Kløjgaard, Mirja Elisabeth; Bech, Mickael; Søgaard, Rikke

Article

Designing a stated choice experiment: The value of a qualitative process

Journal of Choice Modelling

Provided in Cooperation with:
Journal of Choice Modelling

Suggested Citation: Kløjgaard, Mirja Elisabeth; Bech, Mickael; Søgaard, Rikke (2012) : Designing a stated choice experiment: The value of a qualitative process, Journal of Choice Modelling, ISSN 1755-5345, University of Leeds, Institute for Transport Studies, Leeds, Vol. 5, Iss. 2, pp. 1-18

This Version is available at:
https://hdl.handle.net/10419/66822

Journal of Choice Modelling, 5(2), pp 1-18
www.jocm.org.uk

Designing a Stated Choice Experiment: The Value of a Qualitative Process

Mirja Elisabeth Kløjgaard[1,*] Mickael Bech [2,†] Rikke Søgaard [3,Ŧ]

[1] Centre for Applied Health Services Research and Technology Assessment, University of Southern Denmark
[2]COHERE - Centre of Health Economics Research, University of Southern Denmark
[3] Centre for Applied Health Services Research and Technology Assessment, University of Southern Denmark

Received 23 September 2011, revised version received 25 April 2012, accepted 25 April 2012

Abstract

Designing a stated or discrete choice experiment (DCE) involves a process of developing, testing and optimizing the experiment questionnaire. This process is important for the success of the experiment and the validity of the results, but it is often not reported thoroughly. In the field of health care, one faces challenges in relation to what makes sense both for the respondent and what has clinical relevance, especially in situations with little evidence and unclear choices, where the decision making process is not clear or informed. This is the case for degenerative spine diseases, where the selection of candidates for surgical rather than non-surgical treatment has been widely discussed and where surgery rates accordingly vary across settings.

In the present work, we demonstrated how the qualitative process significantly impacted and guided the design, and it was clear that a less thorough qualitative process would have resulted in a less useable and valid design. To elicit relevant attributes and levels for a DCE, fieldwork in clinical departments in Danish hospitals was performed and has been supplemented by qualitative interviews with patients and doctors. Systematic and thorough qualitative investigation of the decision context relevant attributes and levels and appropriate framing appears valuable in the process of designing a DCE for quantitative pilot testing.

Keywords: Design, Discrete Choice Experiment, Qualitative methods, Attribute and Level selection

* Corresponding author, T: +45-65504093, F: + 45-65503880, mirk@cast.sdu.dk
† T: +45-65504234, F: + 45-65503880, mbe@sam.sdu.dk
Ŧ T: +45-65503889, F: + 45-65503880, ris@cast.sdu.dk

ISSN 1755-5345

1 Introduction

Discrete choice experiments (DCE) are increasingly used to elicit preferences about health care interventions because interventions can be described by their attributes and, an individual's valuation depends on the levels of these attributes (Ryan and Gerard 2003). The relative importance of attributes to individuals is usually elicited by presenting the respondent with a series of choice sets where the levels of the attributes are changed across the sets. According to the literature from different research fields, including marketing and transportation, the stepwise design process includes the two initials steps of identifying attributes and assigning levels, both of which can be informed by qualitative work (Ryan 1999, Hensher et al. 2005, Coast and Horrocks 2007, Coast et al. 2012). These initial steps are of the utmost importance for the success of the experiment and the validity of the results.

1.1 Aim of study and description of the context

This paper explores the pros and cons of a series of qualitative approaches used to aid in the design of a DCE. The approaches include a review of the literature, observational fieldwork in a hospital ward, interviews with doctors and patients and a qualitative pilot test. The work was conducted as a step-wise process, from establishing the decision-context through proposing a complete questionnaire that could be applied in a later quantitative pilot test before the large-scale DCE was implemented in patients with degenerative disc diseases (DDD) of the spine.

DDD patients suffer from a very disabling disease and ultimately have to make a difficult choice between the two very different treatment paths of surgical or non-surgical treatment. The choice of treatment modality is distorted by conflicting evidence and by no promise of recovery with either treatment modality (Chou et al. 2009 and Allen et al. 2009). This has resulted in a remarkably variable surgery rate across the world and much discussion of the (cost-) effectiveness of the treatment strategies (Irwin et al. 2005 and Smith et al. 2009).

The paper is structured as follows. First, the introduction describes the process of selecting attributes and levels based on the DCE literature. The introduction also touches upon the use of qualitative methods to inform the process. The methods section describes the methods used in the case and the overall qualitative process of reviewing the literature, observing decision making situations in the clinic, interviewing doctors, interviewing patients, and testing and revising the proposed pilot questionnaire. Next, the results section reports on the identified attributes, the levels and the alternatives and, in particular, on how the thorough qualitative work has changed and coloured the design. This is reported for specified phases of the process. A final section discusses the optimality of the present approach in terms of the choice and the order of qualitative tests, the choice of respondents and their impact on the final design, and the generalisability to other contexts.

1.2 Selecting attributes and levels

Attributes can be quantitative (e.g., waiting time) or qualitative (e.g., the choice of hospitals) and are considered to be based on knowledge gathered from interviews, group discussions, literature reviews, and expert opinions (Coast and Horrocks 2007). The context and goal of choice experiments can be very different, and there is no gold standard for the definition of attributes (Louviere et al. 2000). There is, however, some

agreement on what to consider when approaching the first step of identifying attributes.

DCEs can rarely include all of the important attributes, but it is important that the most important ones relevant to the majority of respondents are included. If this is not the case, respondents can make assumptions about excluded attributes, which can affect the validity of the experiment. This should not be focused upon a priori, but it can also be tested in a pilot (Lancsar and Louviere 2006). The combined set of attributes must describe what the choice consists of, and the attributes must be chosen so that respondents will be willing to make trade-offs between them following the underlying economic theoretical framework with compensatory decision-making. The individual attributes must also reflect the true motivations for the respondents in the given real choice situation (Lancsar and Louviere 2008). Further attributes should be chosen to be either generic for all alternatives or specific for one or more alternatives, and a limited number of attributes must be ensured.

Attributes must be formulated in a way that ensures that the respondents understand the content of the attribute and in a clear and concise manner. Some of the literature directly points to qualitative work as a basis for ensuring this (Mays and Pope 2000, Kuper et al. 2008).

Some of the literature also points to the need for special attention towards attributes involving a description of risk. If such an attribute is relevant and important to the respondents, it should be explained thoroughly in the experiment to avoid mistakes based on the proposed difficulty of respondents in understanding the probabilities (Peters et al. 2006). If the experiment aims to explore a political challenge or has a direct policy question, this must also be considered and included when choosing attributes (Ryan 1999, Bennett and Blamey 2001). In the process of selecting attributes, one has to be aware of causal relationships and interconnections between attributes and their mutual dependence because this can affect the respondents' behaviour in the experiment and muddle the utility measures (Blamey and Bennett 2002). The issue of causality or dependency can simply be handled by excluding one of the attributes. This should be accompanied by a thorough definition in the introductory test, making sure that the respondents all assume the same content of attributes. Alternatively, two or more attributes can be combined, possibly resulting in a loss of information. It is important that these issues are identified and addressed at an early stage.

The second step of the recommended design process is to determine the appropriate levels for each attribute (Ryan 1999 and Lancsar and Louviere 2008). Once again, the levels must be