, households or firms, behave. While this field has mainly used observational and administrative data in the beginning, a different source of data, namely survey data, has increasingly become an important pillar of many empirical analyses during the last century. Survey data is generated by asking randomly chosen units of the economy about their actions, feelings and attitudes. The methods to elicit these responses vary in many ways, such as for example the interviewing mode or the question types used. However, any study that is based upon survey data can only be useful, if the survey that generates this data is well designed and implemented. The current best practice in survey design and implementation has been a moving target and is at the core of the research agenda of survey methodologists. This research area draws upon and combines findings from various fields in the social science community, such as psychology, sociology, marketing and economics.

This dissertation consists of three empirical papers with a strong connection to the survey methodology field. At the core of each paper is the use of hypothetical persons, the so called vignettes. These vignettes can be used in two different contexts. First, vignettes are used in discrete choice experiments to elicit hypothetical choices from respondents. With these methods, survey participants are presented with hypothetical scenarios in which they have to make hypothetical choices. Discrete choice experiments have in recent years been used extensively in many fields (Louviere et al. (2000)). Its applications span a wide range from experiments in health care studies, health economics, environmental and transportation economics as well as in marketing and other fields. These approaches are useful, when actual choices or preferences are hard to observe or fail to be observed with the variation that is needed to correctly assess the topic at hand. The advantage of discrete choice experiments is that the design of the decision environment can be completely controlled by the researcher. With a careful design, it allows for the estimation of the effects of interest that would be analytically intractable otherwise (Lanscar & Louviere (2008)). Data resulting from DCEs belong into the category of stated preference data, where in contrast to revealed preference data no actual decisions or real transactions take place (Hensher et al. (2005)). The first two papers in this dissertation describe the results of a survey experiment that uses vignettes in this context.

The second context in which vignettes are applied is the anchoring vignette methodology. Here, vignettes are used to correct for biases that arise when groups of respondents differ in their response styles. It is well known in the survey methodology literature that people from different countries and even subgroups of people within the same country differ in their mapping from their actual situation to an answering point on a response scale. These systematic differences may bias the comparison of given responses when the differences in response styles are not taken into account (King et al. (2004)). One way to correct for these differences is the anchoring vignette method in which respondents are not only asked to assess their own situation, but also the situation of hypothetical persons, the so called anchoring vignettes. By relating the answers of the self-assessments to the answers provided for the anchoring vignettes, different response styles can be detected and corrected for. Thus, the vignettes are used to anchor the answers given by respondents to a fixed common value and to make the answers given by respondents comparable (King et al. (2004)). Since its original development, anchoring vignettes have found their way into many major surveys and panels, such as for example the Dutch CentERpanel and the Survey of Health, Aging and Retirement in Europe (SHARE). Its applications span a wide range of topics from health and life satisfaction to work disability and public institutions. Despite its growing use in empirical work, some methodological issues are still unsettled (see for example Soest et al. (2011), Datta Gupta et al. (2010), Bago d'Uva et al. (2011b), Vonkova & Hullegie (2011)).

The third paper in this dissertation uses vignettes in the anchoring vignette context and deals with an open methodological issue of this procedure.

Additionally, the first paper in this dissertation has ties to the strand of medical literature that examines patient decision making. Patient decision making nowadays is characterized by an increased role of the patient. The idea is that the patient should decide on the treatment option whereas the physician's task is to inform and consult (Braddock et al. (1999) and Gurmankin et al. (2002)). Information and consultation can however also be sought from other sources, such as decision aids or health related webpages. This invites questions about how efficient decision support strategies can be designed. The discrete choice experiment we conducted examines the influence of different sources of information on patient decision making and therefore adds to the strand of medical literature that examines how patients' decisions are influenced and can be supported by different sources of information.

In the following I will briefly describe each of the three papers. The first two papers are based upon a discrete choice experiment that was conducted with the RAND American Life Panel. This survey experiment was financed by the Foundation for Informed Medical Decision Making, Inc., by the FIMDM Investigator Initiated Grant 0212-1 and was conducted jointly with Frank Caro, Alison Gottlieb, Iris Kesternich and Joachim Winter. A total of 1616 members of the ALP that were 50 years or older at the time of the study participated in this experiment. This discrete choice experiment examines how respondents' choices for or against a medical procedure are influenced by different sources of information. Respondents were presented with three choice tasks in which they had to decide for three randomly chosen vignette persons whether or not to recommend full knee replacement surgery to treat these persons' knee osteoarthritis. These vignette persons varied in terms of their personal and health related situations as well as supplementary information features that may or may not be sought out by a patient when making a medical decision. The supplementary sources of information included in our experiment were a specialist's second opinions, person-specific surgery outcome forecasts and patient testimonials. All vignette persons were randomly generated with randomly varying levels of their personal and health related situations as well as of the supplementary sources of information. The surgery recommendations that are obtained from respondents in this way can then be used to establish the contribution of each of these levels on the decision making process. A major contribution of our experiment is the extensive use of video and audio material to present the vignette persons and the other information features. Since video vignettes have - to the best of our knowledge - never been applied before, a random subsample of respondents was assigned to a text-only version to enable the examination of respondents' survey behavior in the different media versions.

The first paper in this dissertation is joint work with Frank Caro, Alison Gottlieb, Iris Kesternich and Joachim Winter. The paper focuses on the analysis of how the personal and health related situations and especially the supplementary sources of information influence the decision making process in this discrete choice experiment. The central research question that we examine in this paper is to what extend the supplementary sources of information are sought out by respondents and to what degree they influence their decisions. Understanding how these sources of information are sought out and how they impact decisions is important for designing strategies to support informed patient decision making. This research adds to the literature that examines the effects of statistical information and patient testimonials by examining their impact in the presence of personal and health related information. We can thus establish the relative importance of these supplementary sources of information in an all-encompassing environment.

We find that specialist's second opinions, person-specific surgery outcome forecasts as well as patient testimonials with a consistent message are very influential in the decision making process. The impact of these supplementary sources of information sources is practically identical to the impacts of other factors, such as for example a person's pain level. While the finding of an independent influence of the outcome forecasts provides first evidence that these could become a future part of decision support strategies, the finding of a strong influence of patient testimonials in our experimental design is startling. The inclusion of patient testimonials in decision support strategies has been debated, especially in the light of possible biases that may arise due to the subjective nature of the testimonials. The fact that we observe influences of testimonials if two testimonials with a consistently positive or consistently negative message are presented - despite a large amount of person specific, professional and very objective information - is additional evidence that the use of testimonials in decision support strategies deserves caution. Further research needs to address the specific circumstances in which testimonials impact decisions and how biases from the use of testimonials can be avoided.

This paper is accompanied by a methodological appendix that describes various aspects of the survey design, implementation and respondents' survey behavior. The first part presents an example vignette person to illustrate the full design of the vignette persons and exemplify the language and content of the different sources of information. The second part of the appendix describes the randomization of respondents into the video and text versions of our survey and examines the randomization of levels of the information sources for the entire sample as well as for the subsamples that were assigned to the video and text versions. The final part of this appendix examines different aspects of respondents' survey behavior. First, it describes and examines the realized allocation of respondents into the text and video versions of our survey. Next, it analyzes respondents' survey behavior concerning the consideration of the information we presented them with and in making decisions. Finally, an attempt is made to assess the quality of the survey by examining respondents' assessment of the survey, the extent of item non-response, the extent of observed speeding behavior and the extent to which respondents made a good decision for vignette persons for the realized subsamples.

The second paper in this dissertation is also based upon the joint project with Frank Caro, Alison Gottlieb, Iris Kesternich and Joachim Winter and uses the data generated from this survey experiment. This paper exploits another unique design feature of our experiment. After the discrete choice experiment, we also administered a follow-up questionnaire that explicitly asked respondents about their perceptions of the influence and the helpfulness of the different sources of information we presented them with. This paper examines whether the influence of these sources of information in our discrete choice experiment and their contribution to respondents' confidence in decision making varies between subsamples of respondents according to their self-assessment of the influence and the helpfulness. The central research question of this paper is whether or not respondents are actually able to correctly assess the factors that influence their decisions most and that are most helpful in the decision making process. To answer this question, we split up the full sample into different subsamples depending on a respondent's assessment of what constitutes the most influential respectively the most helpful factor in the decision making process.

The estimation results of these subsamples were then compared to examine whether the subsample that considered a specific factor as most influential exhibits a stronger influence of this factor than the subsample that did not consider this specific factor as most influential and whether the subsample that considered a specific factor as most helpful exhibits a more positive/less negative contribution of this factor to the confidence in decision making than the subsample that did not consider this factor as most helpful. Furthermore, we examined whether within the subsample that considered a factor as most influential/most helpful this factor's influence on the decision/contribution to confidence in decision making is larger than for the other factors. If both of these examinations were affirmative for the considered subsample of respondents, this was interpreted as an indication that this subsample correctly assessed its most influential/most helpful factor.

The overall pattern that emerges is that respondents showed a large degree of insight in the factors that most influence their decision making and a lower degree of insight in which factors most help their decision making in our discrete choice experiment. In sum, we found that the self-assessment of the influence of all four assessed factors corresponds closely with the decision making behavior of these subsamples that we observed in our experiment. The self-assessment of the helpfulness of both assessed factors only partially corresponds with these factors' contribution to confidence in decisions. Therefore, respondents' perception of what is most influential and most helpful in their decision making may not be bias free, however respondents appear to have a notion of what influenced their decision making most in our discrete choice experiment.

The third paper is based upon a different survey experiment that was conducted at the Melessa laboratory in Munich and uses vignettes in an anchoring vignette framework to correct for biases in response styles. This paper is based upon a study by Hopkins & King (2010) in which the optimal question administration order of self-assessments and anchoring vignettes was examined. In a large scale online and telephone survey experiment, they applied the anchoring vignette method to the assessment of political efficacy and randomly varied the order in which respondents were presented with self-assessments and the anchoring vignettes. In the control group, respondents were presented with the standard ordering in which they first had to answer the self-assessments and then consider the situation of the vignette persons. The treatment group received the reversed order. Hopkins & King (2010) found that the bias correction with the anchoring vignettes works better when the standard of question administration order was reversed and propose to abandon the current standard. The paper presented here applies their reversal of question administration order to a different domain, the satisfaction with the current living situation and the value for money of the current apartment. The results are based upon an experiment in which 238 randomly chosen members of the Melessa subject pool participated in a paper and pen survey and were faced with the same experimental design as in Hopkins & King (2010).

In contrast to Hopkins & King (2010), we do not find beneficial effects of the reversed question administration order. The beneficial effects of the reversed question administration order in Hopkins & King (2010) are primarily attributed to what the authors call "intentional priming". However, as Tourangeau et al. (2000) note, the size and the presence of priming effects depend on a number of factors, such as the familiarity of the respondent with the issue at hand. It is therefore possible that the reversal of the question administration order may lead to different effects depending on the context to which the anchoring vignettes method is applied. In this application, the familiarity with the research topic was high, so that it can be assumed that priming effects were of lower size than in the Hopkins & King (2010) political efficacy setting.

The contribution of this dissertation to the survey methodology literature is therefore threefold. First, it established that video vignettes can be employed in discrete choice experiments. Second, it provided first evidence that respondents exhibit a surprisingly large degree of insight in their decision making process, at least in our survey experiment. Third, it provides evidence that speaks against a reversal of question administration order in the anchoring vignette method and thus calls for further research in this area. The contribution of this dissertation to the medical literature is also nontrivial. First, it established an independent influence of person-specific outcome forecasts and second, it provides further evidence that the use of patient testimonials in patient decision support strategies deserves further scholarly attention.

Chapter 2

Patient Decisions about Knee Replacement Surgery: Contributions of second opinions, outcome forecasts, and testimonials.[1]

2.1 Introduction

In making medical decisions, patients may draw on information from a number of sources. In part, they are likely to be guided by their own symptoms and life-style preferences. Further, they are likely to receive some guidance from the health care providers who diagnose the disease and offer treatment options. However, with growing consumer skepticism and dissatisfaction with medical care, patients increasingly seek more information beyond the first medical opinion for a variety of reasons. Some patients seek second opinions from independent specialists. Studies on motives for seeking second opinions report that many patients report needing reassurance and more certainty (including more information about the diagnosis, treatment options, and prognosis); others also report having doubts about the advice they received or dissatisfaction with their treatment from the first provider (Mellink et al. (2003), Moumjid et al. (2007), Van Dalen et al. (2001)).

In addition, patients now may have access to a professionally-developed decision guide that provides comprehensive information about the disease and treatment options for many

[1] This chapter is based upon joint work with Frank Caro, Alision Gottlieb, Iris Kesternich and Joachim Winter

chronic conditions. From a health policy perspective, decision guides are a particularly attractive option because their content can incorporate a high level of expertise and they can be made available to patients at a low unit cost. Decision guides are offered to patients through various media including print, video, and the internet (Schwitzer (2002)); the latter is an increasingly attractive option because of its potential for rapid, wide-spread dissemination at low cost. The internet is also attractive because of its potential for inclusion of interactive features that can be useful in engaging patient interest and accommodating patient preference for selectivity in use of information.

Efforts to support patients in their decision-making invite questions about how the supplementary information affects the ways in which patients actually make decisions. Unfortunately, understanding the ways in which patients draw on information to make decisions is not straightforward. When confronted by choice situations, people vary in the extent to which they are receptive to information that would inform their decisions; some tend to approach decision making deliberatively while others tend to rush to decisions (Chaiken et al. (1996), March & Heath (1994)). Moreover, a substantial body of behavioral science research has shown that when people are forced to make choices that involve consideration of information on multiple dimensions, the relative impact of each of the dimensions tends to be complex; further, people often have limited insight into the basis for their choices (Rossi & Anderson (1982), Louviere et al. (2000)).

2.2 Research aims and literature review

This research aims to disentangle the influence of different sources of information on patient decision making for or against full knee replacement surgery to treat knee osteoarthritis. Osteoarthritis of the knee is a widespread chronic condition among middle aged and older people for which various treatment options are available. Among the treatment options is full knee replacement. Although this surgery is often highly effective, the procedure is expensive and requires active patient cooperation during an extended and often painful rehabilitation process. Further, there are some risks of serious side effects. For many patients, nonsurgical strategies are often satisfactory in controlling pain and restoring adequate physical functioning. The decision for or against full knee replacement surgery is thus a good example of a discretionary medical treatment with a strong need for well-informed patient decisions. In seeking guidance for making this decision, patients are likely to draw on information that supplements the information provided by their own health care providers. Of special interest in this study is understanding how sources of supplementary information may influence the decision-making process. The supplementary sources of information considered in this study are specialist's second opinions, patient-specific outcome forecasts, and patient testimonials.

Previous research has addressed questions about the relative effectiveness of statistical outcome predictions and patient testimonials in influencing research subjects. In a recent review article, Winterbottom et al. (2008) examine both the impact that testimonials may have on decision making as well as the bias that testimonials may introduce that reduces the quality of patient decision making. Influences of narratives were found in about one third of the considered studies. Differences in the forms of narratives in these studies however made it difficult to draw conclusions about their impact on decision making. Greene et al. (2010) compared the efficacy of normative messages to other persuasive messages that used anecdotal or statistical risk evidence. Messages were compared regarding their impact on beliefs, attitudes, and normative perceptions regarding tanning bed use. Anecdotal messages were more effective than statistical messages. Braverman (2008) examined the relative effects of testimonials and simple informational messages, and found that mode of delivery affected the impacts of testimonials. When delivered by audio, testimonials were more effective than when delivered by text. Betsch et al. (2011) examined the relative effects of narratives and statistical information provided on an online bulletin board in a recent study of perceived vaccination effects. They found that narratives reporting adverse effects of vaccinations had a stronger effect than statistical information. Narratives were particularly effective when a higher proportion of the narratives reported adverse consequences and when their content was emotional.

Our research is primarily concerned with two questions:

1. When supplementary sources of information are considered, how do they influence decisions? and

2. To what extent are patients willing to consider supplemental sources of information?

These questions are addressed in a context in which we also consider how patients facing knee surgery decisions may be influenced by their level of pain, opportunity costs, and orthopedic surgeon recommendations.

The research was guided by the following hypotheses:

1. patients are more likely to elect surgery when they experience relatively high levels of pain and relatively low opportunity costs associated with surgery;
2. patients are more likely to elect surgery when their orthopedic surgeons make clear recommendations than when their orthopedic surgeons are neutral;
3. the following supplemental information makes patients more likely to elect surgery: [a] Positive recommendations from a specialist providing a second opinion, [b] forecasts of greater than average expectations for successful outcomes, and [c] positive testimonials.

While we hypothesized that patient symptoms and orthopedic surgeon recommendations would have relatively strong effects, we did not have hypotheses about the relative strength of the other experimental variables. We expected that respondents would be selective in considering optional information, but we did not have hypotheses about the extent to which they would consider information.

2.3 Methods, experimental intervention and hypotheses

This study is based upon a discrete choice experiment (DCE) that examines the influences of several factors on respondents' decision making regarding full knee replacement surgery to treat knee osteoarthritis. Discrete choice experiments have in recent years been used extensively in many fields (Louviere et al. (2000)). Its applications span a wide range from experiments in health care studies, health economics, environmental and transportation economics as well as in marketing and other fields. The main purpose of DCEs is to study behavior in markets that do not yet exist or are hard to observe, for example in the health care sector where transaction costs are not born by consumers but largely by health insurances (Lanscar & Louviere (2008)). DCEs are structured survey experiments in which respondents are presented with hypothetical scenarios and are asked to choose between several possible hypothetical alternatives. These alternatives can be situations, products or persons which are presented with varying levels of dimensions. By observing respondents' choices between different alternatives, trade-offs between different dimensions and their levels as well as their individual contribution to overall utility can be analyzed. For example, when a product is described with the two dimensions price and quality, different alternatives can be constructed using varying price-quality combinations. By asking a respondent to choose between different alternatives one can infer how much value is placed on additional quality.

The study employed a fractional factorial survey design (Rossi & Anderson (1982)), a form of discrete choice experiments. We generated a hypothetical scenario in which respondents were asked to make decisions for hypothetical persons, the so called vignettes. These vignettes were presented with varying levels of several dimensions. The fractional factorial survey method makes it possible to show how specific elements in a vignette structure contribute to decision making (Rossi & Anderson (1982)). With this survey design, respondents are randomly assigned to a small fraction of all possible vignettes, where the levels of the vignette dimensions are randomly assigned for each vignette. We employed an innovative form of a fractional factorial survey design developed by the investigators (Caro et al. (2012)). Designed to take advantage of the internet, the technique delivers information to respondents in large part through audio and video files. The technique helps to approximate real-life situations and enabled us to offer respondents with interactive options.[2]

The survey experiment proceeded as follows: Through a video-clip, respondents experienced a mini-lecture (about 2.5 minutes) by a physician on osteoarthritis of the knee, with surgery included among the treatment options. Content information on knee osteoarthritis was drawn largely from a booklet and DVD on knee osteoarthritis published by Health Dialog

[2] To control for possible mode effects, we also included a control treatment in which respondents experienced the entire experiment in a text-only version. In a different analysis we were not able to determine large differences in response behavior between the video and text-only online surveys.

(Health-Dialog (2007)). Respondents were then asked to review the situation of three randomly chosen vignette persons who were considering knee replacement surgery to treat knee osteoarthritis.

After each vignette person, respondents were asked to recommend whether or not the vignette person should have knee replacement surgery using the following question:

1. Do you recommend that Name have full knee replacement surgery now?
Yes
No

The three vignettes were created using vignette persons who were randomly selected from six possible vignette persons (three of the vignette persons were male and three were female). The scenarios were presented with randomly chosen levels of up to eight substantive dimensions. These dimensions and their levels were the following:

1. Pain (high and moderate)
2. Opportunity cost associated with surgery (high and low) linked to the employment status of a person
3. Employment status (employed or retired)
4. An orthopedic surgeon's recommendation (positive or neutral)
5. A second opinion by a specialist (none, strong recommendation, substantial reservations)
6. A patient specific outcome forecast delivered by an online tool (none, above average or below average chances for successful surgery)
7. First patient testimonial (none, positive, or negative)
8. Second patient testimonial (none, positive, or negative)

In this study, dimensions 1 through 4 constitute the basic effects of our design whose purpose is to provide a fully encompassing picture of the vignette person's situation to the study participant and to establish the internal validity of our design. These dimensions were automatically included in all vignettes.

We included pain because we expected that surgery would more often be recommended when vignette persons were experiencing more intense pain. We included opportunity costs because we expected that greater disruption of activities during recovery from surgery and rehabilitation would decrease receptivity to surgery. We anticipated that employed and retired persons would experience opportunity costs differently. For those who were employed, we focused on employment income that might be lost during recovery from surgery. For retired people, we described ways in which life styles would be affected by surgery and rehabilitation. If the level of opportunity costs was high, the vignette person would experience substantial financial hardship (when employed) or experience substantial life-style disruption (when retired) if surgery was chosen. We included surgeon recommendations because these often constitute the starting point for a medical decision and patients are heavily influenced by surgeon recommendations (Gurmankin et al. (2002)). We included two levels of surgeon recommendation: positive and neutral. In the neutral condition, surgery was described by the orthopedic surgeon as an acceptable option but the emphasis was on the need for the patient to make the decision.

Dimensions 5 through 8 are at the core of our research agenda. These substantial dimensions are supplementary sources of information that may or may not be sought out by a patient in a decision making process. Each of these dimensions was shown to a large random subsample of vignette observations (approximately 75%).

We included information from a consulting specialist because patients are frequently encouraged to seek a second opinion as another significant source of expert information before making decisions (Moumjid et al. (2007), Van Dalen et al. (2001)).

Reviewing the consulting specialist's second opinion was optional and was introduced using the following statement:

> [Name] has sought a second opinion from a consulting physician who has carefully reviewed [Name's] health history and the proposed knee replacement operation. Do you want to find out what the specialist has to say?

The consulting specialist either strongly recommended surgery or raised substantial reservations about surgery.

We included outcome forecasts because they are widely used in decision guides. We used outcome forecasts specific to the vignette person because patient-specific forecasts seem likely to have stronger effects than the general outcome forecasts now widely used in decision guides. A study published in 1992 concerned with the effects of a decision guide for patients with benign prostatic hyperplasia that used patient-specific outcome forecasts provides an early precedent for this approach (Kasper et al. (1992)). Large-scale outcome studies for common procedures like knee replacement surgery could provide the basis for outcome projections that could control for such variables as patient age and body mass index. In our study, these outcome forecasts were presented via a tool on the homepage of a hypothetical knee osteoarthritis patient aid group and were introduced to the respondents in the following way:

> [Name] also sought further information on the internet. The homepage of a nonprofit knee osteoarthritis patient aid group offers a tool that predicts likely surgery outcomes. This application uses [Name's] personal and health information and compares these with a large sample of full knee replacement surgery patients to predict [Name's] chances of a successful surgery outcome. The development of this tool was funded by the U.S. Department of Health and Human Services.

Respondents were then shown a screen that informed them that the vignette person's chance of a successful surgery outcome was either "above average" or "below average". The chances of success were described in one of three ways: only in verbal terms, in numeric terms including specific percentages, or in a graphic format that used a bar chart to represent the vignette person's chances of a successful surgery outcome relative to the average chances of a successful surgery outcome. For the third vignette person, reviewing this information was optional.

For the third vignette person, we also included patient testimonials as an additional source of information because there is evidence that personal anecdotes often have a strong impact on decision making, testimonials are often used in decision guides, and testimonials can possibly introduce a serious bias when they are not representative of the population of patients (Winterbottom et al. (2008)). One patient testimonial was automatically presented for a random subsample with the following introduction:

> [Name's] friend recommended [Name] view videos of real patients talking about their experiences with knee surgery. [Name] viewed the following testimonial on the internet.

After viewing the first testimonial, respondents were given the option to see a second testimonial. We included both positive and negative testimonials. Therefore, respondents could either see one testimonial that was either positive or negative or, if they chose to view a second testimonial, they could see two testimonials with either consistently positive or consistently negative messages or two testimonials with mixed (positive and negative) messages.

Before making their recommendations for each vignette person, respondents had an opportunity to view a summary of the information provided to them on all of the dimensions except for the testimonials.[3]

Table 1 summarizes the design, listing all dimensions, their levels, the percentages with which these levels should appear and additional design details. The full design is also illustrated for

[3] We provided the opportunity for review to minimize the potential impact of recency effects. Our use of a video format to provide information on vignette persons contributed to our decision to make a summary available since respondents could not readily go back to review video clips.

a sample vignette person in Appendix A including the actual wording used in the experiment to illustrate the language and clarify the content of the levels within the dimensions.

Dimension	Level	Level frequency (%)	Additional details
Pain	High	50	Mandatory
	Moderate	50	
Opportunity costs and employment status	Employed (high)	25	Mandatory
	Employed (low)	25	
	Retired (high)	25	
	Retired (low)	25	
Orthopedic surgeon's recommendation	Positive	25	Mandatory
	Neutral	75	
Specialist's second opinion	Strong recommendation	37.5	Optional
	Neutral recommendation	37.5	
	Not available	25	
Person-specific outcome forecasts	Above average	37.5	Mandatory at the first and second vignette, optional at the third.
	Below average	37.5	
	Not available	25	
First patient Testimonial	Positive 1	18.75	Displayed at the third vignette. Mandatory.
	Positive 2	18.75	
	Negative 1	18.75	
	Negative 2	18.75	
	Not available	25	
Second patient Testimonial	Positive 1	18.75	Displayed at the third vignette. Optional.
	Positive 2	18.75	
	Negative 1	18.75	
	Negative 2	18.75	
	Not available	25	

Table 2.1: Vignette dimensions, levels, frequency and additional details

The complexity of the design led to a large number of possible vignettes. For the most part, vignettes were created through the random selection of levels from each of the dimensions. To the extent to which that process was used, the total number of possible vignettes is the product of the number of levels within each of the dimensions. Considering only the first six substantive variables, the total number of possible vignettes was 2 * 4 * 2 * 3 * 3 = 144. With the addition of one or two testimonials in the final vignette, the total number of possible combinations increased by a multiple of 7 to 1008.[4]

[4] The multiple of 7 presumes that the two positive testimonials were equivalent and the two negative testimonials were equivalent; those who received either two positive or two negative testimonials received different statements by different actors.

2.4 Study sample

Study participants were members of the RAND American Life Panel (ALP) who were aged 50 and older at the time of the survey and had not participated in our pilot study. Members of the ALP are drawn from the general population and are surveyed periodically in an internet format. Respondents without internet are provided internet access by RAND. They receive modest financial compensation for participating in particular studies.[5] The ALP invited all 2296 of its members who fit our sample selection criteria. A total of 1675 interviews were started, 1622 were completed. This results in a response rate of 70.6%. Background characteristics were incomplete for six respondents, who were subsequently dropped from the analysis. This resulted in a total respondent sample size of 1616.

We had access to standard demographic data about respondents obtained previously based on their participation in the panel. Additionally, we asked respondents to answer questions concerning their own medical histories with respect to chronic knee pain and knee osteoarthritis as well as their friends' and close relatives' experiences with knee osteoarthritis and full knee replacement surgery. Background characteristics of the final sample are summarized in Table 2. Respondents were fairly evenly represented by men and women and reflected a broad age range: between age 50 to 93. Respondents were well educated and relatively well-off financially.[6]

Study participants reported considerable experience with chronic knee pain and knee osteoarthritis. More than 40% of respondents reported experiencing chronic knee pain[7] and roughly half of these have been diagnosed with osteoarthritis in at least one knee. Two

[5] Further information on the American Life Panel, its composition and attrition can be obtained at the homepage of the RAND American Life Panel: https://mmicdata.rand.org/alp/index.php?page=panel (Rand (2012)).

[6] The ALP seeks to be representative of the adult population of the United States. We have to acknowledge that in comparison to CPS 2011 data for the population 60 and over, our specific sample is more highly educated and underrepresents racial and ethnic minority populations (US-Census (2012)). However, the ALP sample we use is much more generalizable than studies based on student or patient samples. Studies by Chang & Krosnick (2009) and Yeager et al. (2011) examined data quality issues with the ALP and another probability sample in comparison to samples obtained via RDD and non-probability samples. Both conclude that the phone sample and the probability sample show the least bias.

[7] In the relevant age range, this is a reasonable percentage of people with chronic knee pain and similar to a recent data from the 2011 Gallup-Healthways Well-Being Index (Gallup (2012)).

Variables	Percent
Gender	
Male	43.4
Age (median=59 years)	
50-59	50.1
60-69	33.7
70 or older	16.3
Income	
below 25,000$	23.4
$\geq 25,000\$, < 50,000\$$	27.3
$\geq 50,000\$, < 75,000\$$	16.1
above 75,000$	33.2
Living Status	
Married or living with a partner	58.2
Employment Status	
Retired	34.3
Working	45.2
Unemployed, disabled and other	20.4
Education	
High school or less	23.9
At most Bachelor's Degree	59.2
Postgraduate	16.9
Ethnicity	
Non-Hispanic white	82.7
Respondents with chronic knee pain	42.1
Respondents with knee osteoarthritis	21.2
Friends/Relatives with knee osteoarthritis	65.8
Friends/Relatives with full knee replacement surgery	52.5

Table 2.2: Descriptive statistics of the study sample

thirds of the sample reported having close relatives or friends who were diagnosed with knee osteoarthritis and about 50% reported having close relatives or friends who have had full knee replacement surgery.

2.5 Empirical analysis

The unit of analysis in this study is the single vignette observation. Since each respondent was presented with three vignette persons, we obtain up to three vignette observations per respondent. Our final sample used for analysis of the surgery recommendations consists of the three vignette observations obtained from each of the 1616 survey respondents with complete information. With three instances of item-nonresponse, our vignette sample consists of a total of 4845 vignette observations.

For each vignette observation, we identified whether or not the respondent recommended surgery for this specific vignette person. Our dependent variable ”surgery recommendation” thus takes the value 1 if surgery was recommended and 0 if surgery was not recommended.

2.5.1 Univariate analysis

Overall, for 52% of the vignette observations, participants recommended surgery for the vignette person. Column 2 of Table 3 presents the recommendation rates (mean of the variable ”surgery recommendation”*100) for the full sample and dimension subsamples. The third and fourth columns present the associated number of vignette observations and the percentage of the relevant total for each subsample. Panel A of table 3 focuses on dimensions available for all three vignette observations, panel B focuses on the patient testimonial dimension available only for the third vignette observation.

In column 2 of panel A, it can easily been seen that the recommendation rates vary with the dimension levels as hypothesized. Surgery recommendation rates for vignettes with high pain were almost 20 percentage points higher than for vignettes with low pain ($P<0.01$).[8] There were similar differences in surgery recommendation rates based on surgeon's recommendations ($P<0.01$). Differences in surgery recommendation rates based on opportunity costs and employment status were less pronounced but as hypothesized and still significant ($P<0.01$).

Differences in recommendation rates for second opinions and the outcome forecasts are as strong as for the basic dimensions. For subsamples where a ”strong recommendation” second opinion or an above average outcome forecast was shown, recommendation rates are at least 25 percentage points higher than for subsamples where a ”substantial reservations” second opinion or a below average outcome forecast was shown ($P<0.01$).

[8] All reported p-values in this section result from t-tests for differences in the means of the variable ”surgery recommendation” between the considered subsamples.

Furthermore, the use of these two dimensions when they were optional was very high. In 3,650 vignette observations the option to view a second opinion was provided, and in 89% of these observations respondents chose to view the second opinion. For the third vignette, a patient-specific outcome forecast was offered for 1187 vignette observations, and in 90% of these cases respondents chose to view the outcome forecast.

	Rec. rate (%)	Observations	% of total
Panel A			
full sample	51.7	4845	100.0
Sample split by:			
Pain level			
high pain	61.3	2491	51.4
low pain	41.5	2354	48.6
Opportunity cost level			
high opportunity costs	47	2441	50.4
low opportunity costs	56.5	2404	49.6
Employment status			
retired	54.4	2477	51.1
employed	48.9	2368	48.9
Surgeon's recommendation level			
positive surgeon's recommendation	66.2	1190	24.6
neutral surgeon's recommendation	47	3655	75.4
Second opinion level seen			
strong recommendation	66.7	1622	33.5
substantial reservation	32.4	1634	33.7
second opinion not chosen	59.6	394	10.8
second opinion not available	57.5	1195	24.7
Person-specific outcome forecast level seen			
forecast above average	64.9	1749	36.1
forecast below average	39.6	1788	36.9
numeric forecast format	54.1	1147	23.7
graphic forecast format	53.9	1231	25.4
verbal forecast format	48.2	1159	23.9
forecast not available	50.4	1188	24.5
forecast not chosen (3rd vignette)	53.3	120	10.1
Panel B			
full sample	54.5	1614	100.0
Sample split by:			
Patient testimonial combination seen			
one positive testimonial	62.1	174	10.8
one negative testimonial	47.5	137	8.5
two positive testimonials	66.4	140	8.7
two negative testimonials	41.4	157	9.7
mixed message testimonials	55.2	598	37.1
testimonials not available	53.7	408	25.3
second testimonial not chosen	55.6	311	25.8

Table 2.3: Distribution of surgery recommendation rates and dimension level frequencies

Panel B presents the distribution of surgery recommendation rates for only the third vignette observations depending on which testimonial combination was seen and whether the testimonials were available. Here, the mean recommendation rate (based on 1614 vignette observations) is similar to that of the full sample. The subsamples where one or two positive testimonials were seen have a higher recommendation rate than the subsamples where one or two negative testimonials were seen ($P<0.01$).The subsample where mixed (positive and negative) testimonials were seen does not differ substantially from the subsample where no testimonial was offered ($P=0.64$). Again, the use of the optional second testimonial is high: A patient testimonial was shown in 1206 cases and 74% of respondents opted to view a second patient testimonial.

The last two columns of panel A show that the actual frequencies of the dimension levels and availability of supplementary information features are approximately as designed.

2.5.2 Multivariate analysis

For our multivariate analysis of the influence of different sources of information on decision making we used linear probability models. Although there are specific regression models designed for binary dependent variables such as ours, for our multivariate analysis, we chose to report results using Ordinary Least Squares models for their ease of interpretation.[9] Thus, all regression models presented in the following section are linear probability models with the dependent variable ”surgery recommendation”. All estimated coefficients describe ceteris paribus percentage point changes in the probability of surgery recommendation. We conducted two independent analyses. The first analysis uses observations from the first and second vignette to examine the effects of the basic dimensions and two supplementary information features available for these vignettes, the second opinions and person-specific outcome forecasts. The second analysis uses observations obtained from the third vignette where patient testimonials were added and person-specific outcome forecasts were optional. The second analysis examines the relative impacts of all three supplementary information features and examines whether the addition of patient testimonials had an impact on the

[9] Our main conclusions are robust to using specific binary dependent variable models; results can be obtained from the authors upon request.

effects of the other dimensions. Estimation results for both analyses are presented in Table 3.

For both analyses, all estimated specifications control for random assignment to the text-only version (21% of the final sample), vignette order effects, and six vignette-specific constants[10] to account for the six different male and female vignette persons. We also control for the following respondent characteristics: gender (male or female), age (continuous), marital status (married or not), education (categories: at most bachelor and postgraduate; reference category: high school or less), respondent's labor force status (employed and retired; reference category: all forms of unemployment), ethnicity (non-Hispanic white, or other ethnicity), yearly household income (middle [$25,000 to $49,999], high [$50,000 to $74,999] and very high [$75,000 or more]; reference category: low [below $25,000],). Additionally, we control for the following respondent health characteristics: whether the respondent has chronic knee pain, whether the respondent was diagnosed with knee osteoarthritis, whether the respondent has friends/relatives who have been diagnosed with knee osteoarthritis, and whether the respondent has friends/relatives who have had full knee replacement surgery. Finally, all models use robust standard errors to account for the fact that each respondent was observed multiple times.[11]

The dimensions of our design are included in the regression models via effects coding. This means that for each of the dimensions, we create n-1 dummy variables for the n levels of this dimension. The nth level constitutes the reference category. For example, the dimension pain has two levels, high and low pain. Thus, we created one dummy variable "high pain" that takes the value 1 if a vignette observation is presented with a high level of pain, and "low pain" is the reference category. The estimated effect for this variable "high pain" is the percentage point difference in the probability of recommending surgery for vignette persons with high and low levels of pain. The same logic applies to the other basic dimensions and supplementary information features. If the supplementary feature was optional, we also include a dummy variable that captures when the feature was offered but not chosen. The

[10] The estimated constants describe the ceteris paribus baseline probability of surgery recommendation for a specific vignette person.

[11] The same analysis was also conducted with standard errors clustered at the individual level. There was no substantial difference in the estimated standard errors.

reference category for all three supplementary information features is always observations where the respective feature was not offered, and the estimated coefficients capture the percentage point difference between the respective level of the supplementary information feature and observations where this feature was not available.

Basic dimensions, second opinions and person-specific outcome forecasts

The first model uses the observations obtained from the first and second vignette observations and thus has a sample size of 3231 observations. The specification of our first linear probability model is as follows:

Pr(Y=1)=$\alpha_i + \beta(VO_i) + \gamma(Text_i) + \delta(FE_i) + \epsilon(X_i) + \theta(Basic_i) + \varphi(SO_i) + \chi(OF_i)$

where α_i stands for the vignette person fixed effects, VO$_i$ includes vignette order effects, Text$_i$ controls for whether the respondent was in the text treatment, FE$_i$ controls for outcome forecast format effects and X$_i$ includes respondent characteristics. The parameter vector θ contains the effect coded levels of our basic dimensions (”high pain”, ”high opportunity costs”, ”retired”, and ”positive surgeon’s recommendation”). The parameter vector φ contains the effect coded levels of the second opinions (”strong recommendation”, ”substantial reservation” and ”second opinion not chosen”, reference category: observations where no second opinion was offered), the parameter vector χ contains the effect coded levels of the outcome forecasts (”forecast above average” and ”forecast below average”, reference category: observations where no outcome forecast was displayed). The vignette person fixed effects, vignette order effects, outcome forecast format effects, and the text control are not reported in Table 4 for the sake of brevity. Furthermore, we will only report the significant background characteristics.

The regression results of Model 1 in Table 4 show that the effects of our basic dimensions are as expected. If a vignette person was characterized by a high level of pain, the probability of surgery recommendation is 20.0 percentage points higher than for a vignette person with low pain. Similarly, a positive surgeon’s recommendation increases the likelihood of a surgery recommendation by 20.7 percentage points compared with observations where a neutral surgeon's recommendation was shown. Vignette persons with high opportunity costs are 10.4 percentage points less likely to receive a surgery recommendation than vignette persons with low opportunity costs, and retired vignette persons are somewhat more likely to be recommended for surgery than employed vignette persons. All of these estimated differences are significant at the one percent level and confirm the hypotheses outlined earlier. Among these basic dimensions, the effects of high pain and a positive surgeon's recommendation are significantly stronger than the other effects ($P<0.01$).[12] Given that our hypotheses are confirmed, we are confident that our method succeeded in providing a valid discrete choice experiment.

The effects of second opinion levels and the outcome forecast levels produced equally strong results consistent with our hypotheses. Both levels of each of these dimensions significantly changed the probability of surgery recommendation in comparison with observations where these dimensions were not presented. If a vignette person was presented with a second opinion that strongly recommended surgery, this vignette person was 12.1 percentage points more likely to be recommended for surgery than vignette persons where no second opinion was offered. When a vignette person was presented with a second opinion that raised substantial reservations, the probability of surgery recommendation was 24.2 percentage points lower than for vignette observations where no second opinion was offered.

For the outcome forecasts, we observe that vignette persons presented with an above average outcome forecast were 12.2 percentage points more likely to be recommended for surgery and vignette persons presented with a below average outcome forecast were 11.6 percentage points less likely to be recommended than vignette persons where no outcome forecast was displayed.

[12] All p-values in the multivariate analysis refer to tests for equality of the strength of the estimated influences using either F-tests or Wald-tests.

Dependent variable	Surgery recommendation			
Model	(1)		(2)	
Basic dimension effects				
high pain	0.200	**	0.199	**
	(0.015)		(0.022)	
high opportunity costs	-0.104	**	-0.094	**
	(0.015)		(0.022)	
retired	0.053	**	0.053	*
	(0.016)		(0.022)	
positive surgeon's recommendation	0.207	**	0.177	**
	(0.017)		(0.025)	
Supplementary information effects				
Second opinion effects				
strong recommendation	0.121	**	0.111	**
	(0.021)		(0.030)	
substantial reservation	-0.242	**	-0.211	**
	(0.021)		(0.029)	
second opinion not chosen	0.050		0.054	
	(0.034)		(0.045)	
Person-specific outcome forecast effects				
forecast above average	0.122	**	0.103	**
	(0.024)		(0.036)	
forecast below average	-0.116	**	-0.168	**
	(0.024)		(0.035)	
forecast not chosen			-0.026	
			(0.050)	
Patient testimonial effects				
two positive testimonials			0.141	**
			(0.044)	
two negative testimonials			-0.129	**
			(0.041)	
mixed message testimonials			0.028	
			(0.028)	
one positive testimonial			0.069	
			(0.038)	
one negative testimonial			-0.062	
			(0.045)	
Respondent's background characteristics				
Postgraduate	-0.059	*	-0.087	*
	(0.026)		(0.038)	
Non-Hispanic white	-0.020		-0.080	**
	(0.022)		(0.031)	
Respondent has chronic knee pain	-0.051	**	-0.019	
	(0.020)		(0.028)	
Respondent's close friends/relatives have had full knee replacement surgery	0.092	**	0.059	
	(0.023)		(0.035)	
N	3231		1614	
Adj. R2	0.621		0.654	

All models control for vignette fixed effects, vignette order effects, outcome forecast format effects, text treatment and respondent characteristics. The stars behind the coefficients indicate significance levels. ** Significant at the 1 percent level, * at the 5 percent level. Standard errors are displayed in brackets below the coefficients.

Table 2.4: Predictors of respondent's decisions regarding full knee replacement surgery

Comparing the strength of the effects of the second opinion and outcome forecast levels, we find that the ”substantial reservations” second opinion is significantly stronger than either an outcome forecast or a second opinion recommending surgery(P<0.01). Compared with the levels of the basic dimensions, we cannot distinguish the strength of the effects of a high pain level and the positive surgeon’s recommendation from the effect of the ”substantial reservations” second opinion. However, these three dimension levels have significantly stronger effects on surgery than a retired vignette person, high opportunity costs and the remaining levels of the outcome forecasts and second opinions (P<0.01). The strength of the effects of both outcome forecast levels, ”strong recommendation” second opinion and high opportunity costs are not significantly different from each other, but all three are stronger than the effect of a retired vignette person (P at least <0.05). Therefore, among all effects the positive effect of a retired vignette person is the weakest in magnitude.

The relative effects of patient testimonials, second opinions and outcome forecasts

This analysis uses only the third vignette observation for which person-specific outcome forecasts were optional and for which a random subset of respondents was shown one testimonial and offered a choice to view a second testimonial. This leaves us with a sample of 1614 observations which we use to estimate Model 2 presented in Table 4. This model expands Model 1 by adding the effects of viewing a specific testimonial version (”one positive testimonial”, ”one negative testimonial”, ”mixed message testimonials”, ”two positive testimonials” or ”two negative testimonials”; reference category: observations where testimonials were not shown) and by controlling for the fact that reviewing the information contained in the outcome forecasts was now optional (”forecast not chosen”).

When we compare the two models, the estimated coefficients on the basic dimensions, second opinions, and outcome forecasts are similar and none of the observed differences between the two models is statistically significant. The estimated effects of high pain, a positive surgeon’s recommendation, and a ”substantial reservations” second opinion in Model 2 are still of equal magnitude. The effect of a ”substantial reservations” second opinion continues to be stronger than the effect of an above average outcome forecast ($P<0.05$), but can no longer be distinguished from the effect of a below average outcome forecast or the second opinion recommending surgery. The effect of high pain continues to be stronger than the effects of the second opinion recommending surgery and the above average outcome forecast ($P<0.05$), but is no longer significantly stronger than the effect of a below average outcome forecast. Furthermore, the effect of a positive surgeon's recommendation no longer statistically dominates the effects of a second opinion recommending surgery, an above average outcome forecast and a below average outcome forecast. However, the overall pattern of dimension effects is similar. High pain, positive surgeon's recommendation and ”substantial reservations” second opinion are the strongest impact factors for both models.

When we focus on the estimation results of Model 2, it can be seen that the viewing of a single testimonial, whether positive or negative, had no effect on recommendations. However, viewing two positive testimonials significantly increases the probability of surgery recommendation by 14.1 percentage points, and viewing two negative testimonials significantly reduces this probability by 12.9 percentage points. The magnitude of these effects is not statistically different ($P=0.86$). The effect of viewing two positive testimonials is significantly stronger than of viewing a mixed testimonial version ($P<0.01$).

2.6 Conclusion and discussion

We conducted a stated choice experiment that was administered to the RAND American Life panel, a large probability sample drawn from the general U.S. adult population. Our aim was to disentangle the effects of personal and supplementary sources of information on the decision making process for full knee replacement surgery to treat knee osteoarthritis and to examine the extent to which respondents are willing to consider these supplementary sources of information. Subjects saw a series of randomly selected video vignettes which varied not only in dimension levels but also in whether specific sources of information were presented, and gave recommendations - knee surgery or not - for each of the persons described in these vignettes.

Through our multivariate analysis of vignettes without patient testimonials, we found that the effects of a high level of pain, a positive surgeon's recommendation, as well as a second opinion that raised substantial reservations about full knee replacement surgery were the strongest impact factors on respondent's decision making. Person-specific outcome forecasts as well as high opportunity costs also strongly affected the decisions. In vignettes which included patient testimonials in the design, these patterns persist, but the effects are not as strong. The patient testimonials were influential when two testimonials with a consistently positive or a consistently negative message were presented.

Respondents showed great interest in receiving supplementary information. When consideration of supplementary information was optional, the vast majority of respondents elected to receive the information. Further, each of the supplementary sources of information affected respondent recommendations. Thus, we established that respondents are deliberately willing to seek out these sources of information that additionally turned out to be influential factors of respondents' decision making processes.

The positive effects of second opinions are reassuring since patients are often encouraged by health educators to seek second opinions in making difficult treatment decisions. The independent effect of a patient-specific outcome forecast is also encouraging for developers of decision guides who hope that patients will be influenced by scientific evidence on treatment outcome.

Interpreting the effects of viewing two testimonials with a consistently positive or consistently negative message is more complex. Introducing testimonials to the decision making process made the dominance of the effects of the other main impact factors less clear. There are two possible reasons for this finding. First, the sample that included testimonials is smaller than the sample that does not include testimonials. This increases the standard errors, which makes it more difficult to establish statistical differences. Second, the introduction of testimonials itself resulted in marginal, although not statistically significant, changes in the impact of the other dimension levels that could be explained by a crowding out effect.

In light of the full research protocol, it may be surprising that testimonials had any effect on recommendations. The respondents were carefully briefed through a physician lecture on knee osteoarthritis, treatment options including full knee replacement surgery, and the risks and side effects associated with this surgery and rehabilitation. They were presented with a good deal of information that described each vignette person's personal situation including the symptoms, treatment history, pain level, employment status, and opportunity costs that would be encountered in the case of surgery. Furthermore, they were presented with the orthopedic surgeon's recommendation as well as the recommendation by a second physician, described as a specialist in treating knee osteoarthritis. They reviewed person-specific outcome forecasts that were presented as based on a large study of full knee replacement surgery patients funded by the U.S. Department of Health and Human Services. Despite this person-specific, professional and objective information, the subjective testimonial of two lay people's negative or positive experiences with full knee replacement surgery had an effect. Additionally, this effect was in magnitude similar to the effects of the most influential dimension levels, including the effects of the other supplementary information features. To provide a specific example: The effect of viewing two subjective negative patient testimonials basically cancels out the effect of an above average person-specific outcome forecast that was based on a very large number of surgery experiences.

Finding testimonial effects in light of the extensive amount of less subjective information presented to respondents and the magnitude of the testimonial effects is noteworthy. The fact that the vast majority of respondents chose to view testimonials when offered is evidence that the public finds testimonials attractive.

Limitations

One limitation that may affect the generalizability of this study is the use of a discrete choice experiment, a stated preference approach. We used hypothetical scenarios to elicit hypothetical choices. Our premise is that these choices approximate the behavior of people in real situations. There is longstanding debate about the degree to which results obtained from stated preference approaches can be generalized to real world decisions (Diamond & Hausman (1994), Hensher (2010), Mark & Swait (2004), Harrison (2006), Kesternich et al. (2012)). In the design of our experiment, we followed as closely as possible the advice put forward in Lanscar & Louviere (2008) to construct a "best-practice" DCE. Our experimental script was carefully developed with the aid of medical practitioners to define salient and realistic attributes and our respondents were extensively debriefed of the medical condition at hand in an effort to provide a common information base for all respondents. Furthermore, we employed a fractional factorial design which allowed us to keep respondent burden to a minimum while preserving orthogonally. We conducted a pilot study with more than 500 respondents and conducted qualitative interviews with volunteers. Our study sample contains respondents from a general population sample in the age range that is usually affected by the medical issue at hand. Additionally, the evidence regarding the biases resulting from the hypothetical nature is less for discrete choice experiments than for other stated preference approaches such as for example contingent valuation (Hensher (2010)). Furthermore, Kesternich et al. (2012) found that hypothetical choices and actual actions in insurance markets are clearly related, so that hypothetical choices can be used to predict and analyze demand. In spite of the uncertainties about the potential for generalizing from stated-choice studies, these designs make it possible to examine the effects of options that could not be studied in research on actual patients because of ethical concerns.

A limitation concerns the findings regarding the testimonial effects. The testimonials were the last piece of information provided to respondents and thus there is the possibility of recency effects (Baddeley (1990)). However, we offered respondents summaries of the other information in vignettes after they viewed the testimonials but before they made their recommendation. We anticipated that the summaries would remind respondents of vignette details, and thus, be as available as the testimonials at the moment of decision making. The question whether testimonial effects were possibly increased by recency effects could be addressed through further experimentation by introducing testimonials earlier. The sequence in which we introduced information was guided generally by our anticipation of the sequence with which patients might obtain information. In future research, the protocol might be modified so that the testimonials would be introduced before the other supplementary information.

Our research did not address questions about the effects of general outcome forecasts. Our findings suggest that patients would be influenced by forecasts specific to their circumstances. The currently available decision guides offer only general outcome forecasts. If general outcome forecasts have as strong an effect on patient choices, the expense of developing patient-specific forecasts would be unnecessary. Future experimental research might systematically compare the effects of patient-specific and general outcome forecasts.

Questions might be raised about the small samples included in the experimental design. Because of time and budget constraints, we used six vignette persons, one orthopedic surgeon, one consulting physician, and two sets of testimonials. Our design would have been stronger if we had used a larger panel of vignette persons with unique presentations of their circumstances, more orthopedic surgeons using varied language in making recommendations, more consulting physicians using their own language in offering second opinions, and a larger set of favorable and unfavorable testimonials. We also limited respondents to two testimonials. The fact that we found that the number of testimonials made a difference (when they were in the same direction) invites questions about the effects of viewing a larger number of testimonials. However, presenting more hypothetical choice tasks could also induce fatigue, and there arises an interesting question about the optimal number of such tasks given the trade-off between obtaining more information and that information potentially becoming less reliable. These issues could be addressed through further experimental research on these questions using stated preference methods.

2.7 Appendix

2.7.1 Appendix A: Example vignette person

This appendix illustrates the full design of our vignette persons for an example vignette person. Table 5 displays for each of the dimensions the language and content for each of the levels using the vignette person Bill. The respondents in the experiment were presented with the introduction and a randomly chosen level of each dimension for a randomly chosen vignette person at each of the three vignette observations. The testimonial dimension was only available at the third vignette observation.

Dimension	Level	Example script
Introduction		I'm Bill. Over the past few years, I've had increasing pain in my left knee. It is difficult to walk very far or to climb stairs. My doctor diagnosed me with knee osteoarthritis. He advised me to lose weight and sent me to physical therapy. I am trying to diet and do the exercises prescribed by the therapist. I also take pain medication to help me sleep. In spite of everything I am doing, my condition has not improved as much as I would like. I'm now thinking about knee replacement surgery.
Pain	A. High	A. My pain has gotten worse. It hurts to do much of anything. I have trouble sleeping even with medication. I'm getting really worn down.
	B. Moderate	B. I experience some pain much of the time. Most of the time, medication relieves the pain.
Opportunity costs and employment status	A. Employed(high)	A. I'm a salesman at a department store. I work long hours and am on my feet all day. My doctor told me I probably won't be able to work for a few months if I have surgery. Since most of my salary is from commissions, this will be tough.
	B. Employed(low)	B. I'm a consultant and have a home office. I do most of my work on the phone or computer. I also can work my own hours. Three months recovery time will be inconvenient, but I can work around it.
	C. Retired(high)	C. Since I retired, I spend most days out with my buddies fixing up old cars and going to road rallies. I live alone so I go out to eat a lot. My doctor said I won't be able to drive or do work on the cars for quite a while. Even if I can get help with meals at home, this will really disrupt my routine.
	D. Retired(low)	D. Now that I'm retired, I live a pretty low-key life. I spend most of the time on the internet or watching TV. I'm pretty handy with a microwave oven. I would have time for rehabilitation.
Orthopedic surgeon's recommendation	A. Positive	A. My name is Dr. Duncan. As Bill's orthopedic surgeon, I recommend that he have total knee replacement surgery. The damage to his knees caused by the arthritis is severe. No other treatment will give this patient the relief from pain that he will get from surgery.
	B. Neutral	B. My name is Dr. Duncan. I have examined Bill as his orthopedic surgeon and talked with him about full knee replacement surgery. The surgery is likely to benefit Bill, but he has to weigh the advantages and disadvantages before making a decision whether to go ahead now. If Bill does not have surgery, he can continue to benefit from other therapies like exercise, weight control, and pain medication. Surgery can also be considered in the future if his condition deteriorates further.

Dimension	Level	Example script
Specialist's second opinion	A. Strong recommendation	A. My name is Dr. Abbott. I am an orthopedic surgeon who specializes in treating patients with knee osteoarthritis. Bill has sought a second opinion from me. I have carefully reviewed Bill's health history and examined his knee. I have determined that in Bill's case, surgery at this time is the best option for long-term pain relief.
	B. Neutral recommendation	B. My name is Dr. Abbott. I am an orthopedic surgeon who specializes in treating patients with knee osteoarthritis. Bill has sought a second opinion from me. I have carefully reviewed Bill's health history and examined his knee. I have determined that in Bill's case surgery is not necessary now. At this time, he will benefit just as well from continuing non-surgical therapies such as targeted exercises and pain medication.
	C. Not available	C. -
Person-specific outcome forecast	A. Above average	A. Verbal format: When Bill entered his personal and health information, the predicted surgical outcome was: Higher than the average patient Numeric format: The average patient has an 80% chance of a successful surgical outcome for knee replacement surgery. When Bill entered his personal and health information, the predicted surgical outcome was: Above average: 95%. Graphic format: The average patient has an 80% chance of a successful surgical outcome for knee replacement surgery. When Bill entered his personal and health information, he saw the following chart that shows his chances of a successful surgical outcome compared with an average patient. Above average: GRAPH (95% and 80%)
	B. Below average	B. Verbal format: When Bill entered his personal and health information, the predicted surgical outcome was: Lower than the average patient Numeric format: The average patient has an 80% chance of a successful surgical outcome for knee replacement surgery. When Bill entered his personal and health information, the predicted surgical outcome was: Below average: 65%. Graphic format: The average patient has an 80% chance of a successful surgical outcome for knee replacement surgery. When Bill entered his personal and health information, he saw the following chart that shows his chances of a successful surgical outcome compared with an average patient. Below average: GRAPH (65% and 80%)
	C. Not available	C. -

Dimension	Level	Example script
First patient testimonial	A. Positive 1	A. I am so happy that I had my knee replaced. I did the five months of physical therapy that was recommended. I took no pain medication other than Tylenol, and I was able to drive and go back to work within a few weeks. I have never had any problems since. I am able to go for long walks and climb stairs comfortably. It is nice to go through a cold winter and not have any excruciating pain and swelling like I had.
	B. Positive 2	B. I put off the surgery for a long time because of concerns that it would not work out well. Now that I have had the surgery, I am very pleased with the result. The rehabilitation asked a lot of me, but I was able to go through with it one day at a time. Now that half a year has gone by since the surgery, I am able to walk normally and my ability to climb stairs has improved a great deal. The pain in my knee has almost completely disappeared.
	C. Negative 1	C. I had my right knee replaced six months ago. It was very painful requiring powerful medications every day. Even after three different series of physical therapy and faithfully doing my exercises at home, I still have constant pain on the right side of my knee. I am able to walk only short distances. Climbing stairs is difficult. This has been going on for six months. I am totally discouraged.
	D. Negative 2	D. I had high hopes for the surgery. The result has been a major disappointment even though I did everything that I was asked to do with the rehabilitation. I have been back to see the doctor several times since the surgery. Nothing the doctor suggests seems to work. Walking is painful. Climbing stairs is worse. Without medication, the pain keeps me awake at night.
	E. Not available	E. -
Second patient testimonial	A. Positive 1 B. Positive 2 C. Negative 1 D. Negative 2 E. Not available	The second testimonials were taken from the same pool of scripts ensuring that no respondent was presented with the same script twice.

Table 2.5: Vignette dimensions, levels and script for an example vignette person

2.7.2 Appendix B: Randomization into the text and video survey versions

This appendix examines the randomization of our study sample into the text and video versions of our survey and the randomization of our dimensions within the full sample and these subsamples.

The sample used in the main part of the paper consists of 1616 respondents. Of these, 339 (21%) were randomly allocated to the text-only version of our survey. The remaining 1277 were allocated to the video version. Table 6 provides descriptive statistics of respondents' personal characteristics for the full sample and the text-only and video subsamples. The first column is identical to the descriptive statistics of the full sample presented in Table 2 in the main part of the paper. The next two columns split up the full sample into the text-only and video subsamples. The final column displays p-values for tests of equality of variable means between the two subsamples.

The means of the personal characteristics of the two subsamples are comparable for most variables. The only statistically significant differences at the 5% level are observed for the category of respondents that are older than 70 years as well as for the categories of working and retired respondents.[13] Since these older respondents are also less likely to be still in the workforce and more likely to be retired, the observed differences in the means of the employment status variables are likely to be a consequence of the difference in age between the two subsamples.

[13] P-values adjusted for multiple testing with the Bonferroni method are all equal to 1. Given that we test 20 hypotheses here, the three significant differences at the 5% level we observe in the unadjusted p-values could also be due to pure chance.

Sample Variable	Full Mean	Text-only Mean	Video Mean	P-value*
Gender				
Male	0.43	0.44	0.43	0.907
Age				
50-59	0.50	0.52	0.49	0.373
60-69	0.34	0.35	0.33	0.528
70 or older	0.16	0.13	0.17	0.044
Income				
below 25,000$	0.23	0.25	0.23	0.334
$\geq$ 25,000\$, $<$ 50,000\$	0.27	0.27	0.27	0.947
$\geq$ 50,000\$, $<$ 75,000\$	0.16	0.14	0.17	0.156
above 75,000$	0.33	0.34	0.33	0.861
Living Status				
Married or living with a partner	0.58	0.58	0.58	0.941
Employment Status				
Retired	0.34	0.29	0.36	0.035
Working	0.45	0.50	0.44	0.030
Unemployed, disabled and other	0.20	0.20	0.21	0.853
Education				
High school or less	0.24	0.25	0.24	0.665
At most Bachelor's Degree	0.59	0.57	0.60	0.401
Post graduate	0.17	0.18	0.17	0.543
Ethnicity				
Non-Hispanic white	0.83	0.80	0.83	0.098
Respondents with chronic knee pain	0.42	0.44	0.42	0.508
Respondents with knee osteoarthritis	0.21	0.20	0.21	0.659
Friends/Relatives with knee osteoarthritis	0.66	0.69	0.65	0.207
Friends/Relatives with full knee replacement surgery	0.53	0.56	0.52	0.146
Number of observations	1616	339	1277	

*P-value for the difference in means of the respective variable between the text-only and video subsamples.

Table 2.6: Descriptive statistics of the full sample and the text-only and video subsamples

Table 7 describes how the dimension levels were allocated to the full sample and the text-only and video subsamples. This table is similar table 3 in the main part of the paper. It is split up into two panels. Panel A provides level frequencies for the full sample, the text-only subsample and the video subsample for the dimensions that were available at all three vignette observations. Panel B displays the same information for the testimonial versions that were available at the third vignette observation only. In contrast to table 3 in the main part of the paper, this table focuses on the assigned levels of the dimensions. For the optional features, the realized frequencies of levels reported in the main paper can be different.

The first column displays the assigned frequencies of the dimension levels in the full sample. These correspond closely to the planned frequencies that were presented in table 1 in the main paper. In Panel B, these frequencies deviate a little, however this is also a much smaller sample. The next two columns present these frequencies for the text-only subsample and for the video subsample. The assigned frequencies are almost identical in both subsamples. The only significant difference is observed for the first testimonial, here the two subsamples were assigned with different frequencies to the positive testimonial version 2 ($P<0.05$).

Sample	Full	Text-only	Video
Panel A (n=4845)			
Sample split by:			
Pain level			
high pain	51.41	51.67	51.34
low pain	48.59	48.33	48.66
Opportunity cost level			
high opportunity costs	50.38	50.00	50.48
low opportunity costs	49.62	50.00	49.52
Employment status			
retired	51.12	51.67	50.98
employed	48.88	48.33	49.02
Surgeon's recommendation level			
positive surgeon's recommendation	24.56	24.80	24.50
neutral surgeon's recommendation	75.44	75.20	75.50
Second opinion level assigned			
strong recommendation	37.52	35.73	38.00
substantial reservation	37.81	38.58	37.61
second opinion not available	24.66	25.69	24.39
Person-specific outcome forecast level assigned			
forecast above average	37.38	38.68	37.03
forecast below average	38.10	37.50	38.26
numeric forecast format	24.54	25.30	24.34
graphic forecast format	26.36	26.97	26.19
verbal forecast format	24.58	23.92	24.76
forecast not available	24.52	23.82	24.71
Panel B (n=1614)			
Sample split by:			
Patient testimonial version assigned			
Testimonials not available	25.28	28.02	24.55
First testimonial			
positive 1	17.53	17.40	17.57
positive 2	18.96	14.16	20.24
negative 1	19.02	18.88	19.06
negative 2	19.21	21.53	18.59
Second testimonial			
positive 1	18.15	16.22	18.67
positive 2	20.32	20.35	20.31
negative 1	17.47	17.70	17.41
negative 2	18.77	17.70	19.06

Table 2.7: Randomization of dimension levels in the full sample and the text-only and video subsamples

The statistics presented in this appendix are strong evidence that our randomization procedure succeeded in randomly assigning respondents into the different media versions and in randomly assigning the dimension levels as they were intended for the full sample as well as the subsamples.

2.7.3 Appendix C: Respondents' survey behavior

This appendix examines various aspects of respondents' survey behavior. First, while the previous appendix presented the distribution of respondents in the assigned media treatments, this section now presents and examines the realized allocation of respondents into the text version and the video version of our survey. The assigned and realized media version may vary because we had to enable respondents with difficulties viewing videos a possibility to continue the survey. The group of respondents that made use of this possibility will from now on be called "deviators". All subsequent analysis maintains this division of the full sample into text-only, deviator and realized video-only subsamples. The appendix then proceeds to examine the duration of the survey as well as the time respondents spent on reviewing the vignette person's situation, the surgeon's recommendation, the supplementary information features as well as the time respondents spent on the summaries and the decision screens. Finally, this appendix examines the overall quality of the survey for the full sample as well as for the three subsamples. First, we describe respondents' assessment of their survey experience. Then, we describe the extent of item non-response. Next, we examine the extent to which respondents speed through the survey and whether this can be associated their personal characteristics. Finally, we develop measures for good decisions which will then be used to gauge the extent of "good" decisions in our discrete choice experiment and summarize our results.

Realized distribution of respondents in text and video versions

As noted in the introduction to this appendix, this subsection describes and examines the final allocation of respondents to text and video versions as it was realized. To enable respondents to continue the survey even if they could not watch the videos, we provided a video setup check at the beginning of the survey and offered an opt-out opportunity for all videos. If the video setup failed, respondents were switched to the text version of our survey entirely. If they stated at a single video that they could not watch this video, the respective text version of this video was presented. The group of respondents that failed the video setup or opted out of the videos at several instances will be called "deviators" since they deviate from our original randomized assignment to media versions. In total 229 respondents failed the video setup and were thus allocated to the text version. This is the first group of respondents that makes up the deviator subsample. Additionally, respondents that skipped out of the vignette person's video and the surgeon's recommendation for at least two vignette observations are also part of this group. This definition applies to a total of 129 respondents.[14] In sum, 358 respondents deviated from our original assignment to the video survey version.

Table 8 divides the sample into the three groups of respondents that were observed in our study. The first column presents the personal characteristics of the text-only subsample, the second of the "deviator" subsample and the third of the realized video-only subsample. The last column displays the p-value of group comparison mean tests between the deviator and the video-only subsample.

The main differences between the deviator and the video-only subsamples are observed for age and employment status. The deviator subsample is older and consequently also more respondents in this group are retired. Additionally, a larger fraction of these respondents is in the low income category which might also be related to age and retirement. Further differences are observed ethnicity and marital status. For the older respondents in this subsample, it is likely that viewing the video files constituted a challenge. We do however still observe a substantive fraction of respondents that is younger than sixty years in this subsample. These belong to an age group that we expected to be familiar with the technology we used. Therefore, for a subgroup of these respondents it is questionable whether they truly failed the video setup or truly opted out of the video version because of having trouble viewing the videos.

[14] With this definition, we ensure that only respondents are assigned to the deviator group that requested the text versions for a large part of the survey. Most of these respondents also requested text versions for all three vignette person's videos and their surgeon's recommendations as well as for the videos that followed the surgeon's recommendation.

Sample Variable	Text-only Mean	Deviators Mean	Video-only Mean	P-value*
Gender				
Male	0.43	0.42	0.44	0.447
Age				
50-59	0.50	0.45	0.51	0.021
60-69	0.34	0.29	0.35	0.049
70 or older	0.16	0.26	0.14	0.000
Income				
below 25,000 $	0.23	0.32	0.19	0.000
$\geq$ 25, 000\$, < 50, 000\$	0.27	0.29	0.27	0.397
$\geq$ 50, 000\$, < 75, 000\$	0.16	0.16	0.17	0.922
above 75,000 $	0.33	0.23	0.37	0.000
Living Status				
Married or living with a partner	0.58	0.52	0.61	0.004
Employment Status				
Retired	0.34	0.43	0.33	0.000
Working	0.45	0.34	0.48	0.000
Unemployed, disabled and other	0.20	0.23	0.20	0.187
Education				
High school or less	0.24	0.27	0.22	0.080
At most Bachelor's Degree	0.59	0.58	0.61	0.464
Post graduate	0.17	0.15	0.17	0.301
Ethnicity				
Non-Hispanic white	0.83	0.78	0.86	0.012
Respondents with chronic knee pain	0.42	0.44	0.41	0.441
Respondents with knee osteoarthritis	0.21	0.23	0.21	0.305
Friends/Relatives with knee osteoarthritis	0.66	0.67	0.64	0.678
Friends/Relatives with full knee replacement surgery	0.53	0.54	0.51	0.639
Number of observations	339	358	919	

*P-value for the difference in means of the respective variable between the deviator and video-only subsamples.

Table 2.8: Descriptive statistics of the realized text-only, deviator and video-only subsamples

Description of respondents' survey behavior for reviewing content and making decisions

This subsection discusses the survey behavior of respondents with respect to the time taken by respondents for completing the survey, reviewing vignette content and making decisions. Table 9 presents the mean time spent on the entire survey and the standard deviation for the full sample and the three subsamples. The distribution of the time spent on completing the entire survey has a high standard deviation and is heavily skewed with a long right tail. Therefore, the mean time spent on completing the survey is high and amounts to approximately 34 minutes. This moment is however driven by extreme outliers. Therefore, this table displays in addition the 25th, the 50th and the 75th percentile to reduce the distortions caused by respondents that took extremely long to complete the survey. For the full sample, the median time spent on the survey is 905 seconds or approximately 15 minutes.

Sample	Full	Text-only	Deviators	Video-only
Total time in seconds				
Mean	2007	4438	1505	1306
Std. Dev.	21470	46402	4146	2951
Percentiles				
25th	689	481	665	802
50th	905	637	877	987
75th	1225	988	1317	1267

Table 2.9: Distribution of total time spent on the survey

A similar pattern is observed for the subsamples. The mean total survey times in the three subsamples suggest that respondents in the text-only subsample spent approximately three times as much time on completing the survey than respondents in the deviator and video-only subsample. However, this is again misleading, since the standard deviation in the text-only sample is driven up by three extreme observations where the total time spent on the survey exceeds six hours. Therefore, the following discussion focuses on the percentiles of the distribution. When we look at the percentiles of total time spent on the survey in the three subsamples, the text-only subsample has the shortest duration for all three percentiles. On average the respondents in the text-only subsample finished the survey faster with 25 percent of respondents completing the survey in eight minutes or less. The median of the video-only subsample is approximately six minutes higher than the median of the text-only subsample($P<0.01$)[15]. The median of the deviator subsample is a little lower than of the video-only subsample ($P<0.01$) and higher than of the text-only subsample ($P<0.01$). This suggests that at least a part of the respondents in the deviator subsample truly had issues with the survey and did not switch to the text version to save time.

Next, the time respondents spent on reviewing the vignette content is examined. Table 10 presents the distribution of the time spent on reviewing the narration of the vignette

[15] The p-values in this subsection are based upon non-parametric k-sample tests for equality of medians.

person's personal situation, the surgeon's recommendation, the specialist's second opinion, the outcome forecast and the first and second patient testimonial for the full sample and the three subsamples.[16]

[16] For the deviator subsample the time spent on a single vignette dimension is the video time if only the video was viewed and the time spent on the text version if the text version was requested. This way we can isolate the time spent on reviewing content and do not observe a mixture between time needed to request the text version and the time actually spent on reviewing vignette content.

Sample	Full	Text-only	Deviators	Video-only
Total time in seconds spent on:				
Personal situation				
Mean	91	126	49	95
Std. Dev.	1043	2232	109	267
Percentiles				
25th	33	23	23	62
50th	61	32	34	70
75th	74	46	48	81
Surgeon's recommendation				
Mean	38	22	22	50
Std. Dev.	159	36	22	208
Percentiles				
25th	16	11	11	31
50th	31	16	17	39
75th	41	24	26	44
Second opinion				
Mean	29	20	17	38
Std. Dev.	48	78	14	39
Percentiles				
25th	14	10	10	29
50th	27	14	14	32
75th	34	19	20	37
Outcome forecast				
Mean	33	25	47	31
Std. Dev.	301	29	605	121
Percentiles				
25th	10	10	10	11
50th	22	19	20	25
75th	36	31	34	38
First patient testimonial				
Mean	559	2561	31	58
Std. Dev.	17803	39581	94	96
Percentiles				
25th	21	16	14	40
50th	39	22	23	48
75th	51	30	31	56
Second patient testimonial				
Mean	35	22	24	45
Std. Dev.	48	49	73	29
Percentiles				
25th	16	10	11	37
50th	34	15	16	42
75th	43	22	22	47

Table 2.10: Distribution of time spent on vignette content

Again the distribution is characterized in a number of cases by extreme observations that drive the mean time spent and the standard deviation up. When we look at the median time spent on the single vignette content parts, the median time spent on reviewing the vignette person's personal situation is highest and the median time spent on reviewing the content of the second opinion is the lowest in the full sample.

Furthermore, the distribution of the time spent on the single parts is very similar for the text-only and deviator subsample[17] and the median time spent in these subsamples is considerably lower than for the video-only subsample ($P<0.01$). The only dimension where the median time spent in all three subsamples is fairly close is the outcome forecast dimension. This can be explained by the fact that this information was displayed in text format in all subsamples and was only supported by a short automatically played narration in the video-only subsample. However, the video-only subsample's median is still significantly lower than the median of the other subsamples for this dimension ($P<0.01$). The variation of the time needed for a single content part in all three subsamples is in part due to the different length of the scripts for the single vignette persons and for the different levels of the other content parts. For the respondents in the text-only and deviator subsample additional variation arises because of differing reading abilities and possibly different reading strategies of respondents. For the video-only subsample additional variation arises because of different internet connections and computer hardware which varied the time it took to load the videos.

To assess the extent to which respondents completely read the text we presented them with in the text version, we now by way of example compare the time spent on the vignette person's personal situation with the time necessary to read this amount of text assuming an average reading speed of 250 words per minute (McNair (2009)). The shortest possible text of a vignette's personal description in our study contained 111 words. With the average reading speed assumption, respondents should have needed at least 26.6 seconds to fully read the content provided in the personal situation descriptions (since all other scripts were longer). However, roughly 32% of respondents in the text-only subsample and roughly 31% of respondents in the deviator subsample spent less than 26 seconds on this content. This suggests that substantive fraction of the respondents that reviewed the text version did not completely read through the provided information. A similar behavior is observed for the

[17] For these two groups the only statistically significant difference is observed in the median time spent on reviewing the content provided in the surgeon's recommendation ($P<0.05$).

other dimensions. In the video-only subsample some respondents also exhibit an extremely low time spent on viewing the videos suggesting that they did not watch the videos entirely. In this subsample, we observe a steady amount of single observations that each spent little time on the videos. This is then followed by a clear clustering of at least 50 observations, which suggests that at this point the first videos were done loading and playing. For the videos containing the personal situation of the vignette persons, the first cluster is observed after roughly 57 seconds. Only 14.3% of respondents spent less time reviewing this dimension. A similar extend of this behavior is observed for the other dimensions that contained videos.

Next we turn to the analysis of the time spent on reviewing the summaries that were displayed before the respondent was presented with the decision task. For reviewing the summaries, there is the possibility of learning effects in the sense that after the first vignette observation, respondents know that the summaries exist. They might therefore reduce the time spent on reviewing vignette content and instead rely on the information given in the summaries. If that was the case, we should observe that the time spent on the summaries increases for the following vignette observations while the time spent on the vignette dimensions should decrease by a large extend. Table 11 presents the distribution of time spent on this survey feature separately for each of the three vignette observations in consecutive order. In the full sample, the median time spent on the summaries decreases from 15 seconds at the first vignette observation to 11 at the third vignette observation.

The distribution of time spent on the summaries is similar for the three subsamples. None of the observed differences between the text-only and deviator subsamples are statistically significant at the 5%-level. The median time spent of the video-only subsample is significantly lower than the median time spent of the deviator and text-only subsample at the first and second vignette observation and lower than the median time spent of the deviator subsample at the third vignette observation. Furthermore, during the course of our discrete choice experiment the median time spent on the summaries exhibits a significant downward trend all three subsamples (P<0.01). In a separate analysis not tabulated here, we also examined whether the time respondents spent on the single dimensions of the vignette content decreases from the first vignette observation to the third. For most dimensions, we observe only a slight downward trend in the time spent on reviewing these that is similar for all three

Sample	Full	Text-only	Deviators	Video-only
Total time in seconds spent on summaries:				
At the first vignette observation				
Mean	18	18	20	18
Std. Dev.	17	11	14	20
Percentiles				
25th	11	12	12	11
50th	15	16	17	15
75th	21	21	23	20
At the second vignette observation				
Mean	15	16	17	15
Std. Dev.	18	25	15	16
Percentiles				
25th	8	9	9	8
50th	12	12	13	12
75th	18	18	19	17
At the third vignette observation				
Mean	15	14	17	14
Std. Dev.	20	16	20	21
Percentiles				
25th	7	7	8	7
50th	11	11	12	11
75th	17	17	19	16

Table 2.11: Distribution of time spent on vignette summaries

subsamples. The differences in median time spent observed between the first and third vignette observations are statistically significant at the 1%-level. However, the magnitude of most these differences is very low, ranging from 2 to 6 seconds. This decrease can be attributed to the fact that the dimensions and their introductions were more familiar at the second and third vignette observations. A substantive downward trend was only found for the time taken to review the information contained in the outcome forecast. This time decreases by approximately 20 to 25 seconds from the first vignette observation for all three subsamples. This decrease can be explained by the fact that the outcome forecast was presented with a rather lengthy and complex introduction. At the first vignette observation, respondents needed some time to read and understand when they were confronted with the outcome forecast dimension for the first time. The evidence presented here does not support the hypothesis that respondents increasingly relied on the summaries instead of reviewing vignette content during the course of the interview.

Finally, we analyze the time spent on deciding whether or not to recommend surgery for a specific vignette person. Table 12 presents the distribution of time taken for the decision separately for each of the three vignette observations in consecutive order.

Again, we observe that the distributions are comparable across groups and the median time spent on the decision screen also exhibits a slight downward trend after the first vignette (P<0.01). This can be explained by the fact that by the second vignette observation the decision task and the question wording was already known and thus less time was needed to read the information contained in the decision screen. Respondents in the video-only subsample did not take longer to decide than respondents in the text version of the survey.

Sample	Full	Text-only	Deviators	Video-only
Total time in seconds spent on decision screen:				
At the first vignette observation				
Mean	10	10	13	9
Std. Dev.	18	8	31	12
Percentiles				
25th	5	5	6	5
50th	7	7	8	7
75th	10	10	12	9
At the second vignette observation				
Mean	8	8	9	8
Std. Dev.	9	5	9	10
Percentiles				
25th	5	4	5	5
50th	6	6	6	6
75th	8	8	10	8
At the third vignette observation				
Mean	8	7	9	7
Std. Dev.	10	5	10	11
Percentiles				
25th	4	4	5	4
50th	6	6	6	5
75th	8	8	9	7

Table 2.12: Distribution of time spent on decision screen

This subsection examined the time respondents spent on reviewing the information we provided them with and on making decisions. The median time spent on this survey was approximately 15 minutes but differed substantially between the video-only and text-only subsamples. Respondents in the video-only subsample needed on average 6 minutes longer than respondents in the text-only subsample. The same applies to the time spent on reviewing the single vignette dimensions. In the subsamples that reviewed the text version of our survey, a non-trivial fraction of respondents exhibits extremely low times spent on the dimension content, suggesting that they did not read the provided text entirely. For the respondents in the video-only subsample we also observe some respondents with extremely low times. However, in all three subsamples, we could not find evidence that supports the hypothesis that many respondents solely relied on the information provided in the summaries instead of reviewing the full dimension content. Additionally, we found that the time spent on the decision screens is comparable between all three subsamples. Since response times can be used as an indicator for respondents' cognitive burden (Bassili & Scott (1996)), we could not find an indication for an increased respondent burden in the video-only subsample.

Description of overall survey quality

This final part of this appendix examines the overall survey quality and summarizes the results of this appendix. First, the prevalence of item non-response is described, then respondents' assessment of the interest in our survey as well as the degree of identification with our vignette persons is presented. Next, the extent of speeding through the survey content is examined. Finally, we develop a measure for good decisions in our application and assess the extent of good decisions that we observed.

Item-nonresponse in this survey is a rather minor issue. For all surgery recommendations, we only miss a response in three instances. For the questions asked in a follow up and background questionnaire, a similar picture arises. A total of 123 respondents did not provide an assessment of at least one question. Of these respondents, 116 only skipped one item, the remaining 7 skipped two items. The probability that a respondent does not respond to an item is positively associated with the deviator group of respondents ($P<0.05$).

Table 13 presents respondents' assessment of the interest in our survey as well as the degree of identification with our vignette persons. The assessment of the interest in the survey was elicited by the American Life Panel using their following standard question:

> Could you tell us how interesting or uninteresting you found the questions in this interview?
> Very interesting
> Interesting
> Neither interesting nor uninteresting
> Uninteresting
> Very uninteresting

The degree of identification with our vignette persons was elicited by us in a follow-up question using the following question wording:

> Overall, during the course of this interview, to what extend were you able to identify with the people described in the survey?
>
> Very
>
> Somewhat
>
> Slightly
>
> Not at all

The first column table 13 describes the distribution of the assessment of the interest in the survey, the second of the assessment of the degree of identification with the vignette persons. Panel A presents the distribution for the full sample, Panel B for the text-only subsample, Panel C for the deviator subsample and Panel D for the video-only subsample.

Overall, 93% of respondents assessed the interview as either interesting or very interesting and only 2% stated that they considered the interview uninteresting or very uninteresting. Furthermore, 75% of respondents stated that they could at least somewhat identify with the vignette persons we had presented them with. In all three subsamples, the percentages of respondents picking the top two or bottom two categories are not statistically significant at the 5%-level. In sum, this survey has thus managed to generate a high respondent interest and a high degree of identification with our vignette persons for both media versions.

Question: Could you tell us how interesting or uninteresting you found the questions in this interview?			Question: Overall, during the course of this interview, to what extend were you able to identify with the people described in the survey?		
Panel A: Full sample					
Categories	Frequency	Percent	Categories	Frequency	Percent
Very interesting	854	53.01	Very	650	40.25
Interesting	638	39.60	Somewhat	557	34.49
Neither interesting nor uninteresting	92	5.71	Slightly	270	16.72
Uninteresting	15	0.93	Not at all	138	8.54
Very uninteresting	12	0.74			
Total	1611	100.00	Total	1615	100.00
Panel B: Text-only subsample					
Categories	Frequency	Percent	Categories	Frequency	Percent
Very interesting	174	51.48	Very	142	41.89
Interesting	137	40.53	Somewhat	109	32.15
Neither interesting nor uninteresting	25	7.40	Slightly	59	17.40
Uninteresting	1	0.30	Not at all	29	8.55
Very uninteresting	1	0.30			
Total	338	100.00	Total	339	100.00
Panel C: Deviator subsample					
Categories	Frequency	Percent	Categories	Frequency	Percent
Very interesting	182	50.98	Very	148	41.34
Interesting	149	41.74	Somewhat	125	34.92
Neither interesting nor uninteresting	20	5.60	Slightly	59	16.48
Uninteresting	2	0.56	Not at all	26	7.26
Very uninteresting	4	1.12			
Total	357	100.00	Total	358	100.00
Panel D: Video-only subsample					
Categories	Frequency	Percent	Categories	Frequency	Percent
Very interesting	498	54.37	Very	360	39.22
Interesting	352	38.43	Somewhat	323	35.19
Neither interesting nor uninteresting	47	5.13	Slightly	152	16.56
Uninteresting	12	1.31	Not at all	83	9.04
Very uninteresting	7	0.76			
Total	916	100.00	Total	918	100.00

Table 2.13: Assessment of the interest in the survey and the identification with vignette persons

However, we have to acknowledge that a small subgroup of respondents appears to not have taken the survey seriously. Some of our respondents spent extremely little time on reviewing vignette content. This group and its survey behavior is now described in more detail.

While the previous subsection of this appendix examined the overall distribution of the time respondents spent on the single parts of our survey, we now focus on the extent to which respondents in the three subsamples speed through different parts of the survey. For reviewing the personal situations of the vignette persons, an observation was identified as speeding if the time spent on reviewing this information was 10 seconds or less. Within these 10 seconds, no video could be watched and the text-version of a video could also not be read. For reviewing the surgeon's recommendation, an observation was identified as speeding if the time spent on reviewing this information was 5 seconds or less. The threshold is lower for this information since the amount of information presented in the surgeon's recommendation was considerably less than for the personal situations. With the same argument, the thresholds for speeding behavior were also set to five seconds or less for reviewing the information contained in the second opinion, the outcome forecast, the first and second patient testimonial. For the time spent on the summaries and on the decision screen, an observation was identified as speeding if the respondent spent three seconds or less on the respective screen. Within this time, it is unlikely that the content or the question posed was read and considered.

Table 14 presents the percentage of observations with speeding behavior according to the definition given above for the full sample and the three subsamples for all dimensions. For the full sample, we obtain the highest amount of speeding observations for the description of the vignette person's personal situation and the outcome forecast. For both, roughly 5% of observations are characterized by speeding behavior. This is significantly higher than for all other dimensions (P at least <0.1). The lowest percentage of speeding behavior is observed for the second patient testimonial dimension. The respondents in the second patient testimonial dimension are however a selected sample of respondents since reviewing this information was optional. When we compare the percentages of speeding behavior across the subsamples, the lowest incidence of this behavior is observed for the video-only subsample. The difference to the other subsamples is significant at least at the 5%-level except for the outcome forecast and first testimonial dimension. There are subtle differences between the text-only and deviator subsample. For most dimensions, the incidence of speeding behavior is however not significantly larger in the deviator subsample than in the text-only subsample.

Furthermore, the overall incidence of speeding behavior and the incidence at the single vignette observations and across groups are not consistently associated with respondent's background characteristics.

The only persistent association is a negative link between speeding behavior and higher education.[18]

Sample	Full	Text-only	Deviators	Video-only
Percentage of speeding for:				
Personal situation	5.47	6.89	9.03	3.56
Surgeon's recommendation	4.38	6.20	6.98	2.69
Second opinion	2.33	4.63	3.82	0.91
Outcome forecast	5.14	5.41	5.59	4.86
First patient testimonial	3.53	3.83	6.70	2.18
Second patient testimonial	1.98	3.54	4.19	0.55

Table 2.14: Percentage of observations with speeding behavior for vignette dimensions

Table 15 presents the percentage of observations with speeding behavior for the full sample and the three subsamples for the summaries and the decision screen. As in the previous subsection of this appendix, we hypothesized that the time taken for these survey parts may change during the course of the interview. Therefore, this table is again split up according to the vignette observation.

Sample	Full	Text-only	Deviators	Video-only
Percentage of speeding through summaries:				
At the first vignette observation	1.36	1.47	1.12	1.41
At the second vignette observation	2.79	2.96	2.51	2.83
At the third vignette observation	4.89	3.83	4.75	5.34
Percentage of speeding through decision screen:				
At the first vignette observation	1.73	2.36	1.40	1.63
At the second vignette observation	6.07	7.99	4.75	5.88
At the third vignette observation	9.23	10.03	6.70	9.92

Table 2.15: Percentage of observations with speeding behavior for summaries and decision screens

For both survey parts, we observe that the incidence of speeding observations increases as the survey progressed for the full sample as well as for the three subsamples. At the first vignette observation, almost no speeding is observed for the summaries and the decision screens. At the third vignette observation, roughly 5% of respondents speed through the summaries and roughly 10% of respondents speed through the decision screen. This increase in speeding behavior as the survey progresses is highly significant ($P<0.01$). Additionally,

[18] The same analysis was performed using a stricter threshold for speeding behavior of three or less seconds for all dimensions. The incidence of speeding behavior is then lower ranging from 1 to 2 percent depending upon the considered dimension. However, the same pattern emerges. The lowest incidence of speeding behavior is observed for the video-only subsample and the text-only and deviator subsamples' speeding behavior exhibits small but insignificant differences.

there are subtle differences between the three subsamples. There is no clear pattern for the speeding behavior regarding the summaries. When it comes to speeding through the decision screen, the deviator subsample clearly exhibits a significantly lower extend of speeding behavior than the text-only subsample ($P<0.05$). Furthermore, the overall incidence of speeding behavior, the incidence at the single vignette observations and across subsamples is consistently negatively associated with higher education for the summaries and with being older for the decision screens.

Furthermore, the incidence of speeding behavior is positively correlated between all vignette dimensions, the summaries and the decision screens. Thus, if for example a person speeds through the vignette content, this person is also more likely that to speed through the other dimensions.

The final part of this appendix now develops measures for good decisions in our discrete choice experiment and examines how these measures vary between the three observed respondent subsamples. Due to the randomized assignment of dimension levels to the single vignette persons, we are not able to clearly determine for every generated vignette person whether this specific vignette person should have been recommended for surgery or not. However, this can be done for some of the generated vignette persons. The measures for the good decisions were generated using the following logic: If a vignette person was characterized by a specific combination of dimension levels that are favorable for surgery recommendation, this vignette person was identified as a prime surgery candidate. Consequently, if this vignette person was recommended for surgery by the respondent, this constitutes a good decision. If not, this is a bad decision. Alternatively, if a vignette person was characterized by a specific combination of dimension levels that speak against surgery recommendation, this vignette person was identified as an inferior surgery candidate. Thus, for these vignette persons a good decision was when they were not recommended for surgery and a bad decision when they were recommended for surgery. Vignette persons that do not fall into the category of prime or inferior surgery candidates are dropped from the subsequent analysis. In deciding which dimension levels should be used to generate these measures, we were guided by the strongest impact factors observed in the main part of this paper.

The first measure is a conservative attempt to define prime and inferior surgery candidates. It is entirely based upon the vignette person's personal situation and the surgeon's recommendation. A prime candidate in this measure is defined as a vignette person that has high pain, low opportunity costs and a positive surgeon's recommendation. This definition applies to 312 vignette observations in our sample. Of these observations, 241 were recommended for surgery and are thus vignette observations where a good decision was made. An inferior candidate in this measure is defined as a vignette person with the opposite dimension levels (i.e. low pain, high opportunity costs and a neutral surgeon's recommendation). This applies to 895 vignette observations of which 604 were not recommended for surgery. Thus in total, we observe 845 good decisions and 362 bad decisions. Therefore, the clear majority of decisions (70.0%) made for these vignette observations are good. However, the percentage of good decisions varies within our three subsamples. It is the highest for the video-only subsample (73.1%), followed by the text-only subsample (68.8%). The lowest percentage is observed within the deviator subsample with only 63.3% of good decisions. The difference between the percentage of good decisions between the video-only and deviator subsample is significant at the 1%-level. This difference remains significant when we control for the personal characteristics of the respondents. Thus, if we take this measure of good decisions as a measure for the overall quality of decisions in our survey, the deviator subsample exhibits a lower quality of decisions than the video-only subsample.

The second measure uses a less conservative definition of prime and inferior surgery candidates that is based upon various combinations of dimensions levels. A prime candidate is now defined as a vignette observation for which at least three dimension levels of the following dimensions are favorable for a surgery recommendation: pain, opportunity costs, surgeon's recommendation, second opinion and outcome forecast. With this definition, 1460 vignette observations are considered prime surgery candidates. Of these observations, 1135 (or 77.7%) were recommended for surgery and are thus observations where the respondent made a good decision. An inferior candidate is now defined as a vignette observation for which at least three dimension levels of the same dimensions speak against surgery recommendation. In our sample, this applies to 2370 vignette observations of which 1592 did not get a surgery recommendation. Thus, for the inferior candidates a good decision was made in 67.1%

of vignette observations. In total, we observe a good decision in 71.2% of these vignette observations. Again, the percentage of good decisions varies between our subsamples. The same pattern as with the first measure emerges. The video-only subsample has the highest percentage of good decisions with 73.4%, followed by the text-only subsample with 70.2%. The lowest percentage is again observed for the deviator subsample with 66.8% of good decisions. For this measure, the percentage of good decisions is again significantly larger in the video-only subsample than in the deviator subsample ($P<0.01$) and this difference remains significant when we control for the personal characteristics of the respondents. Again, if we use this measure as a proxy for the quality of decision making, the deviator subsample again exhibits a lower quality of decisions than the video-only subsample.

Overall, this survey has managed to generate a high survey quality. Item-nonresponse was only a minor issue. Respondents' assessment of the interest in the survey and the degree of identification with our hypothetical persons was very high in both survey versions. Some respondents speed through the survey very fast, but this behavior was only observed to a low extend. Our measures for good decisions revealed that approximately 70% of decisions can be classified as good decisions.

With respect to the behavior of the three subsamples, the video-only group exhibits the lowest extend of speeding behavior and the highest amount of good decisions. Despite the fact that respondents in this subsample experienced a considerably longer interview duration which could lead to increased fatigue therefore lower effort in decision making (Tourangeau et al. (2000)), we found no indication of a lower quality of survey behavior in this subsample. Additionally, we found no indication of an increased respondent burden in this subsample in the previous subsection of the appendix. This serves as another indication that the use of video vignettes is at least feasible. Further research should try to examine whether the quality of surveys employing video vignettes differs from text-only versions when the total survey time is held constant.[19] Comparing the behavior of the text-only and deviator subsamples, we conclude that these subsamples' survey behavior is to a large degree comparable. This appendix raises the concern that possibly not all of the respondents in the

[19] This finding is subject to the limitation that our assignment into the text-only and video version of the survey is hampered by the deviator subsample. Since respondents self-selected into this group, the video-only subsample is no random subsample.

deviator subsample truly experienced problems with the video version of our survey. Instead some of these respondents may have switched to the text version to save time. The fact that the deviator subsample's survey behavior is not considerably worse than the behavior of the text-only subsample suggests that the prevalence of respondents skipping out of the video version to save time did not result in a tremendously worse survey quality in the deviator group.

Chapter 3

Respondents' decision making: Can people assess what influences and helps their decision making process?

3.1 Introduction

Patient decision making nowadays is characterized by an increased role of the patient. The idea is that the patient should decide on the treatment option whereas the physician's task is to inform and consult (Braddock et al. (1999) and Gurmankin et al. (2002)). Information and consultation can however also be sought from other sources, such as decision aids or health related webpages. Furthermore, the kind of information that is sought may also vary from professional advice, data on effectiveness of procedures and incidence of side effects to narratives of other patients describing their subjective experience with procedures. For policy makers, it is important to know how patients weigh information obtained from different sources and of different kinds and whether this information benefits the decision making process. The fundamental question is how to obtain the necessary data to answer these questions regarding the decision making process in a reliable way.

One possible approach is to directly ask patients to what degree the different forms of information influenced their decisions and how helpful the different forms of information were. It is however a well-established fact in the psychological literature that when people are faced with decisions, they are susceptible to a number of biases such as bounded rationality

(Simon (1972)). Furthermore, people are unaware of these biases in a number of cases. Therefore, the question arises whether people can assess what influenced their decisions and what helped them decide without bias. This research constitutes a first attempt to provide an answer to this question. This study is based upon a discrete choice experiment (DCE) that examines the influence of several factors on respondents' decisions and the contribution of these factors to respondents' confidence in their decisions. These findings are compared with respondent's self-assessed main factor of influence and helpfulness as reported in a follow-up questionnaire.[1]

Our approach to this question is of an explorative nature and tries to assess whether respondents' assessment of the influence and helpfulness of factors can be matched what we observed in our DCE. If this is the case, asking patients after medical decisions about the factors that influenced their decisions and the helpfulness of these factors could be used in diverse research settings. Thus, if patients for example report that reviewing statistical information concerning the effectiveness of certain medical procedures is very influential and helpful in their decision making process, this finding could then be used to promote the use of this kind of information in medical counseling.

We find that the self-assessment of the influence of four factors can to a large degree be matched with what we observed in our DCE. The self-assessment of the helpfulness of two factors can only partially be matched. Therefore, respondents' perception of what is most influential and most helpful in their decision making may not be bias free, however respondents appear to have a notion at least of what influenced their decision making in our discrete choice experiment.

This paper proceeds as follows: We first describe the study design and sample (Section2), then provide descriptive statistics of our sample and the follow-up questions (Section3). Finally, the empirical results are presented in which we compare respondents' perceived most influential and most helpful factor with the observed influences of this factor in our DCE and its contribution to respondents' confidence in decision making (Section 4). Section 5 concludes and discusses limitations.

[1] The full results of this discrete choice experiment are presented in Caro et al. (2012) and can also be found in chapter 2 of this dissertation.

3.2 Study design

This study links the results of a discrete choice experiment that examines respondents' behavior in a medical decision to respondents' answers to a follow-up questionnaire relating to the same experiment.[2]

The discrete choice experiment we conducted aims to disentangle factors that influence patients' decision making regarding full knee replacement surgery. We generated a hypothetical scenario in which survey participants were presented with three randomly chosen hypothetical persons, the so called vignette persons. All these vignette persons suffer from knee osteoarthritis and have to decide whether they want to pursue full knee replacement surgery or not.

At first, each vignette person described his or her personal history with knee osteoarthritis as well as the treatments that have been tried so far and failed. The vignette person then proceeded to describe his or her personal situation regarding the experienced pain level (high or moderate), the employment status (employed or retired) as well as the opportunity costs that would be encountered if he or she would have the surgery (high or low).

After this description of a vignette's personal situation, up to three health related factors were presented. First, an orthopedic surgeon assessed the vignette person's situation and either recommended surgery or stressed the possibility of continuing alternative treatments leaving the decision for or against surgery up to the vignette person. Then additional information parts were presented for random subsamples of vignette persons. These factors were specialist's second opinions and person specific surgery outcome forecasts. Respondents could then choose whether they want to review a specialist's second opinion for a specific vignette person and/or whether they want to review the third vignette person's outcome forecast.[3] The specialist's second opinion was either a strong recommendation that clearly

[2] This a brief description of the experiment we conducted that focuses mainly on the aspects needed for this paper's purpose. For a detailed description of discrete choice experiments in general and our full experimental setup, see chapter 2 of this dissertation.

[3] At the third vignette person, respondents were also shown one patient testimonial and offered the option to see a second patient testimonial. Since testimonials were at the last choice task and qualitative interviews revealed a minor role of these, the follow up questionnaire focused on the information parts that were shown at every choice task.

identified surgery now as the best option or a statement of substantial reservation against surgery which stressed equal benefits of continuing non-surgical therapies.

The person specific outcome forecasts were introduced in the following way:

> [Name] also sought further information on the internet. The homepage of a nonprofit knee osteoarthritis patient aid group offers a tool that predicts likely surgery outcomes. This application uses [Name's] personal and health information and compares these with a large sample of full knee replacement surgery patients to predict [Name's] chances of a successful surgery outcome. The development of this tool was funded by the U.S. Department of Health and Human Services.

Respondents were then shown a screen that informed them that the vignette person's chance of a successful surgery outcome was either "above average" or "below average". The chances of success were described in one of three ways: only in verbal terms, in numeric terms including specific percentages, or in a graphic format that used a bar chart to represent the vignette person's chances of a successful surgery outcome relative to the average chances of a successful surgery outcome.

For each of these vignette persons, the survey participants had to decide whether they would recommend surgery for this person or not and how confident they are this decision. The wording of the decision questions was the following:

> 1. Do you recommend that Name have full knee replacement surgery now?
> Yes
> No
>
> 2. How confident are you of your recommendation?
> Please use this scale where 1 means you are not at all confident and 10 means you are absolutely confident.

Not at all									Absolutely
( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
1	2	3	4	5	6	7	8	9	10

Once the actual DCE was completed, respondents filled out a follow-up questionnaire regarding their perceptions of the factors we presented them with. This paper uses the questions that establish which factor respondents considered the most influential in their decisions and which factor they considered to have been most helpful in their decision making process. To find out whether respondents considered the personal situation or one of the health related factors the most influential factor when deciding, the following question was asked:

> In general, which of the following most influenced your recommendations for the people considering full knee replacement? (Select only one)
>
> People's personal situation
> Physician's recommendation
> Second opinion
> Probability of successful surgery

The following question was used to elicit whether respondents considered the additional information conveyed in two of the health related factors, the second opinions and the person specific surgery outcome forecasts as most helpful:

> Aside from people's personal situations and the physician's recommendation, which additional information was most helpful to you for making your recommendations? (Select only one)
>
> Second opinion
> Probability of successful surgery
> Not applicable (I did not review any additional information).

3.3 Study sample and descriptive statistics

This study was conducted with members of the RAND American Life Panel, which recruits its members from the general U.S. population and provides all panel members with the means to take part in internet surveys.[4] The sampling base of our study was members of the RAND American Life Panel that were 50 years old or older and had not participated in our pilot study. In total, 2296 ALP members were invited to take the survey. Of these, 1675 took part in the interview and 1622 respondents completed the survey. This results in a response rate of 70.6%. Of the 1622 completed interviews, complete information on the background variables was available for 1616 respondents. These respondents serve as the final sample studied here.

The demographic characteristics of our final sample are summarized in Table 1. Respondents were fairly evenly represented by men and women and reflected a broad age range: between age 50 to 93.

[4] Further information on the American Life Panel, its composition and attrition can be obtained at the homepage of the RAND American Life Panel: https://mmicdata.rand.org/alp/index.php?page=panel (Rand (2012)).

Variables	Percent
Gender	
Male	43.4
Age (median=59 years)	
50-59	50.1
60-69	33.7
70 or older	16.3
Income	
below 25,000$	23.4
$\geq$ 25,000\$, < 50,000\$	27.3
$\geq$ 50,000\$, < 75,000\$	16.1
above 75,000$	33.2
Living Status	
Married or living with a partner	58.2
Employment Status	
Retired	34.3
Working	45.2
Unemployed, disabled and other	20.4
Education	
High school or less	23.9
At most Bachelor's Degree	59.2
Post graduate	16.9
Ethnicity	
Non-Hispanic white	82.7
Respondents with chronic knee pain	42.1
Respondents with knee osteoarthritis	21.2
Friends/Relatives with knee osteoarthritis	65.8
Friends/Relatives with full knee replacement surgery	52.5

Table 3.1: Descriptive statistics of the study sample

Respondents were well educated and relatively well-off financially. About half of the respondents are still active in the workforce and roughly 50% have an annual income exceeding 50,000$.[5] Furthermore, 42% of respondents have experienced chronic knee pain[6] and 21% are diagnosed with knee osteoarthritis. Additionally most respondents reported having friends

[5] The ALP seeks to be representative of the adult population of the United States. We have to acknowledge however, that in comparison to CPS 2011 data for the population 60 and over, our specific sample is more highly educated and underrepresents non-Hispanic white population members (US-Census (2012)). However, the ALP sample we use is much more generalizable than studies based on student or patient samples. Studies by Chang & Krosnick (2009) and Yeager et al. (2011) examined data quality issues with the ALP and another probability sample in comparison to samples obtained via RDD and non-probability samples. Both conclude that the phone sample and the probability sample show the least bias.

[6] In the relevant age range, this is a reasonable percentage of people with chronic knee pain and similar to a recent data from the 2011 Gallup-Healthways Well-Being Index (Gallup (2012)).

or close relatives that were diagnosed with knee osteoarthritis and/or that have had full knee replacement surgery.

The distribution of the answers given to the follow-up questions is presented in Table 2. 1614 respondents assessed the factor they perceived as the most influential in their decision making. 46% of respondents indicated that the vignette person's personal situation constituted the most important factor of influence in their decision making. In contrast to this, only 13 % chose the surgeon's recommendation as the most influential factor. The second opinions and the outcome forecasts were chosen as the most influential factor by 14%, respectively 27%, of respondents. Thus, by far the largest fraction of respondents considered the personal situation of a vignette as the most influential factor. The second largest fraction considered the outcome forecast as most influential. When it comes to the perceived helpfulness of the second opinions and the outcome forecasts, a similar picture arises. 56% of respondents considered the outcome forecast as the most helpful, whereas only 41% chose the second opinion.

Variables	Percent	No. of respondents
Most influential factor		
Personal situation	46.1	744
Surgeon's recommendation	13.3	214
Second opinion	13.7	221
Outcome forecast	26.9	435
Most helpful factor		
Second opinion	41.2	665
Outcome forecast	56.2	909
Did not review additional information	2.6	42

Table 3.2: Descriptive statistics of the follow-up questions

It follows that most respondents considered the personal situation or the outcome forecast as the most influential factors for their decisions and the information contained in the outcome forecast as the most helpful information in their decision making process.

3.4 Empirical analysis

This section examines the correspondence between the self-assessed most influential factor and the self-assessed most helpful factor with the actual influence of these factors observed in our DCE and the contribution of these factors to respondents' confidence in decision making. Subsection 3.4.1 analyzes whether the self-assessed main factor of influence corresponds with the actual most influential factor observed in our DCE. Subsection 3.4.2 examines whether self-assessed most helpful factor increases respondents' confidence in their decisions most.

The unit of analysis in both subsections is the single vignette observation. All reported p-values in these subsections refer to tests for equality of the strength of the estimated coefficients using either F-tests or Wald-tests.

3.4.1 Correspondence of most influential factors

The estimation results presented in this subsection are linear probability models estimated on the binary dependent variable "surgery recommendation". This variable takes the value 1 if the respondent recommended surgery for the respective vignette person and 0 otherwise.

The specification of all estimated models is the following: We include the effects of the levels of the personal situation of the vignette person as dummy variables (high pain, retired and high opportunity costs) and the effect of the level of the surgeon's recommendation (positive surgeon's recommendation). Also included are the effects of the second opinion levels (recommending or substantial reservations second opinion) controlling for whether the option was chosen or not (reference category: observations where no second opinion was offered) and the effects of the outcome forecast levels (forecast above average or forecast below average) controlling for format effects (reference category: observations where no outcome forecast was offered). Furthermore, we include dummy variables for each vignette person and choice order effects that capture whether the surgery recommendation was the first, second or third decision a respondent made. Additionally, we control for the following respondent characteristics: gender, age, marital status, education (categories: at most bachelor and postgraduate, reference category: high school or less), respondent's labor force status (employed or retired, reference category: all forms of unemployment), ethnicity (Non-Hispanic white, reference category: all other ethnicities), household income (25.000$ to 49.999$, 50.000$ to 74.999$ or above 75.000$, reference category: below 25.000$), whether the respondent has chronic knee pain, was diagnosed with knee osteoarthritis, whether the respondent has friends/relatives that were diagnosed with knee osteoarthritis and whether the respondent has friends/relatives that have had full knee replacement surgery. All models use robust standard errors to take into account that each respondent was observed up to three times.[7] With three instances of item non response, our total vignette sample used for estimation consists of 4845 observations.

The analysis presented here is guided by the following logic: If respondents' perceptions of the most influential factor reflect their actual behavior, we should observe that the influences of different factors vary according to respondents' self-assessment of what constitutes the most influential factor. Thus, for example for the subsample that stated the outcome forecasts were the most influential factor, we should observe a larger impact of these than in the subsample that did not consider the forecasts as the most influential factor. Furthermore, within the subsample that considered the forecasts as the most influential, the estimated impact of these should be stronger than the impacts of other factors. Therefore, our analysis proceeds along the following lines for each assessed factor: First, the full sample is split up into two subsamples according to respondents' assessment of whether this factor was the most influential or not. Then, the estimation results are compared between the subsample that assessed a specific factor as most influential and the subsample that did not to find out whether this specific factor's influence is indeed larger in the first subsample than in the latter. Finally, only the estimation results within the subsample that considered this factor as most influential are examined to find out whether the influence of this one factor is really the most important influence within that subsample. If both of these examinations are affirmative, this is interpreted as an indication that these respondents correctly assessed their most influential factor.

[7] The same analysis was also conducted with standard errors clustered at the individual level. There was no substantial difference in the estimated standard errors.

All estimated models are reported in Table 3 and only vary with respect to the sample used in estimation. The first column presents the estimation results obtained when we use the full sample, the following columns then split up the sample into two mutually exclusive subsamples depending upon the respondent's assessment of whether a specific factor (for example the personal situation) is most influential or not (Table 3, columns 2 and 3). Table 3 only reports the estimated coefficients that are of interest for this research.[8]

When we first look at the estimation results of the entire sample in Table 3, we observe that the estimated coefficients on all personal factors as well as the health related factors are highly significant. In the overall sample, the estimated effect of the "substantial reservations" second opinion is the largest followed by the effect of high pain and a positive surgeon's recommendation. These estimated coefficients are not significantly different from each other. However, all of these effects are significantly stronger than the effects of high opportunity cost ($P<0.01$), a retired vignette person ($P<0.01$), a "recommending" second opinion ($P<0.01$), a below average outcome forecast ($P<0.01$) and an above average forecast ($P<0.01$). These relationships however change when we turn to the estimation results of the different subsamples.

At first we examine the patterns of influences when the full sample is split up according to whether the personal factors were assessed as most influential or not. Column 2 presents the estimation results of the subsample that considered the personal factors as most influential, column 3 presents the results of the subsample that did not consider the personal factors as most influential. By comparing the estimation results between these two subsamples, a clear pattern emerges. The subsample that considered the personal factors as most influential has higher parameter estimates for high pain ($P<0.01$) and high opportunity costs ($P<0.01$) and lower parameter estimates for the "recommending" second opinion ($P<0.05$), the "substantial reservations" second opinion ($P<0.05$) and the above average outcome forecast ($P<0.1$) than the other subsample. In the subsample that stated personal factors as most influential, the effect of high pain significantly dominates the effects of all health related factors of influence (P at least <0.05). The effect of high opportunity costs

[8] Two respondents did not provide an assessment of their most influential factor and their six vignette observations are therefore not included in the models with the split samples.

is also significantly stronger than the effects of a "recommending" second opinion ($P<0.01$) and of an above average outcome forecast ($P<0.05$). It follows that in this subsample, high pain is the most influential factor, and this factor is also more influential than the surgeon's recommendation, the second opinions and the outcome forecasts. Thus, in this subsample the personal factors exhibit a stronger influence than in the other subsample and the influence one personal factor within this subsample is stronger than the influence of all other factors.

Columns 4 and 5 then present the estimation results when we split up the sample according to whether the surgeon's recommendation was assessed as the most influential factor (column 4) or not (column 5). The subsample that stated that the surgeon's recommendation was the most important factor of influence exhibits a stronger influence of the positive surgeon's recommendation ($P<0.01$) and the "substantial reservations" second opinion than the other subsample. Furthermore, the impact of high opportunity costs and a below average outcome forecast is lower ($P<0.01$). Within the subsample that perceived the surgeon's recommendation as the most influential factor, the parameter estimate of the surgeon's recommendation is among the largest factors of influence, but can't be distinguished statistically from the parameter estimate of the "substantial reservations" second opinion. Thus, this subsample exhibits a stronger effect of the surgeon's recommendation than the other subsample but within the subsample the effect of the surgeon's recommendation is not clearly stronger than the effects of all other factors.

Sample	Full	Personal factors		Surgeon's recommend.		Second opinion		Outcome forecast	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Explanatory variables									
high pain	0.199**	0.256**	0.148**	0.203**	0.201**	0.070*	0.217**	0.148**	0.212**
	(0.013)	(0.019)	(0.017)	(0.033)	(0.014)	(0.031)	(0.014)	(0.023)	(0.015)
retired	0.052**	0.048*	0.043*	0.026	0.053**	0.061	0.049**	0.045	0.053**
	(0.013)	(0.019)	(0.017)	(0.034)	(0.014)	(0.031)	(0.014)	(0.023)	(0.015)
high opportunity costs	-0.100**	-0.162**	-0.040*	0.005	-0.115**	-0.067*	-0.104**	-0.049*	-0.121**
	(0.013)	(0.019)	(0.017)	(0.033)	(0.014)	(0.031)	(0.014)	(0.023)	(0.015)
positive surgeon's recommendation	0.198**	0.172**	0.211**	0.335**	0.176**	0.116**	0.206**	0.184**	0.202**
	(0.014)	(0.022)	(0.019)	(0.034)	(0.015)	(0.036)	(0.015)	(0.026)	(0.017)
second opinion was not chosen	0.046	0.071	0.038	0.007	0.054	0.180*	0.030	-0.005	0.075*
	(0.027)	(0.038)	(0.037)	(0.065)	(0.029)	(0.082)	(0.028)	(0.054)	(0.031)
strong recommendation second opinion	0.118**	0.078**	0.154**	0.103*	0.122**	0.380**	0.081**	0.073*	0.135**
	(0.017)	(0.025)	(0.023)	(0.046)	(0.018)	(0.047)	(0.018)	(0.030)	(0.020)
substantial reservations second opinion	-0.230**	-0.189**	-0.265**	-0.318**	-0.215**	-0.299**	-0.213**	-0.188**	-0.239**
	(0.017)	(0.025)	(0.023)	(0.046)	(0.018)	(0.046)	(0.018)	(0.031)	(0.020)
above average outcome probability	0.121**	0.085**	0.150**	0.109*	0.114**	0.011	0.133**	0.204**	0.086**
	(0.020)	(0.029)	(0.026)	(0.049)	(0.021)	(0.048)	(0.021)	(0.038)	(0.023)
below average outcome probability	-0.126**	-0.110**	-0.144**	0.006	-0.152**	-0.058	-0.137**	-0.270**	-0.076**
	(0.020)	(0.029)	(0.026)	(0.051)	(0.021)	(0.050)	(0.021)	(0.037)	(0.023)
N	4845	2230	2609	642	4197	663	4176	1304	3535
adj. R2	0.629	0.598	0.670	0.730	0.621	0.695	0.633	0.704	0.616

All models are linear probability models on the binary dependent variable "surgery recommendation" and control for vignette person fixed effects, vignette order effects and respondent's personal and health characteristics. The stars behind the parameter estimates indicate significance levels. *prob< .05; **prob< .01. Robust standard errors are displayed in brackets.

Table 3.3: Estimation results conditional on the assessment of the most influential factor

Columns 6 and 7 report the estimation results of the subsample that considered the second opinions as most influential (column 6) and the subsample that did not (column 7). Here, we again observe a clear difference between the estimation results. The impacts of both levels of the second opinion are significantly larger in the subsample that considered the second opinions as the most influential factor than in the other subsample (P at least <0.1). Furthermore, in the subsample that reported the second opinions as their most influential factor, the parameter estimates of high pain, the positive surgeon's recommendation as well as of the above average outcome forecast are significantly lower than in the other subsample (P at least <0.05). Additionally, within the subsample that perceived the second opinions as their most influential factor, the estimated coefficients on the "recommending" and "substantial reservations" second opinions are by far stronger than the estimated coefficients of all other factors (P at least <0.01). Thus, in this subsample the second opinion levels exhibit stronger influences than in the other subsample and the influences of the second opinion levels within this subsample are stronger than the influences of all other factors.

When we compare the estimation results of the subsample that assessed the outcome forecasts as the most influential factor (column 8) to the subsample that did not (column 9), we again observe a clear difference in the patterns of influence. The subsample that assessed the outcome forecast as the most influential factor shows a stronger influence of the outcome forecast levels than the other subsample (P at least <0.01). It also differs with respect to the effects of the other factors. The effects of high pain, high opportunity costs and of a "recommending" second opinion are significantly lower (P<0.05, P<0.01 and P<0.1). When we examine the parameter estimates of the different factors within the subsample that considered the outcome forecasts as the most influential factor, the outcome forecasts are among the strongest influences, but do not completely exhibit a stronger effect than all other factors. The effects of both outcome forecast levels are significantly stronger than the effect of high opportunity costs (P <0.01) and the "recommending" second opinion (P <0.01). Furthermore, the effect of the below average outcome forecast is also significantly stronger than the effect of high pain (P<0.05), a positive surgeon's recommendation (P<0.1) and a "substantial reservations" second opinion. There is no significant difference between the effect of an above average outcome forecast and the effects of the positive surgeon's recommendation, high pain and a ”substantial reservations” second opinion. Thus, this subsample exhibits stronger effects of the outcome forecast levels than the other subsample but within the subsample the effects of the outcome forecast levels are not entirely stronger than all other effects.

In sum, we found that the subsamples that considered either the personal factors or the second opinions as the most influential factors completely assessed the role of these factors in their decision making process correctly. The decision making process of subsamples that reported either that the surgeon's recommendation or the outcome forecast were most influential also differ from the subsamples that did not assess these factors as the most influential in the expected direction. Within these subsamples, the dominance of the self-assessed most influential factor is however not as prevalent as in the other two subsamples.

3.4.2 Correspondence of most helpful factors

The models reported in this subsection use the dependent variable ”confidence in decision” which is the confidence level that a respondent reported regarding his or her decision to recommend surgery for each vignette recommendation. All models are ordinary least squares models using robust standard errors to take into account that each respondent was observed up to three times.[9] The estimated specification is again the same for all three presented models and only differs in the modeling of the outcome forecasts from the models presented in the previous subsection. These outcome forecasts were presented in two levels (above or below average) and in three different formats (verbal, numeric or graphic). Thus, each vignette observation could be displayed with one of six possible level/format combinations. These combinations are captured by six dummy variables that measure the contribution of an above average outcome forecast and a below average outcome forecast in each format to the confidence levels in comparison to observations where no outcome forecast was displayed.

In this subsection, the contribution of the second opinion levels and the outcome forecast levels to respondents' confidence in their decisions is used as a measure for the helpfulness of

[9] The same analysis was also conducted with standard errors clustered at the individual level. There was no substantial difference in the estimated standard errors.

these factors in respondents' decision making. If the contribution of a factor to confidence levels is positive, this speaks for the helpfulness of this factor in respondents' decision making process. Conversely, if the contribution is negative, this speaks against the helpfulness of this factor. Therefore, if respondents' perception of the helpfulness of the second opinions or the outcome forecasts holds true, we should observe that the contribution of these factors to respondents' confidence in decision making varies according to respondents' self-assessment of what is the most helpful factor. Similar to the previous subsection, we again split up the total sample into subsamples according to whether the second opinion or the outcome forecast was considered as most helpful in the decision making process. Again, we examine for each subsample whether the contribution of the self-assessed most helpful factor is more positive/less negative in this subsample than in the other and whether the contribution of this factor has the most positive/ least negative effect on confidence levels within this subsample. If both of these examinations are affirmative, this is taken as an indication that the respective subsample correctly assessed the helpfulness of this factor.

The estimation results are presented in Table 4 which essentially follows the same setup as Table 3. Column 1 reports the estimation result for the full sample, column 2 for the subsample that considered the second opinions as most helpful in the decision making process and column 3 for the subsample that considered the outcome forecasts as most helpful.[10]

When we first look at the estimation results of the full sample (column 1), we observe that if a second opinion was offered, but not viewed, the confidence level tends to be lower than for observations where no second opinion was offered. Furthermore, viewing a second opinion of any level tends to increase confidence levels. These effects are however insignificant. For the outcome forecasts, the picture is mixed. While above average outcome forecasts in a numeric format significantly increase confidence levels, a below average outcome forecast significantly decreases these in all three formats. These patterns again change when we examine the estimation results of the subsamples that either considered the second opinions as most helpful or the outcome forecasts.

[10] Note that it was possible that a respondent did not view any of these two features in the course of our DCE. These respondents were able to state this at the relevant question and the 42 respondents that reported not having seen both the second opinions and the outcome forecasts were subsequently dropped from analysis. This leaves us with a total sample size of 4719 vignette observations.

Comparing the estimation results of these two subsamples, we can observe that the contribution of the second opinions to confidence levels tends to be higher in the subsample that considered the second opinions most helpful (column 2) than in the subsample that considered forecasts as most helpful (column 3). Additionally, the positive contributions of the above average forecasts to confidence levels tend to be lower and the negative contributions of below average forecasts to respondents' confidence levels tend to be less pronounced. However, the only significant parameter difference is observed for the positive effect of viewing an above average outcome forecast in verbal format. This effect is weaker in the subsample that considers the second opinions as most helpful ($P<0.05$).

Within the subsample that considered the second opinions most helpful, the effects of both levels of the second opinions are significantly more positive the effect of a below average forecast that was displayed in verbal format ($P<0.1$). The estimated coefficients of both second opinions are however not significantly more positive than the effects of the other outcome forecast versions. Thus, while the positive contribution to confidence levels tends to be stronger in this subsample than in the other, within this subsample the contribution of the second opinion levels to respondents' confidence is not the most positive.

In the subsample that considered the outcome forecasts as most helpful, only the positive effect of viewing an above average outcome forecast in numeric format is significantly more positive than the effects of a ”recommending” or ”substantial reservations” second opinion (P at least < 0.05). On the other hand, the negative effects of viewing a below average forecast in verbal or numeric format are significantly more negative than the effect of viewing a ”substantial reservations” or ”recommending” second opinion (P at least <0.05). Therefore, we neither find a more positive contribution of the outcome forecast levels to respondents' confidence in their decisions than in the other subsample, nor can the contribution of the outcome forecast levels within this subsample be considered to be the most positive/ least negative.

In total, we found that both subsamples were not completely able to assess the most helpful factor. The second opinions tend to increase confidence levels in the subsample that considered these the most helpful factor, but the levels of the second opinions are not

Sample	Full (1)	Second opinion (2)	Outcome forecast (3)
Explanatory variables			
second opinion was not chosen	-0.170	-0.041	-0.251
	(0.112)	(0.192)	(0.139)
strong recommendation second opinion	0.072	0.120	0.039
	(0.062)	(0.102)	(0.079)
substantial reservations second opinion	0.047	0.133	-0.013
	(0.063)	(0.099)	(0.083)
above average outcome probability, verbal	0.039	-0.167	0.216
	(0.087)	(0.138)	(0.111)
above average outcome probability, numeric	0.290**	0.150	0.385**
	(0.080)	(0.123)	(0.107)
above average outcome probability, graphic	0.120	0.062	0.167
	(0.082)	(0.124)	(0.109)
below average outcome probability, verbal	-0.407**	-0.406**	-0.387**
	(0.086)	(0.132)	(0.115)
below average outcome probability, numeric	-0.278**	-0.170	-0.331**
	(0.083)	(0.135)	(0.107)
below average outcome probability, graphic	-0.230**	-0.220	-0.230*
	(0.081)	(0.117)	(0.110)
N	4719	1993	2726
adj. R2	0.958	0.958	0.959

All models are ordinary least squares models estimated on the dependent variable "confidence in decision" and control for vignette person fixed effects, vignette order effects and respondent's personal and health characteristics. The stars behind the parameter estimates indicate significance levels. *prob< .05; **prob< .01. Robust standard errors are displayed in brackets.

Table 3.4: Estimation results conditional on the assessment of the most helpful factor

the most positive contributors to these respondents' confidence levels. The impact of the outcome forecasts on the other hand is not significantly more positive in the subsample that considered them to be the most helpful than in the subsample that did not. In fact, the below average outcome forecasts exhibit the strongest negative effects on confidence levels in the subsample that considered them to most helpful.

3.5 Conclusion and discussion

This paper used the results of a survey in which respondents of the RAND American Life Panel participated in a discrete choice experiment and completed a follow-up questionnaire relating to the DCE. In the discrete choice experiment, respondents had to decide whether or not to recommend surgery for three vignette persons that were presented with varying levels of personal situations and additional health related factors. Furthermore, they had to indicate at each of the three choice tasks how confident they were in their decision. Following the DCE, the follow-up questionnaire was administered in which we asked respondents to identify whether personal situations or one of the health related factors was their most important factor of influence in decision making and which of two health related factors (second opinion and person-specific surgery outcome forecast) was most helpful in their decision making process. The answers to these questions were then compared with these factors' influences on the decision to recommend surgery and with these factors' contributions to respondents' confidence in their decision making in our DCE.

This was done by splitting up the sample into subsamples of respondents according to respondents' assessment of the most influential or most helpful factor. The estimation results of these subsamples were then compared to find out whether the estimated coefficients differ between the subsamples in a way that supports these respondents' self-assessments. Furthermore, we examined whether the self-assessed most influential factor in a subsample is the most influential factor for these respondents in our discrete choice experiment and whether the self-assessed most helpful factor contributes most to these respondents' confidence levels in their decisions.

In the analysis of the most influential factors, we found that the subsamples that perceived the personal factors and the second opinions as the most influential factors completely assessed their decision making behavior correctly in retrospective. These two subsamples exhibit a stronger influence of these factors than the subsamples that did not consider these as the most influential. Additionally, in these two subsamples, these factors also exhibit the strongest influence on the decision to recommend surgery. The subsample that perceived the surgeon's recommendation as the most influential factor also shows a stronger impact of this factor than the subsample that did not consider the surgeon's recommendation as most influential. However, within this subsample, the effect of the "substantial reservations" second opinion is equally strong as the effect of the surgeon's recommendation. The same conclusion applies to the subsample that considered the person specific outcome forecasts as the most influential factor. The estimated effects for the outcome forecasts are stronger than in the subsample that did not consider them as most influential. However, here too the "substantial reservations" second opinion is equally strong as the above average outcome forecast. In sum however, respondents showed a surprisingly large degree of awareness when it comes to assessing the most important influences on their decisions in our discrete choice experiment.

In the analysis of the most helpful factors, we found that the second opinions tend to increase the confidence respondents have in their decisions for both levels in the subsample that considered them the most helpful. Additionally, this positive contribution of the second opinions to confidence levels is a little more pronounced in the subsample that considered them as the most helpful factor than in the subsample that did not. Concerning the person specific outcome forecasts, we neither found that the contribution of these to respondent's confidence in their decisions is clearly more positive in the subsample that considered them most helpful than in the subsample that did not nor could we establish that the contribution of these outcome forecasts in the subsample that considered them the most helpful is the most positive/least negative. Therefore, this subsample miss-assessed the helpfulness of the outcome forecasts in their decision making process.

The overall pattern that emerges is that respondents showed a large degree of insight in the factors that most influence their decision making and a lower degree of insight in which factors most help their decision making in our DCE. Regarding the factors of influence, the strong effect of a "substantial reservations" second opinion was underappreciated a little by two subgroups. At least for the group that considered the surgeon's recommendation as the most influential factor this could possibly be explained by the fact that the second opinions were also described as being provided by a doctor. Thus, respondents who stated that the surgeon's recommendation was the most influential factor may have summed up these two factors into one.

Concerning the helpfulness of factors, the outcome forecasts were considered more helpful than what we observed in our DCE. One possible explanation for this finding could be that this is due to recency effects (Baddeley (1990)) since the outcome forecasts were displayed at the end of each vignette. Additionally, they were extensively described as being based upon a large sample of knee replacement surgery patients' outcomes and as being funded by the

U.S. Department of Health and Human Services. Thus, this finding could partially be due to an availability bias in the sense that this information was more vividly present at the moment we asked respondents about the most helpful factor and this facilitated retrieval (Tversky & Kahneman (1973)). This may have been promoted for some respondents by seeing the outcome forecasts in the graphic format. In this case, the forecast filled an entire screen. This format was thus more distinctive and this may have increased retrieval of this factor at the moment of answering the question relating to the helpfulness of factors (Tourangeau et al. (2000)). Additionally, the role of doctors in medical decisions may be suspect to a number of criticisms especially when it comes to their motivation for recommending certain procedures where they benefit directly in financial terms if the procedure is undertaken (Cunningham (2009)). In the aftermath of this experiment, respondents may have considered this ambiguity and thus refrained from reporting that the second opinion by the specialist was most helpful.

Overall however, we were positively surprised by the degree of insight respondents had in their decision making process in this experiment. Most impressing is that we observe for all subsamples an increased role of the self-assessed most influential factor whereas the role of the other factors is reduced in these respondents' decisions. This is a first indication that respondents can at least self-assess what influenced their decisions. Further research should examine whether what we found regarding respondent's insight into their behavior in our discrete choice experiment also holds up when it comes to assessing real decisions.

Limitations

The results of this study have to be seen in the light of the following limitations. First, the use of a stated preference approaches poses a threat to external validity (Diamond & Hausman (1994), Hensher (2010), Mark & Swait (2004) and Harrison (2006)).[11] However, we do believe that the choices we observed approximate the behavior of people in real situations. The second limitation concerns our measurement of the helpfulness of factors in the decision making. The reported confidence levels can clearly only serve as a proxy for real helpfulness. Finally, we have to acknowledge that the variables we used to generate the subsamples

[11] An extended discussion of this limitation can is presented in chapter 2 of this dissertation.

are endogenous since they were reported by respondents themselves after the experiment. However, the purpose of this study is not to establish consistent parameter estimates for the dimensions of our design, but rather to check for differences in the patterns of influences in the estimation results of the different subsamples.

Chapter 4

Question order and intentional priming in the anchoring vignettes methodology.

4.1 Introduction

The use of self-assessments in statistical analysis can lead to biased results if respondents from different countries or subgroups of people exhibit a differential use of response scales. This is commonly referred to as Differential Item Functioning (DIF). The result is that the given answers are incomparable and using them in statistical analysis may lead to biased estimation results. The anchoring vignette method promises to ameliorate the extent of this bias. Originally introduced by King et al. (2004), this method can directly correct for DIF. The method uses respondents' self-assessments of their personal situation and respondents' assessment of the situation of hypothetical persons, the so called anchoring vignettes to generate an adjusted measure that corrects for this bias.

Since its original development, anchoring vignettes have found their way into many major surveys and panels, such as the Dutch CentERpanel, the Survey of Health, Aging and Retirement in Europe (SHARE), the RAND American Life Panel (ALP) and the Health and Retirement Study (HRS) to name a few. Its applications span a wide range of topics from health (see Peracchi & Rossetti (2012) and Bago d'Uva et al. (2011a)), political efficacy (see King et al. (2004)) and life satisfaction (see Kristensen & Johansson (2008) and Kapteyn et al. (2012)), to work disability (see Kapteyn et al. (2007) and Angelini et al. (2011)) and public institutions (Rice et al. (2010)).

Despite its growing use in empirical work, some methodological issues are still unsettled. Recently, there is a number of studies examining the validity of the underlying assumptions of this method (Soest et al. (2011), Datta Gupta et al. (2010), Bago d'Uva et al. (2011b), Vonkova & Hullegie (2011), Jürges & Winter (2011), and Kapteyn et al. (2011)). All of these papers conclude on cautious notes and Kapteyn et al. (2011) stress the need for pre-survey experimentation to validate the design of vignettes.

The purpose of this paper is to expand this methodological strand of the anchoring vignette literature. The central question that will be dealt with is the optimal order of self-assessment and anchoring vignette questions. The standard procedure up to now is to ask the self-assessment question first followed by the anchoring vignette part. Hopkins & King (2010) find in a survey experiment that placing the vignettes before the self-assessment improves the DIF-correction when applied to the political efficacy domain. Thus, they propose a reversal of the question administration order to make use of what they call intentional priming. This research further examines the effects of intentional priming in an anchoring vignette framework as proposed by Hopkins & King (2010). Their reversal of question administration order was applied to a new anchoring vignette domain.

We conducted a survey experiment in which we use a set of anchoring vignettes in the domain of satisfaction with living circumstances. Respondents were randomly assigned to two different survey versions. The respondents of the control group received the standard administration question order where they first had to answer two self-assessment questions and were then asked to rate several anchoring vignettes. The treatment group started out with rating the vignettes and was then asked to assess its own situation. In a supplementary questionnaire, a number of respondents' living situation characteristics were surveyed to generate an objective validation dataset. These validation variables were used to examine whether the reversal of question administration order improves the DIF-correction.

This paper proceeds as follows. Section two provides a short description of the anchoring vignette method and the DIF-correction procedure used in this paper. Section three outlines the motivation and the experimental design. Section four describes the study sample and presents the univariate analysis of the self-assessments, the vignette ratings and the adjusted measures. This is then followed by a critical comparison of the estimation results obtained by using the adjusted and unadjusted measures obtained from the two survey versions. The last section discusses the obtained results and concludes.

4.2 Anchoring vignettes

Anchoring vignettes are a method developed by King et al. (2004) to correct for biases that arise if people from different subgroups of a population or countries use a different mapping from their situation, for example their health status, to the answer scale. In other words, people that are equal with respect to their situation differ in the probability of choosing the same point on the answer scale. This is commonly referred to as Differential Item Functioning (DIF) and results in answers that are not comparable. Using these answers in a regression analysis may lead to biased results.[1]

The key idea of the anchoring vignette method is to ask respondents in addition to a self-assessment of their own situation also for an assessment of the situation of hypothetical persons, the anchoring vignettes. In sum, this method generates two types of responses. First, the self-assessment responses that differ across respondents both with respect to the actual situation of the respondent as well as the respondent's use of the response scale. Second, the vignette responses that only differ with respect to the use of the response scale as the actual level of the vignette person's situation is the same for all respondents. Relating (”subtracting”) the responses from the self-assessment to the vignette assessments enables the researcher to create an adjusted measure of the respondent's situation. If the assumptions of response consistency and vignette equivalence hold, this adjusted measure is free of DIF (King et al. (2004)).

By response consistency, King et al. (2004) understand that respondents use the response scale in both question types in an identical way. This means the mapping used from the

[1] As an example of the potential gravity of the bias, consider the analysis presented in King et al. (2004). Respondents in China and Mexico were asked about how much say they had in government affairs. Using only the original answers to the self-assessment question, the result was that respondents in China reported significantly higher levels of political efficacy than respondents in Mexico. Using the anchoring vignette ratings to correct for the differential use of response scales, this result is reversed.

level of the situation to the response category has to be the same irrespective of whether the respondent assesses its own situation or the situation of the hypothetical person. Only then, DIF is present in the same way in both question types and the adjusted measure will be bias free.

The second key assumption King et al. (2004) describe, vignette equivalence, demands that the concept presented in the vignette description must be understood by all respondents in the same way. That means, while it is no problem if respondents differ in how they assess a vignette, they all have to understand that this vignette represents a level of a situation on the continuum of all possible levels.

To develop an intuition for how the DIF-correction works, the nonparametric approach described in King et al. (2004), that is also employed in this paper, is now briefly summarized.

Let j (j=1,2,...,J) be the number of vignettes presented to respondent i (i= 1, 2,..., n). Then z_{ij} is the response of person i to vignette j and y_i is the response of person i to the self-assessment question. Now the answers to the self-assessment question can be recoded to create a DIF free, adjusted measure C_i by relating the self-assessment answer to the vignette answers in the following way:

$$\mathrm{C}_i = \begin{cases} 1 & \text{if } \mathrm{y}_i < \mathrm{z}_{i1} \\ 2 & \text{if } \mathrm{y}_i = \mathrm{z}_{i1} \\ 3 & \text{if } \mathrm{z}_{i1} < \mathrm{y}_i < \mathrm{z}_{i2} \\ \ldots & \\ 2J+1 & \text{if } \mathrm{y}_i > \mathrm{z}_{iJ} \end{cases}$$

Thus, a person gets the best category, if it assesses its own situation better than the situation of the best vignette (j=J) and it is assigned to the worst category, if it assesses itself worse than the situation of the worst vignette (j=1). The values in between are assigned depending on the relative position of the respondent's self-assessment to the assessment of the remaining vignettes. This adjusted measure can then be used with standard methods for categorical variables (King et al. (2004)).

4.3 Motivation and experimental design

The experiment that is basis for this paper addresses the issue which question administration order works best for creating the adjusted measure. The explore this issue, the anchoring vignette method is applied to the context of satisfaction with living circumstances.

Hopkins & King (2010) hypothesize that worries about priming or other question order effects promoted the idea that the self-assessment question has to be asked before the vignette ratings are eluded. However, this intuitive approach has been challenged recently. Buckley (2008) calls for a complete randomization of the anchoring vignettes as well as of the order in which the self-assessment question and the anchoring vignettes are presented as a means to reduce question order biases. Hopkins & King (2010) conclude after a large scale survey experiment that question order effects in the context of anchoring vignettes can be a useful tool and should be exploited by asking the self-assessment question after the vignette task is completed.

Their argument is that the exposure to the vignettes familiarizes the respondents with the underlying research concept and the response scale. In two survey experiments, respondents were randomly assigned to two survey versions. The respondents either completed a survey with the standard question administration order (In this case, they had to answer the self-assessment question first and were then presented with the vignette assessment task) or they received a survey version with the reversed question administration order. Hopkins & King (2010) found that the reversed question administration order almost always increased the correlation between the explanatory variables and the adjusted measure. In some cases, it even shifted the sign of the correlation so that it had the expected direction. Therefore, they argue that the reversed question administration order improved the measurement of the theoretically expected relationship and thus increased the construct validity of the procedure.[2]

4.3.1 Experimental setup and questionnaire description

We conducted a paper and pencil survey experiment in the Melessa laboratory in Munich, as part of another, independent, experiment in March 2010.[3] The topics of the survey were the satisfaction with living conditions in general and with the value for money of the current apartment. These survey topics were chosen primarily because it allowed for the collection of objective measures in a supplementary questionnaire. Furthermore, these topics were more suitable to the respondent pool than the typical anchoring vignette applications since this survey was conducted with the typical lab population that mainly consists of students.

The setup of our survey experiment was the following. Respondents were first handed a vignette survey that randomly varied the question administration order in the same fashion as in the Hopkins & King (2010) experiment. Half the sample randomly received a vignette survey where the self-assessment questions came first and thus constituted the control group. The treatment group received a vignette survey where the vignette task was presented first. Upon completion of the vignette survey, all respondents were handed an identical supplementary questionnaire that elicited a number of objective characteristics of the respondents' current living situation and their apartments.

The vignette survey consisted of two self-assessment questions and of four anchoring vignettes that had to be rated. The self-assessment component asked for a rating on a 5-point likert-scale of the general satisfaction with the living circumstances and the satisfaction of the value for money of the current apartment. Both answer scales ranged from very satisfied (=1) to very dissatisfied (=5).

[2] Construct validity means that supposedly influential variables correlate or correlate more with the theoretical construct to be evaluated in the hypothesized direction (convergent validity) and that supposedly independent variables do not correlate (discriminant validity).

[3] The subjects were recruited randomly using ORSEE (Greiner (2004)).

The vignette component consisted of four hypothetical women of roughly the average age of a typical lab pool respondent.[4] The vignette Anna described the general situation in a living community, with two roommates, a good atmosphere, enough private space and a separate living kitchen. Respondents were then asked to rate Anna's satisfaction with her living situation in general. The three other vignettes gave a description of the main apartment features of three other hypothetical people, Jana, Lena, and Sarah. These three vignettes each described the apartment's size, the number of rooms, the location of the apartment and its base rent. The base rent used in these vignettes was calculated using the 2009 "rent index" for Munich.[5] The reported base rent was calibrated so that Jana's apartment had a very good value for money, Lena's a medium one and Sara's a bad one. The rating questions then asked for an assessment of the vignette person's satisfaction with the value for money of her current apartment. All vignettes were introduced using the same question wording as in the self-assessment component they accompanied and had to be rated on the same 5-point likert-scale.

The supplementary questionnaire consisted of 14 question items that covered a wide range of characteristics of the living situation and the apartment including the dimensions presented in the vignette persons.[6] Furthermore, information about respondent's age, gender, study program and lab experience was collected.

4.4 Results

4.4.1 Summary statistics and univariate analysis

The full dataset used in the analysis contains respondent's personal characteristics as well as the variables generated from the follow up questionnaire, the self-assessment variables,

[4] We chose only women in order to avoid the possibility of gender differences in the rating (see Jürges & Winter (2011)). The names of the vignettes were chosen by picking popular, yet not negatively connotated birth names in the middle of the 1980s according to http://www.beliebte-vornamen.de/3776-1980er-jahre.htm.

[5] The "rent index" is an online tool that allows future residents to calculate the typical base rent of a flat in Munich and can be found at http://www.mietspiegel-muenchen.de/dienst/ms2009.html.

[6] A translation of the original vignette and supplementary questionnaire can be found in Appendix A.

the vignette ratings and the adjusted measures generated from the self-assessments and the vignette ratings. We observe item-nonresponse for 15 respondents so that our final study sample consists of 223 respondents with complete information. Of these, 109 are in the control group and 114 are in the treatment group.

The observed personal characteristics are the respondent's gender (1 if male), age, lab experience and indicators for the respondent's study area (business science, mathematical/natural sciences, social studies, languages, teaching and legal studies). Table 1 summarizes the personal characteristics for the full sample and the treatment and control group as well as the p-values resulting from t-tests for differences in means between the treatment and control group. The 223 respondents are on average 24 years old, are fairly evenly represented by male and female persons and have participated in a number of lab experiments before. Additionally, we observe some variation in respondent's study area. None of the observed differences between the treatment and control groups are statistically significant.

Sample	Full		Treatment		Control		△ in means
Variable	Mean	St.D.	Mean	St.D.	Mean	St.D.	p-value
gender	0.435	0.497	0.456	0.500	0.413	0.495	0.517
age	23.987	3.539	23.947	3.809	24.028	3.250	0.866
experience	4.682	3.301	4.474	3.191	4.899	3.413	0.337
business science	0.269	0.444	0.263	0.442	0.275	0.449	0.840
math/natural sciences	0.229	0.421	0.246	0.432	0.211	0.410	0.541
social studies	0.166	0.373	0.193	0.396	0.138	0.346	0.269
languages	0.117	0.322	0.123	0.330	0.110	0.314	0.769
teaching	0.108	0.311	0.114	0.319	0.101	0.303	0.753
legal studies	0.081	0.273	0.061	0.241	0.101	0.303	0.281

Table 4.1: Summary statistics of respondents' personal characteristics

The validation variables generated from the supplementary questionnaire are: the size of the apartment and the size of the private area (in square meters), the number of rooms, indicators for the living status (alone, in a living community, with partner, with parents), indicators for apartment characteristics (living kitchen, living room, wooden floor, balcony, basement compartment, yard, parking), the condition of the apartment and the house (1 if in very good condition, 3 if in need of renovation), distance to the university (in kilometers), characteristics of the surrounding area (vibrant, neutral or boring), base rent (in Euro), price per square meter (calculated as the base rent divided by the size of the private area), the probability to move out of the current apartment during the remainder of the respondent's studies (in percent) and the self-assessment of the base rent (1 if very appropriate, 10 if very inappropriate) and if the respondent was not living alone also the number of people living with him or her and the atmosphere of the living arrangement (1=very harmonic, 5=very tense).

Table 2 summarizes the validation variables for the full sample and the treatment and control group as well as the p-values resulting from t-tests for differences in means between the treatment and control group. The 223 respondents on average live in a private area of roughly 29.5 square meters about 9.6 kilometers away from the main building of the university and approximately pay a base rent of 379 euro per month. In this sample, about 30% of respondents live in a living community, roughly 20% live alone, with their partners, or with their parents.

There is little difference in the variables between treatment and control group, suggesting that the randomization was successful. P-values adjusted for multiple testing using the bonferroni method are all equal to 1.

Finally, the self-assessment variables and the generation of the adjusted measures are described and analyzed. The variable *gensat* contains the self-assessed general satisfaction with the living conditions, and the variable *valsat* contains the self-assessed satisfaction with the value for money of the current apartment. Both variables were measured on a 5-point likert-type scale where the value 1 indicates very satisfied and 5 indicates very dissatisfied. These self-assessments are complemented by the vignette ratings on the same scale: Anna for the general satisfaction topic, Jana, Lena, and Sarah for the satisfaction with the value for money.

Table 3 presents the summary statistics of the self-assessments and the vignette ratings for the full sample and the treatment and control group subsamples as well as the p-values resulting from t-tests for differences in means between the treatment and control group. On average respondents are fairly satisfied with their living situation in general and the value for money of their current apartment. Furthermore, there is close to no difference in the means of the two self-assessments between the treatment and control group. The distribution of

Sample	Full		Treatment		Control		△ in means
Variable	Mean	St.D.	Mean	St.D.	Mean	St.D.	p-value
size	74.684	56.610	76.346	62.593	72.945	49.822	0.655
private area	29.493	16.652	28.557	14.542	30.472	18.624	0.392
number of rooms	3.413	2.950	3.561	3.339	3.257	2.484	0.442
living status							
alone	0.220	0.415	0.202	0.403	0.239	0.428	0.510
in a community	0.341	0.475	0.360	0.482	0.321	0.469	0.546
with partner	0.224	0.418	0.202	0.403	0.248	0.434	0.413
with parents	0.215	0.412	0.237	0.427	0.193	0.396	0.425
apartment characteristics							
living kitchen	0.242	0.429	0.263	0.442	0.220	0.416	0.456
living room	0.489	0.501	0.456	0.500	0.523	0.502	0.321
wooden floor	0.646	0.479	0.588	0.494	0.706	0.458	0.064
balcony	0.610	0.489	0.614	0.489	0.606	0.491	0.897
basement	0.789	0.409	0.798	0.403	0.780	0.416	0.737
yard	0.274	0.447	0.263	0.442	0.284	0.453	0.724
parking	0.381	0.487	0.404	0.493	0.358	0.482	0.485
distance to university	9.591	10.534	9.346	9.734	9.847	11.350	0.724
condition apartment	1.408	0.585	1.368	0.537	1.450	0.631	0.301
condition house	1.417	0.562	1.439	0.565	1.394	0.561	0.560
area characteristics							
vibrant	0.318	0.467	0.281	0.451	0.358	0.482	0.219
neutral	0.534	0.500	0.570	0.497	0.495	0.502	0.265
boring	0.148	0.356	0.149	0.358	0.147	0.356	0.961
base rent	378.536	299.128	348.601	315.104	409.844	279.469	0.127
price per square meter	16.551	16.566	15.362	14.916	17.795	18.118	0.274
move probability	42.936	40.416	44.167	41.013	41.648	39.931	0.643
self-assessed base rent	2.771	2.739	2.509	2.522	3.046	2.936	0.144
number of people	2.892	2.599	3.167	3.159	2.606	1.811	0.107
atmosphere	1.619	0.790	1.596	0.688	1.642	0.887	0.667

Table 4.2: Summary statistics of validation variables

the self-assessment answers is slightly tighter for the treatment group. This might be due to the fact that this group saw the vignettes first and was therefore more familiar with the research topic and the response scale.[7] Thus, the treatment may have set off a directional context effect which in this setting would mean that changing the question order resulted in less variation in the self-assessment (Buckley (2008)).

The summary statistics for the single vignette ratings support the assumption of vignette equivalence, since the value for money vignettes Jana, Lena, and Sarah are ordered as expected according to their assigned values for money. The treatment itself however did not fundamentally affect the vignette ratings at first glance. While Anna is assessed marginally

[7] This difference in variance is not significant at conventional significance levels.

Sample	Full		Treatment		Control		△ in means
Variable	Mean	St.D.	Mean	St.D.	Mean	St.D.	p-value
gensat	2.094	0.918	2.070	0.890	2.119	0.950	0.691
valsat	1.969	0.993	1.930	0.966	2.009	1.023	0.552
Anna	1.830	0.628	1.789	0.631	1.872	0.625	0.330
Jana	2.242	0.998	2.254	1.037	2.229	0.959	0.852
Lena	2.386	0.887	2.351	0.922	2.422	0.853	0.551
Sarah	4.413	0.671	4.456	0.612	4.367	0.729	0.323

Table 4.3: Summary statistics of self-assessment and vignette responses

better by the treatment group, Sarah is assessed marginally worse and there is virtually no difference in the assessment of the other two vignettes.

Table 4 provides sample means for all categories of the vignette ratings for the full sample and the treatment and control group subsamples as well as the p-values resulting from t-tests for differences in means between the treatment and control group.

Sample		Full	Treatment	Control	△ in means
Variable	Category	Mean	Mean	Mean	p-value
Anna	1	0.287	0.316	0.257	0.333
	2	0.605	0.588	0.624	0.583
	3	0.099	0.088	0.110	0.578
	4	0.009	0.009	0.009	0.975
	5	0.000	0.000	0.000	.
Jana	1	0.247	0.272	0.220	0.373
	2	0.413	0.368	0.459	0.173
	3	0.202	0.193	0.211	0.739
	4	0.130	0.167	0.092	0.097
	5	0.009	0.000	0.018	0.148
Lena	1	0.161	0.175	0.147	0.563
	2	0.404	0.430	0.376	0.416
	3	0.323	0.263	0.385	0.052
	4	0.112	0.132	0.092	0.348
	5	0.000	0.000	0.000	.
Sarah	1	0.000	0.000	0.000	.
	2	0.013	0.000	0.028	0.075
	3	0.063	0.061	0.064	0.931
	4	0.422	0.421	0.422	0.988
	5	0.502	0.518	0.486	0.642

Table 4.4: Category means of vignette responses

This closer look reveals slight, although mainly not significant, differences in the response behavior of the two groups. The treatment group more often assigns the best category to the general satisfaction vignette Anna. The worst vignette concerning the value for money assessment, Sarah, is seen more critically by the treatment group with less people providing a positive assessment and more people assigning the worst category. Concerning the other two vignettes, Jana and Lena, the categorical analysis reveals that here also the response frequencies of the single categories slightly differ between the groups. Jana, the best vignette in terms of value for money, produced mixed responses in the treatment group, with more people picking the best and worst categories than in the control group. The same picture can be observed with the assessment of Lena, here also the treatment group tends toward the extremes while the control group tends more toward moderate assessments. In conclusion, it appears that the reversal of question administration order affects the self-assessments and the vignette ratings differently. On the one hand, it tightens the distribution of the self-assessment answers. On the other hand, the distribution of the vignette ratings is driven more toward the extremes.

The nonparametric adjusted measure of the satisfaction with the living situation can now be generated using *gensat* and Anna. This measure C_{gen} is generated by assigning the value 1 if a respondent's self-assessment is better than this respondent's assessment of Anna's situation (i.e. smaller in value), 2 if the self-assessment is equal to Anna's assessment and 3 if the self-assessment is worse than the assessment of Anna (i.e. larger in value).

The same procedure can be applied to the value for money self-assessment and the answers to the three remaining vignettes to generate the adjusted measure of the satisfaction with the value for money. Here the recoding to generate the adjusted measure C_{val} uses the variable *valsat* and the three vignette variables Jana, Lena and Sarah and proceeds as follows:[8]

[8] Some respondents provided inconsistent and/or tied answers. Inconsistent answers arise when better vignettes are judged worse than the worse vignettes, tied answers when two vignettes are given equal ratings. In these cases, the best possible category was assigned to the observation.

$$\mathrm{C}_{val}=\begin{cases} 1 & \text{if } valsat<Jana \\ 2 & \text{if } valsat=Jana \\ 3 & \text{if } Jana< valsat<Lena \\ 4 & \text{if } valsat=Lena \\ 5 & \text{if } Lena< valsat<Sarah \\ 6 & \text{if } valsat=Sarah \\ 7 & \text{if } valsat>Sarah \end{cases}$$

Table 5 summarizes the adjusted measures for the full sample and the treatment and control group subsamples as well as the p-values resulting from t-tests for differences in means between the treatment and control group.

While the self-assessment of the general satisfaction with the living situation does not differ across groups, the adjusted measure C_{gen} reveals that the treatment group slightly assesses its situation worse than the control group. In the case of the adjusted measure of the satisfaction with the value for money of the current apartment, C_{val}, the adjusted measure is almost identical for both groups.

Sample	Full		Treatment		Control		△ in means
Variable	Mean	St.D.	Mean	St.D.	Mean	St.D.	p-value
C_{gen}	2.135	0.753	2.184	0.736	2.083	0.771	0.315
C_{val}	2.323	1.584	2.307	1.575	2.339	1.600	0.879

Table 4.5: Summary statistics of the adjusted measures

Table 6 provides sample means for all categories of the adjusted measures for the full sample and the treatment and control group subsamples as well as the p-values resulting from t-tests for differences in means between the treatment and control group. Here the findings are consistent with the broad impressions described above. The treatment group less often falls into the best category of C_{gen} and more often in the middle or worst category than the control group. Concerning the adjusted measure C_{val}, there is close to no difference apparent in how the groups assess themselves after DIF has been corrected for.

Sample		Full	Treatment	Control	△ in means
Variable	Category	Mean	Mean	Mean	p-value
C_{gen}	1	0.224	0.193	0.257	0.255
	2	0.417	0.430	0.404	0.694
	3	0.359	0.377	0.339	0.559
C_{val}	1	0.413	0.421	0.404	0.793
	2	0.323	0.316	0.330	0.818
	3	0.009	0.009	0.009	0.975
	4	0.076	0.070	0.083	0.729
	5	0.148	0.158	0.138	0.672
	6	0.027	0.026	0.028	0.956
	7	0.004	0.000	0.009	0.308

Table 4.6: Category means of the adjusted measures

4.4.2 Multivariate analysis

The following subsection proceeds with the analysis of the effects of a reversal of the question administration order on the adjusted measures C_{gen} and C_{val}. Each of these measures is used as dependent variable in estimation models using the validation variables and respondent characteristics that are estimated separately for the treatment and control group. The obtained estimation results are then compared between the two groups to examine whether the results of Hopkins & King (2010) are also supported in this application.

In order to support the results of Hopkins & King (2010) the comparison of the estimation results between the control and treatment group should lead to the following results. First, if the estimated coefficients differ between the two groups, the estimated coefficients in the treatment group should be closer to the expected association between a person's satisfaction and these variables. Second, the explanatory power of the models estimated in the treatment group should be higher than in the control group. The underlying hypothesis of this argument is that a correlation with the hypothesized sign or higher correlation of explanatory variables with the dependent variables is better in the sense that it speaks for construct validity.

The following table outlines the expected association of our validation variables with persons' satisfaction of their living situation and the value for money of their apartment. A positive association signifies that an increase in the validation variable is associated with a decrease of the satisfaction.

Validation variable	expected association
move probability	positive
size	negative
number of rooms	negative
number of persons	positive
private area	negative
atmosphere	positive
living room	negative
condition apartment	positive
condition house	positive
distance to university	positive
base rent	positive
self-assessed base rent	positive
price per square meter	positive
living kitchen	negative
wooden floor	negative
balcony	negative
yard	negative
parking	negative
vibrant	negative
boring	positive

Table 4.7: Expected associations between the validation variables and respondents' satisfaction measures

All estimation models presented in this paper are ordered probit models using robust standard errors and control for respondent's personal characteristics and the living status.[9] In all regressions, a positive coefficient signifies that an increase in the explanatory variable is associated with a decrease of the satisfaction. A negative coefficient on the other hand implies that an increase in the explanatory variable is associated with an increase in satisfaction.

Table 8 presents the estimation results for the adjusted general satisfaction measure, C_{gen}, table 9 for the adjusted value for money measure, C_{val}. The stars behind the variable names specify the significance level at which the estimated coefficients differ between the treatment and control group.

In both tables, it can be seen that the treatment results in statistically different coefficients for most of the significant explanatory factors for both adjusted measures. The clearly insignificant coefficients on some validation variables are mainly indistinguishable between both groups.[10]

[9] The coefficients of the respondent characteristics and the living status are not displayed here because there is no a priori hypothesis as to what their association with the self-assessments should be.

[10] For the remaining significant explanatory factors the difference between treatment and control group is almost significant and it cannot be excluded that a larger sample would have rendered these differences significant.

As to whether these changes in coefficients are desirable is open to dispute. In some cases, the estimated results in the treatment group make more sense. In table 8 this is the case for example for the effect of the condition of the apartment. The estimated coefficient in the control group is -0.510 which means that in this estimation a worse condition of the apartment is associated with an increase in the satisfaction with the living situation. In the treatment group, this coefficient is positive and significant signaling that a worse condition of the apartment in this estimation is associated with a lower satisfaction with the living situation. The same argument can be made for the changes in the estimated coefficients for the having a living kitchen, access to a yard and living in a vibrant area. Here, the coefficients change signs and move closer to their expected direction when looking at the treatment group.

Sample	Control		Treatment		
Explanatory variable	Coefficient	z	Coefficient	z	P - value
move probability	0.009	2.65	0.011	2.85	0.685
size	-0.003	-0.63	-0.002	-0.66	0.936
number of rooms	-0.294	-2.00	-0.053	-0.77	0.137
number of persons*	0.264	1.96	0.012	0.24	0.079
private area	-0.024	-1.98	0.003	0.23	0.128
atmosphere	0.582	2.16	0.078	0.36	0.143
living room**	-0.935	-2.55	0.107	0.33	0.034
condition apartment***	-0.510	-1.55	0.992	2.59	0.003
condition house*	0.586	1.76	-0.312	-0.99	0.051
distance to university**	0.070	2.93	0.008	0.53	0.031
base rent	0.000	0.21	0.000	0.21	0.976
self-assessed base rent	0.164	3.31	0.128	2.12	0.646
price per square meter	-0.007	-1.08	0.001	0.05	0.612
living kitchen***	0.433	1.15	-0.766	-2.81	0.010
wooden floor	-0.538	-1.42	-0.297	-0.98	0.620
balcony	0.541	1.76	0.193	0.69	0.402
yard**	0.806	2.21	-0.421	-1.28	0.013
parking	-0.082	-0.25	-0.113	-0.39	0.944
vibrant	0.601	1.87	-0.166	-0.57	0.076
boring	-0.363	-0.75	0.150	0.38	0.412
LL	-77.338		-90.090		
Number of observations	109		114		
LR chi2(31)	81.22		58.81		
Prob > chi2	0.000		0.002		
Pseudo R2	0.344		0.246		

All models are ordered probit models using the dependent variable C_{gen} and control for respondent's personal characteristics and respondent's living status. The stars indicate significance levels of the observed coefficient differences based upon robust standard errors. *** Significant at the 1 percent level, ** at the 5 percent level, * at the 10 percent level.

Table 4.8: Estimation results using the adjusted general satisfaction measure C_{gen}

However, there are also cases where the change in coefficients induced by the treatment does not make sense. This is for example the case for the estimated coefficient on the living room variable. In the control group, having a living room is associated with higher satisfaction, in the treatment group however, this association is no longer significant and the estimated coefficient changed signs. The same argument can be made for the changes observed for the coefficients on the condition of the house and the distance to the university.

In table 9, a similar picture arises. The changes in the coefficients induced by the treatment make sense for some variables, such as the number of rooms, the price per square meter, and the self-assessment of the base rent. In all of these cases, the estimated coefficients in the treatment group more close to the expected association of these variables with the satisfaction with the value for money.

However, we also observe changes that do not make sense. This is for example the case for the effect of the condition of the house. If we believe the estimated coefficient of the treatment group, then a worse condition of the house would be associated with a higher satisfaction with the value for money. Other variables where the coefficient differences do not make sense are the size of the apartment, the base rent, having a separate living room and living in a boring neighborhood.

The above comparison revealed that not all of the changes in coefficients that are brought about by the reversed question administration order in the treatment group make sense. Furthermore, the explanatory power of the models estimated in the treatment group is lower than in the control group. Thus, the hypothesis that reversing the question administration order improves the DIF correction is not supported in this application.

One possible explanation for this observation is that the reversal of the question administration order may have resulted in a change in respondent behavior concerning the vignette task by taking the anchoring vignette questions out of their context. This behavioral change may have crowded out beneficial effects on the self-assessment questions from the increased familiarity with the research concept and the response scale that are used as arguments by

Sample Explanatory variable	Control Coefficient	z	Treatment Coefficient	z	P - value
move probability	0.005	1.19	0.004	0.97	0.871
size**	-0.016	-3.16	-0.003	-1.18	0.030
number of rooms***	0.701	5.06	0.148	1.79	0.001
number of persons*	-0.488	-3.49	-0.169	-1.83	0.057
private area	-0.003	-0.20	0.009	0.79	0.490
atmosphere	-0.058	-0.31	0.160	0.71	0.460
living room*	-0.683	-1.91	0.085	0.29	0.097
condition apartment	0.521	1.91	0.112	0.32	0.356
condition house*	0.351	1.20	-0.482	-1.47	0.058
distance to university	-0.007	-0.76	0.008	0.50	0.421
base rent***	0.003	4.46	-0.002	-2.78	0.000
self-assessed base rent*	0.117	2.14	0.256	4.30	0.084
price per square meter***	-0.008	-0.91	0.039	2.64	0.006
living kitchen	-0.399	-0.98	0.403	1.40	0.107
wooden floor	0.005	0.01	0.278	1.06	0.524
balcony	-0.317	-0.99	0.010	0.03	0.461
yard	-0.565	-1.42	0.107	0.26	0.237
parking	-0.657	-1.91	-0.357	-1.03	0.538
vibrant	0.337	1.07	0.070	0.26	0.522
boring***	0.741	1.81	-0.911	-2.13	0.005
LL	-110.770		-126.251		
Number of observations	109		114		
LR chi2(31)	82.77		53.79		
Prob > chi2	0.000		0.007		
Pseudo R2	0.272		0.176		

All models are ordered probit models using the dependent variable C_{val} and control for respondent's personal characteristics and respondent's living status. The stars indicate significance levels of the observed coefficient differences based upon robust standard errors. *** Significant at the 1 percent level, ** at the 5 percent level, * at the 10 percent level.

Table 4.9: Estimation results using the adjusted value for money measure C_{val}

Hopkins & King (2010) for the reversal of question administration order.[11] If this holds true, we should observe that while the estimation results for the adjusted measures do not make more sense in the treatment group than in the control group, the reliability of the responses to the self-assessment questions should be higher in the treatment group.

This is the case if we observe the following:

1. When we compare the changes in coefficients between models estimated for the unadjusted and adjusted measures within both groups, fewer meaningful changes are observed for the treatment group than for the control group and

[11] Another possible explanation could be that the use of the anchoring vignettes procedure was not indicated in this domain. Appendix B presents a simple test for DIF that speaks against this hypothesis.

2. The explanatory power of the models using the unadjusted measures should be higher than for models using the adjusted measures in the treatment group while the reverse is observed for the models estimated in the control group.

For this comparison we now examine the estimation results when the model is estimated for the adjusted and unadjusted measures first within the treatment group and then within the control group.

Table 10 summarizes this comparison of estimation results for the treatment group. The left panel compares the results when *gensat* and C_{gen} are used as dependent variables, the right panel compares the results when *valsat* and C_{val} are used as dependent variables. In both panels, the first column specifies the expected association of the dependent variable with the respective validation variable. The second column specifies whether the change resulting from using the adjusted as opposed to the unadjusted measure moved the coefficient closer to this expected association. Finally, the last column presents the p-value on the null hypothesis that this change in coefficients is zero. Looking at the general satisfaction topic, the changes resulting from using the adjusted measure as dependent variable as opposed to the unadjusted measure are only significant for three variables (bold). In two of these cases, the change moved the estimated coefficient further away from the hypothesized direction when the adjusted measure was used. Almost the same picture arises when looking at the value for money results. Here all three significant coefficient changes moved the coefficients to the wrong direction.

Abstaining from the significance of the changes, a look at the variables where the hypotheses concerning the direction of the association are strongest (underlined in gray) does not alter the result. In almost two thirds of all these cases, the coefficients moved toward the wrong direction as a result of using the adjusted measure. Furthermore, the explanatory power is extremely higher in the models using the unadjusted measures. Using *gensat* as dependent variable, the R-squared is 0.349 as opposed to only 0.246 when using C_{gen}. In the value for money context, the corresponding R-squareds using *valsat* versus C_{val} are 0.357 as opposed to 0.176. The higher explanatory power is thus more pronounced in the value for money context.[12]

[12] Theoretically, the correction for DIF should have been more powerful here as a total of three anchoring vignettes was used in contrast to one in the general satisfaction context. However, while using three vignettes might make the correction better, it may also increase subject's understanding of the research topic more and thus increase the effect of intentional priming on the reliability of the self-assessment.

Comparison: gensat vs. C_{gen}				Comparison: valsat vs. C_{val}			
Explanatory variable	expected association	change ok	P - value	Explanatory variable	expected association	change ok	P - value
move probability	positive	Yes	0.618	**move probability**	**positive**	**No**	**0.042**
size	negative	Yes	0.808	size	negative	Yes	0.138
number of rooms	negative	Yes	0.185	**number of rooms**	**negative**	**No**	**0.031**
number of persons	**positive**	**Yes**	**0.003**	number of persons	positive	No	0.150
private area	negative	No	0.145	private area	negative	Yes	0.612
atmosphere	**positive**	**No**	**0.000**	atmosphere	positive	No	0.890
living room	negative	Yes	0.726	living room	negative	Yes	0.417
condition apartment	positive	No	0.283	condition apartment	positive	Yes	0.306
condition house	positive	No	0.427	condition house	positive	No	0.323
distance to university	positive	No	0.910	distance to university	positive	No	0.581
base rent	positive	No	0.255	base rent	positive	No	0.930
self-assessed base rent	positive	Yes	0.527	**self-assessed base rent**	**positive**	**No**	**0.011**
price per square meter	positive	Yes	0.853	price per square meter	positive	No	0.269
living kitchen	negative	Yes	0.265	living kitchen	negative	Yes	0.758
wooden floor	negative	No	0.204	wooden floor	negative	Yes	0.500
balcony	**negative**	**No**	**0.088**	balcony	negative	Yes	0.685
yard	negative	Yes	0.311	yard	negative	No	0.611
parking	negative	Yes	0.262	parking	negative	No	0.129
vibrant	negative	Yes	0.622	vibrant	negative	Yes	0.840
boring	positive	No	0.143	boring	positive	No	0.487

All models are ordered probit models controlling for respondent's personal characteristics and respondent's living status and were estimated using robust standard errors.

Table 4.10: Summary of coefficient changes in treatment group

Table 11 presents the same analysis for the control group. In the general satisfaction context, six coefficients changed significantly when we used the adjusted measure as opposed to the unadjusted measure as dependent variable. Of these changes, three move the estimated coefficient closer to the hypothesized association. In the value for money context, ten coefficients changed, five of them move the estimated coefficient toward the hypothesized direction.

Abstaining again from the significance of changes, but looking at the strong hypotheses, we find a reversal of the picture obtained in the treatment group. The observed changes in coefficients move toward the hypothesized direction in almost 70 percent of cases. Here, the explanatory power is lower in the models using the unadjusted measures. Using C_{gen} as dependent variable, the R-squared is 0.344 as opposed to only 0.234 when using *gensat*. In the value for money context, the corresponding R-squareds using C_{val} versus *valsat* are 0.272 as opposed to 0.242.

Comparison: gensat vs. C_{gen}				Comparison: valsat vs. C_{val}			
Explanatory variable	expected association	change ok	P - value	Explanatory variable	expected association	change ok	P - value
move probability	positive	No	0.945	**move probability**	**positive**	**No**	**0.019**
size	negative	No	0.225	size	negative	Yes	0.395
number of rooms	negative	Yes	0.440	**number of rooms**	**negative**	**No**	**0.000**
number of persons	positive	No	0.543	**number of persons**	**positive**	**No**	**0.000**
private area	**negative**	**Yes**	**0.017**	private area	negative	Yes	0.167
atmosphere	positive	Yes	0.273	atmosphere	positive	No	0.633
living room	**negative**	**Yes**	**0.037**	living room	negative	Yes	0.239
condition apartment	positive	No	0.135	condition apartment	positive	No	0.224
condition house	positive	Yes	0.401	**condition house**	**positive**	**Yes**	**0.046**
distance to university	**positive**	**Yes**	**0.012**	distance to university	positive	Yes	0.830
base rent	positive	Yes	0.243	**base rent**	**positive**	**Yes**	**0.000**
self-assessed base rent	positive	Yes	0.523	self-assessed base rent	positive	No	0.830
price per square meter	**positive**	**No**	**0.064**	**price per square meter**	**positive**	**No**	**0.052**
living kitchen	negative	No	0.474	**living kitchen**	**negative**	**Yes**	**0.079**
wooden floor	negative	Yes	0.162	wooden floor	negative	Yes	0.529
balcony	negative	No	0.205	balcony	negative	Yes	0.750
yard	**negative**	**No**	**0.016**	yard	negative	Yes	0.708
parking	negative	No	0.533	**parking**	**negative**	**Yes**	**0.011**
vibrant	negative	No	0.210	**vibrant**	**negative**	**No**	**0.052**
boring	**positive**	**No**	**0.073**	**boring**	**positive**	**Yes**	**0.000**

All models are ordered probit models controlling for respondent's personal characteristics and respondent's living status and were estimated using robust standard errors.

Table 4.11: Summary of coefficient changes in control group

Wrapping up, the DIF-correction resulted in less significant coefficient changes in the treatment group, the explanatory power of the models estimated on the unadjusted measures was higher than for the adjusted measures and the observed coefficient changes when using the adjusted as opposed to the unadjusted measure are not hypothesis confirming in a number of cases. The lower number of coefficient changes can be taken as evidence of a clearer understanding of the research matter or a higher familiarity with the response scale. Thus, the need for DIF-correction was consequently lower. However, the fact that the changes in coefficients are not rational in a number of cases hints at less beneficial effects of the reversal of question administration order. In the control group, more significant changes are observed, the explanatory power of the models on the adjusted measures is higher than for the unadjusted measures and more coefficient changes make sense. This indicates that the correction worked for the better with the original question administration order. These findings confirm our impression that the behavioral change regarding the vignette assessments induced be the reversed question administration order offset the beneficial aspects on the self-assessments.

As pointed out in the beginning when analyzing the distribution of answers to the single vignettes, there was more extreme ratings of the vignettes in the treatment group, hinting at possible respondent confusion by this task when the self-assessments are not introducing the topic. From the last two tables, it becomes obvious that this confusion may have totally offset the gains from the clearer understanding of the research question and a higher familiarity with the response scale.[13]

The analysis presented in this section calls into question whether changing the question administration order and using the resulting DIF-corrected measure was the way to go in this application. The evidence gathered here does point toward positive effects of changing the question administration order with respect to the reliability of the self-assessment answers in the form of higher explanatory power and more reliable estimation results. However, it points to the opposite direction concerning the reliability of the adjusted measures.

[13] One could wonder whether it is the tied and inconsistent answers that drive these results and a compound ordered probit as described in Hopkins & King (2010) should be used. However, performing the above analysis of the value for money satisfaction with only the best and the worst vignette, which gets rid of ties and inconsistencies, leaves all central results unchanged. Therefore, the cause for these results is unlikely to be the tied and inconsistent answers. The corresponding tables can be found in Appendix C

As an additional result, we find that for both adjusted measures that the coefficients on the dimensions presented in the vignettes differ significantly between treatment and control group in most cases and changes are close to significance in the remaining cases.[14][15] This can be taken as evidence that the priming focused respondents' attention on the vignette content in this application.

4.5 Conclusion and discussion

This research further examined the effects of intentional priming in an anchoring vignette framework as proposed by Hopkins & King (2010). Their reversal of question administration order was applied to a new anchoring vignette domain, the satisfaction with living circumstances.

We conducted a survey experiment in which we randomly assigned two survey versions with different question administration orders to respondents. The respondents of the control group completed the survey with the standard question administration order whereas the treatment group received a questionnaire with the reversed question administration order. For both groups, the answers given to the self-assessments and anchoring vignettes were used to construct two adjusted measures. These adjusted measures were then used as dependent variables in ordered probit models using the validation variables separately for each group.

The results we obtained comparing the estimation results between the treatment and control groups call into question whether intentional priming is beneficial for improving the anchoring vignette procedure in this application. A better estimation result for the treatment group is at best doubt worthy. The treatment group's estimation indeed results in some different coefficients than the control group's, however, not all of these changes are plausible. Furthermore, the models estimated in the treatment group have lower explanatory power than those estimated in the control group.

[14] see estimation results presented in tables 8 and 9.

[15] These dimensions are: for the general satisfaction: number of people, atmosphere, private area and living kitchen, for the satisfaction with the value for money: size, number of rooms, living area and base rent.

Comparing the results of ordered probit models estimated using the adjusted and unadjusted measures directly within the treatment group reveals that the DIF-correction does lead to some changes in coefficients. However, most of these changes are in the opposite direction of the hypothesis and the models on the unadjusted measures have higher explanatory power than their adjusted counterparts. The exact opposite is observed for the control group.

A look at the summary statistics provides a hint at why this may be the case. Moving the vignettes first has had two opposite effects on the treated group. It tightens the distribution of the self-assessment answers but also drives the distribution of the vignette ratings more toward the extremes. Since the answers to these two question types are related in construction of the adjusted measure, this had undesired effects on the adjusted measure. In this experiment, it is possible that respondent confusion from seeing the anchoring vignettes unintroduced by the self-assessments offset the gains from the clearer understanding of the self-assessment and the familiarity with the response scale. In the end, the unadjusted measure provided more thrust worthy results despite possibly left over bias resulting from DIF.

While the reversal of question administration order resulted in beneficial effects in the Hopkins & King (2010) setup, we cannot confirm their results in this application. As Tourangeau et al. (2000) note, the size and the presence of priming effects depend upon a number of factors, such as for example familiarity of the respondent with the issue at hand. It is therefore possible that the reversal of question administration order may lead to different effects depending on the context to which the anchoring vignettes method is applied. In this application the familiarity with the research topic was presumably high, so that it can be assumed that priming effects were of lower size than in the Hopkins & King (2010) political efficacy setting. This may explain why we were not able to obtain positive results of intentional priming.

Also, we found that the vignette content itself affects the effect of intentional priming. The validation variables that were also included in the vignette descriptions were most affected by the treatment. This fact will make the task of designing vignettes even more difficult and crucial for the DIF correction. Before the reversal of question administration order is implemented in surveys more research is needed, possibly in the form of qualitative interviews and experimental studies on changes in response behavior and design issues should be conducted.

Furthermore, the question arises why intentional priming is needed in the first place. The point of using anchoring vignettes is to correct for the differential mapping of respondents from their situation to a point on the answering scale. Intentional priming does however not only familiarize with the response scale. It may also change the perception of the domain to be assessed and this may depend upon the choice and content of the vignettes. The danger is that in the end, we cannot really be sure what exactly we measure with the self-assessment questions when they are preceded by the vignettes.

Limitations

The results of this study are subject to the following limitations. First off, the small sample size makes it impossible in some cases to obtain significant differences due to a lack of precision. Furthermore, the specific sample (mainly students of the University of Munich) and the specific research topic (satisfaction with the living circumstances) prevent external validity, so that the results obtained here will have to be validated in further studies using a general population sample and standard vignette applications. Finally, since no qualitative interviews were conducted, the cause of the results remains subject to some speculation. Further research should determine which cognitive processes are set off by reversing the question administration order. Also, the impact of intentional priming on respondent's understanding of the research topic are left for further examination.

4.6 Appendix

4.6.1 Appendix A: Questionaires

a) Vignette questionnaire

Questions regarding your current living circumstances

1. How satisfied are you in general with your living situation?

   □ very satisfied
   □ satisfied
   □ neither satisfied nor dissatisfied
   □ dissatisfied
   □ very dissatisfied

2. How satisfied are you with the value for money of your current apartment?

   □ very satisfied
   □ satisfied
   □ neither satisfied nor dissatisfied
   □ dissatisfied
   □ very dissatisfied

We will now describe a couple of persons in different living circumstances. We are interested in how you rate the living circumstances of these people. Please imagine that these persons are of your age and find themselves in similar life situations.

3. Jana lives in a newly decorated one room apartment with separate living Kitchen and a small balcony in a quiet neighbourhood close to the university. For the 40 square meter apartment Jana pays a monthly base rent of 480 Euro.

   In your opinion, how satisfied is Jana with the **value for money** of her current apartment?

   □ very satisfied
   □ satisfied
   □ neither satisfied nor dissatisfied
   □ dissatisfied
   □ very dissatisfied

4. Anna lives together with two roommates. She gets along fine with her roommates, except for the typical small arguments with roommates. Her room offers her enough space for her needs, the small kitchen serves as a meeting point of the apartment.

   In your opinion, how satisfied is Anna with her **general living situation**?

   □ very satisfied
   □ satisfied
   □ neither satisfied nor dissatisfied
   □ dissatisfied
   □ very dissatisfied

5. Sarah lives in a run down one room apartment with an integrated kitchen. The 25-squaremeter flat is located in a remote neighbourhood in some distance to the university. For this apartment Sarah pays a monthly base rent of 425 Euro.

   In your opinion, how satisfied is Sarah with the **value for money** of her current apartment?

   □ very satisfied
   □ satisfied
   □ neither satisfied nor dissatisfied
   □ dissatisfied
   □ very dissatisfied

6. Lena lives in a modern 35 square meter apartment with separate kitchen in a vibrant neighbourhood. For the apartment close to the university she pays a monthly base rent of 465 Euro.

   In your opinion, how satisfied is Lena with the **value for money** of her current apartment?

   □ very satisfied
   □ satisfied
   □ neither satisfied nor dissatisfied
   □ dissatisfied
   □ very dissatisfied

Thank you very much for your cooperation!

b) Supplementary questionnaire

Questions regarding your current living circumstances

7. How likely do you think it is that you will move into a different appartment again during the remainder of your studies? (Please state the probability in per cent. A low value indicates a low probability of another move, a high value for a high one.)

   __________ %

8. Do you live

   □ alone
   □ with roommates
   □ with your partner
   □ with your parents?

9. How large is the living space of your apartment? (Please make a guess if you do not know the exact answer.)

   ________ sq. meters

10. And how many rooms does your apartment have excluding kitchen and bathroom?

    □□ rooms

    ↱ If you live alone, please jump to questions 16.

11. How many people, including you, live in this apartment?

    □□ people

12. How large is the area of your apartment that you primarily use alone? (By this we mean your private area plus your share of the jointly used rooms. Please make a guess if you do not know the exact answer.)

    ________ sq. meters

13. How would you best describe your relationship with your roommates?

□ very harmonic
□ harmonic
□ neither harmonic nor tense
□ tense
□ very tense

14. How is your apartment equipped? Please check all answers that apply.

□ separate kitchen
□ separate living kitchen
□ separate living room
□ wooden floor
□ central heating
□ balcony/patio
□ garage storage place
□ garden plot
□ parking space

15. How do you assess the general condition of your flat?

□ in good condition
□ needs a little redecoration
□ needs extensive redecoration

16. How do you assess the general condition of the building your flat is in?

□ in good condition
□ needs a little redecoration
□ needs extensive redecoration

17. How would you best describe your residential area?

□ vibrant neighbourhood with a lot of shopping and leisure time facilities
□ quiet neighbourhood with some shopping and leisure time facilities
□ rather remote neighbourhood with few shopping and leisure time facilities

18. How large is the distance from your apartment to your university?
(If you are working full time please state the distance to your employer.)

________ kilometers

19. How much is your monthly base rent? (Please make a guess if you do not know the exact answer.)

□□□□□ Euro/Month

20. Thinking about comparable apartments, for example of your friends, for how appropriate do you consider your monthly rent on a scale from 0 to 10? (Please check one box on the scale, where the value “0” means “totally appropriate” and “10” means “totally inappropriate”. With the values in between you can grade your judgement)

totally appropriate										totally inappropriate
□	□	□	□	□	□	□	□	□	□	□
0	1	2	3	4	5	6	7	8	9	10

Thank you very much for your cooperation!

4.6.2 Appendix B: A simple test for DIF

Since the domain of satisfaction with living circumstances has never been analyzed with anchoring vignettes before, an alternative explanation for why intentional priming may not have worked in this application could be that there is no DIF to start off with. If this was the case, the models using the unadjusted and adjusted measures should not result in fundamentally different regression results within each group.

In order to test this hypothesis, ordered probit models are estimated using the unadjusted and adjusted measures as dependent variables for the control group and the obtained coefficients are compared.[16] This group provides the strongest test base since it uses the established ordering of self-assessments and the anchoring vignette task. Table 12 presents the estimation results. In both analyzed topics, the correction for DIF results in a number of coefficients that are statistically different (indicated by bold p-values) when using the adjusted measure instead of the unadjusted measure as dependent variable. Therefore, given the assumptions of response consistency and vignette equivalence hold, the self-assessments must have been biased by DIF.

Furthermore, for both models using the adjusted measures, the pseudo R-squareds are higher than in the counterparts using the unadjusted measures. Thus, the relationship between the adjusted measure and the explanatory variables is stronger here, a further indicator that the DIF-correction improved the estimation results.

This analysis provides evidence against the hypothesis that this research topic might be unsuited for applying the anchoring vignette procedure.

[16] This is the same model specification as in the main part of the paper.

Dep. Variable	gensat	C_{gen}		valsat	C_{val}	
Explanatory variable	Coefficient	Coefficient	P - value	Coefficient	Coefficient	P - value
move probability	0.009	0.009	0.949	0.013	0.005	**0.019**
size	-0.007	-0.003	0.225	-0.011	-0.016	0.395
number of rooms	-0.206	-0.294	0.440	0.213	0.701	**0.000**
number of persons	0.335	0.264	0.543	-0.045	-0.488	**0.000**
private area	0.000	-0.024	**0.017**	0.010	-0.002	0.167
atmosphere	0.383	0.582	0.273	0.015	-0.058	0.633
living room	-0.255	-0.935	**0.037**	-0.352	-0.683	0.239
condition apartment	-0.083	-0.510	0.135	0.823	0.521	0.224
condition house	0.360	0.585	0.401	-0.150	0.351	**0.046**
distance to university	0.019	0.070	**0.012**	-0.009	-0.007	0.830
base rent	-0.001	0.000	0.243	0.001	0.003	**0.000**
self-assessed base rent	0.136	0.164	0.523	0.126	0.117	0.830
price per square meter	0.003	-0.007	**0.064**	0.005	-0.008	**0.052**
living kitchen	0.219	0.433	0.474	0.128	-0.399	**0.079**
wooden floor	-0.011	-0.537	0.162	0.205	0.005	0.529
balcony	0.244	0.541	0.205	-0.240	-0.316	0.750
yard	0.058	0.806	**0.016**	-0.439	-0.564	0.708
parking	-0.253	-0.082	0.533	0.062	-0.657	**0.011**
vibrant	0.272	0.601	0.210	-0.227	0.337	**0.052**
boring	0.347	-0.363	**0.073**	-0.614	0.741	**0.000**
Log likelihood	-107.306	-77.338		-108.491	-110.770	
Number of observations	109	109		109	109	
LR chi2(31)	65.72	81.22		69.08	82.77	
Prob > chi2	0.000	0.000		0.000	0.000	
Pseudo R2	0.234	0.344		0.242	0.272	

All models are ordered probit models using robust standard errors and controlling for respondent's personal characteristics and respondent's living status. Bold p-values indicate a significant change between the estimated coefficients.

Table 4.12: Test for DIF in the control group

4.6.3 Appendix C: Tables for the corrected measure $C_{val}2$

The following pages present the descriptive statistics and estimation results for the adjusted measure $C_{val}2$, which was generated using only the best and the worst vignettes (Jana and Sarah). This measure may be less precise, however it contains no tied or inconsistent answers.

In table 13, it can be seen that the adjusted measure is only marginally better in the treatment group. This adjusted measure in the treatment group is in line with its self-assessment that was presented in the main part of the paper.

Sample	Full		Treatment		Control		△ in means
Variable	Mean	St.D.	Mean	St.D.	Mean	St.D.	p-value
$C_{val}2$	1.888	0.060	1.860	0.083	1.917	0.089	0.633

Table 4.13: Summary statistics of the adjusted measure $C_{val}2$

Concerning the single categories, table 14 only shows a slight reversal of the frequencies in the first two categories between the treatment and control group.

Sample		Full	Treatment	Control	△ in means
Variable	Category	Mean	Mean	Mean	p-value
$C_{val}2$	1	0.417	0.430	0.404	0.694
	2	0.323	0.316	0.330	0.818
	3	0.220	0.219	0.220	0.987
	4	0.036	0.035	0.037	0.949
	5	0.004	0.000	0.009	0.308

Table 4.14: Category means of the adjusted measure $C_{val}2$

The estimation results in Table 15 are coefficients obtained from ordered probit models using the adjusted measure $C_{val}2$ as dependent variable separately for control and treatment group.[17] The conclusion is the same as in the main part of this paper. The changes in the coefficients induced by the treatment have the expected direction for the effects of the number of rooms, the price per square meter, and the self-assessment of the base rent. However, again they don't make sense for example for the effect of the condition of the house, the size of the apartment, the base rent, and having a separate living room. Furthermore, the explanatory power is again higher in the control group.

[17] This is the same model specification as in the main part of the paper.

Sample Explanatory variable	Control Coefficient	z	Treatment Coefficient	z	P - value
move probability	0.003	0.82	0.004	0.92	0.924
size*	-0.014	-2.96	-0.004	-1.14	0.064
number of rooms***	0.658	4.85	0.172	1.84	0.003
number of persons	-0.437	-3.03	-0.215	-1.99	0.219
private area	-0.002	-0.18	0.008	0.62	0.574
atmosphere	-0.045	-0.23	0.120	0.54	0.582
living room*	-0.700	-1.96	0.189	0.63	0.056
condition apartment	0.460	1.82	0.085	0.24	0.390
condition house	0.258	0.89	-0.273	-0.84	0.222
distance to university	-0.003	-0.26	0.011	0.70	0.461
base rent***	0.003	4.18	-0.002	-2.81	0.000
self-assessed base rent	0.122	2.16	0.251	4.21	0.116
price per square meter***	-0.009	-1.04	0.042	2.70	0.004
living kitchen*	-0.430	-1.09	0.491	1.71	0.059
wooden floor	0.036	0.11	0.463	1.62	0.324
balcony	-0.379	-1.20	-0.143	-0.49	0.582
yard	-0.568	-1.34	0.151	0.38	0.215
parking	-0.769	-2.23	-0.248	-0.72	0.282
vibrant	0.500	1.56	0.177	0.67	0.437
boring***	0.798	1.94	-0.925	-1.99	0.006
LL	-93.781		-107.362		
Number of observations	109		114		
LR chi2(31)	80.49		53.68		
Prob > chi2	0.000		0.007		
Pseudo R2	0.300		0.200		

All models are ordered probit models using the dependent variable $C_{val}2$ and control for respondent's personal characteristics and respondent's living status. The stars indicate significance levels based upon robust standard errors. *** Significant at the 1 percent level, ** at the 5 percent level, * at the 10 percent level.

Table 4.15: Estimation results using the outcome measure $C_{val}2$

Finally, table 16 presents the comparison of estimation results within each group when the ordered probit models are estimated on the adjusted and unadjusted measures. This table follows the same setup as tables 10 and 11 in the main part of the paper. The left panel presents the results for the treatment group and the right panel for the control group. All significant changes observed when using the adjusted as opposed to the unadjusted measure are in the wrong direction and a majority of all changes is misdirected. This is again reversed in the control group. Here only half of the significant changes are in the wrong direction and less than half of all changes are misdirected. Furthermore, we again observe that the explanatory power of the models using the adjusted measure in the control group is higher than for the unadjusted measure, whereas in the treatment group, the explanatory power is only 0.200 when $\mathrm{C}_{val}2$ is used as dependent variable, but 0.357 when *valsat* is used in this place.

Treatment group				Control group			
Comparison: valsat vs. $C_{val}2$				Comparison: valsat vs. $C_{val}2$			
Explanatory variable	expected association	change ok	P - value	Explanatory variable	expected association	change ok	P - value
move probability	**positive**	**No**	**0.046**	**move probability**	**positive**	**No**	**0.005**
size	negative	Yes	0.130	size	negative	Yes	0.535
number of rooms	**negative**	**No**	**0.023**	**number of rooms**	**negative**	**No**	**0.001**
number of persons	**positive**	**No**	**0.089**	**number of persons**	**positive**	**No**	**0.002**
private area	negative	Yes	0.545	private area	negative	Yes	0.188
atmosphere	positive	No	0.780	atmosphere	positive	No	0.722
living room	negative	Yes	0.596	living room	negative	Yes	0.215
condition apartment	positive	Yes	0.364	condition apartment	positive	No	0.138
condition house	positive	No	0.690	condition house	positive	Yes	0.113
distance	positive	No	0.710	distance	positive	Yes	0.511
base rent	positive	-	0.792	**base rent**	**positive**	**Yes**	**0.001**
self-assessed base rent	**positive**	**No**	**0.010**	self-assessed base rent	positive	No	0.939
price per square meter	positive	No	0.384	**price per square meter**	**positive**	**No**	**0.043**
living kitchen	negative	Yes	0.975	**living kitchen**	**negative**	**Yes**	**0.070**
wooden floor	negative	Yes	0.946	wooden floor	negative	Yes	0.602
balcony	negative	Yes	0.426	balcony	negative	Yes	0.584
yard	negative	No	0.509	yard	negative	Yes	0.716
parking	**negative**	**No**	**0.070**	**parking**	**negative**	**Yes**	**0.004**
vibrant	negative	No	0.911	**vibrant**	**negative**	**No**	**0.016**
boring	positive	No	0.489	**boring**	**positive**	**Yes**	**0.000**

All models are ordered probit models controlling for respondent's personal characteristics and respondent's living status and were estimated using robust standard errors.

Table 4.16: Summary of coefficient changes with respect to the outcome measure $C_{val}2$

Since the same conclusion as in the main part of this paper is reached, this speaks against the hypothesis that our results are driven by using the wrong model given tied and inconsistent answers.

References

Angelini, V., Cavapozzi, D., & Paccagnella, O. (2011). Dynamics of reporting work disability in europe. *Journal of the Royal Statistical Society Series A*, 174, 621–638.

Baddeley, A. D. (1990). *Human memory: Theory and practice.* Needham Heights, MA, US: Allyn and Bacon.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O., & Doorslaer, E. K. V. (2011a). Education-related inequity in healthcare with heterogeneous reporting in health. *Journal of the Royal Statistical Society Series A*, 174, 639–664.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O., & Doorslaer, E. K. V. (2011b). Slipping anchor? testing the vignettes approach to identification and correction of reporting heterogeneity. *Journal of Human Resources*, 46(4), 875–906.

Bassili, J. N. & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, 60, 390–399.

Betsch, C., Ulshofer, C., Renkewitz, F., & Betsch, T. (2011). The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making*, 31(5), 742–753.

Braddock, C., Edwards, K., Hasenberg, N., Laidley, T., & Levinson, W. (1999). Informed decision making in outpatient practice: Time to get back to basics. *Journal of the American Medical Association*, 282(24), 2313–2320.

Braverman, J. (2008). Testimonials versus informational persuasive messages: The moderating effect of delivery mode and personal involvement. *Communication Research*, 35(5), 666–694.

Buckley, J. (2008). Survey context effects in anchoring vignettes. *mimeo. New York University, NY.*

Caro, F., Gottlieb, G., Hoffmann, S., Kesternich, I., & Winter, J. (2012). Patient decisions about knee replacement surgery: Contributions of second opinions, outcome forecasts, and testimonials. Unpublished working paper.

Chaiken, S., Wood, W., & Eagly, A. (1996). Principles of persuasion. In E. Higgins & A. Kruglanski (Eds.), *Social Psychology: Handbook of Basic Principles* (pp. 702–742). NY: Guilford Press.

Chang, L. & Krosnick, J. (2009). National surveys via rdd telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73, 641–678.

Cunningham, P. (2009). High medical cost burdens, patient trust, and perceived quality of care. *Journal of General Internal Medicine*, 24(3), 415–420.

Datta Gupta, N., Kristensen, N., & Pozzoli, D. (2010). External validation of the use of vignettes in cross-country health studies. *Economic Modelling*, 27, 854–865.

Diamond, P. & Hausman, J. (1994). Contingent valuation: Is some number better than no number. *The Journal of Economic Perspectives*, 8(4), 45–64.

Gallup (2012). Gallup healthways well-being index 2011. http://www.gallup.com/poll/154169/chronic-pain-rates-shoot-until-americans-reach-late-50s.aspx. Last retrieved 11-10-2012.

Greene, K., Campo, S., & Banerjee, S. C. (2010). Comparing normative, anecdotal, and statistical risk evidence to discourage tanning bed use. *Communication Quarterly*, 58(2), 111–132.

Greiner, B. (2004). The online recruitment system orsee 2.0 - a guide for the organization of experiments in economics. *University of Cologne, Working paper series in economics*, 10.

Gurmankin, A., Baron, J., Hershey, J., & Ubel, P. (2002). The role of physicians' recommendations in medical treatment decisions. *Medical Decision Making*, 22(3), 262–271.

Harrison, G. (2006). Hypothetical bias over uncertain outcomes. In J. List (Ed.), *Using Experimental Methods in Environmental and Resource Economics* (pp. 41–69). Edward Elgar, Northampton, MA.

Health-Dialog (2007). Treatment choices for knee osteoarthritis. Boston: Foundation for Informed Medical Decision Making.

Hensher, D. (2010). Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B: Methodological*, 44(6), 735–752.

Hensher, D., Rose, J., & Greene, W. (2005). *Applied Choice Analysis: A Primer*. New York: Cambridge University Press.

Hopkins, D. J. & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, 74, 201–222.

Jürges, H. & Winter, J. (2011). Are anchoring vignette ratings sensitive to vignette age and sex? *Health Economics*, forthcoming.

Kapteyn, A., Smith, J. P., & Soest, A. V. (2007). Vignettes and self-reports of work disability in the united states and the netherlands. *American Economic Review*, 97, 461–473.

Kapteyn, A., Smith, J. P., & Soest, A. V. (2012). Are americans really less happy with their incomes. *Review of Income and Wealth*.

Kapteyn, A., Smith, J. P., Soest, A. V., & Vonkova, H. (2011). Anchoring vignettes and response consistency. *Rand working paper*, WR-840.

Kasper, J., Mulley, A., & Wennberg, J. (1992). Developing shared decision making programs to improve the quality of health care. *Quality Review Bulletin*, 18(6), 183–190.

Kesternich, I., Heiss, F., McFadden, D., & Winter, J. (2012). Suit the action to the word, the word to the action: Hypothetical choices and real decisions in medicare part d. *Journal of Health Economics, in press.*

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98, 191–207.

Kristensen, N. & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics*, 15, 96–117.

Lanscar, E. & Louviere, J. (2008). Conducting discrete choice experiments to inform healthcare decision making: A user's guide. *Pharmacoeconomics*, 26(8), 661–677.

Louviere, J., Hensher, D., & Swait, J. (2000). *Stated Choice Methods: Analysis and Applications*. New York: Cambridge University Press.

March, J. & Heath, C. (1994). *A primer on decision making: How decisions happen*. Simon and Schuster.

Mark, T. & Swait, J. (2004). Using stated preference and revealed preference modeling to evaluate prescribing decisions. *Health Economics*, 13, 563–573.

McNair, J. (2009). http://ezinearticles.com/?what-is-the-average-reading-speed-and-the-best-rate-of-reading?&id=2298503. Last retrieved: 11-23-2012.

Mellink, W., van Dulmen, A., Wiggers, T., Spreeuwenberg, P., Eggermong, A. M., & Bensing, J. (2003). Cancer patients seeking a second surgical opinion: Results of a study on motives, needs, and expectations. *Journal of Clinical Oncology*, 21(8), 1492–1497.

Moumjid, N., Gafni, A., Bremond, A., & Carrere, M. (2007). Seeking a second opinion: Do patients need a second opinion when practice guidelines exist? *Health Policy*, 80(1), 43–50.

Peracchi, F. & Rossetti, C. (2012). Heterogeneity in health responses and anchoring vignettes. *Empirical Economics*, 42(2), 513–538.

Rand (2012). Rand american life panel. https://mmicdata.rand.org/alp/index.php?page=panel. Last retrieved: 11-02-12.

Rice, N., Robone, S., & Smith, P. C. (2010). International comparison of public sector perfomance: The use of anchoring vignettes to adjust self-reported data. *Evaluation*, 16, 81–101.

Rossi, P. & Anderson, A. (1982). The factorial survey approach: an introduction. In P. Rossi & S. Nock (Eds.), *Measuring Social Judgments: The Factorial Survey Approach*. Beverly Hills: Sage.

Schwitzer, G. (2002). A review of features in internet consumer health decision-support tools. *Journal of Medical Internet Research*, 4(2), e11.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, 1, 161 – 176.

Soest, A. V., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. (2011). Validating the use of vignettes for subjective threshold scales. *Journal of the Royal Statistical Society Series A*, 174, 575–595.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232.

US-Census (2012). http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml. U.S. Census Bureau American FactFinder. POPULATION 60 YEARS AND OVER IN THE UNITED STATES. 2011 American Community Survey 1-Year Estimates. Last retrieved 11-10-2012.

Van Dalen, I., Grothoff, J., Stewart, R., Spreeuwenberg, P., Groenewegen, P., & van Horn, J. (2001). Motives for seeking a second opinion in orthopaedic surgery. *Journal of Health Services Research and Policy*, 6(4), 195–201.

Vonkova, H. & Hullegie, P. (2011). Is the anchoring vignette method sensitive to the domain and choice of the vignette? *Journal of the Royal Statistical Society Series A*, 174, 597–620.

Winterbottom, A., Bekker, H., Conner, M., & Mooney, A. (2008). Does narrative information bias individual's decision making? a systematic review. *Social Science and Medicine*, 67, 2079–2088.

Yeager, D., Krosnick, J., Chang, L., Javitz, H., Levindusky, M., Simpser, A., & Wang, R. (2011). Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709–747.

List of Tables

Curriculum Vitae

10/2008 - present	Research and Teaching Assistant, Ph.D. Program in Economics Munich Graduate School of Economics Ludwig-Maximilians-Universität, Munich, Germany
04/2007 - 05/2008	Studies in Economics, Master of Arts Ludwig-Maximilians-Universität, Munich, Germany
08/2005 - 05/2006	Studies in Economics, Visiting Student Missouri Western State University St. Joseph, Missouri, USA
10/2003 - 09/2006	Studies in European Economic Studies, Bachelor of Arts Otto-Friedrich Universität Bamberg, Germany
03/2003	Abitur Gymnasium im Alfred - Grosser - Schulzentrum Bad Bergzabern, Germany
02/27/1984	Born in Landau in der Pfalz, Germany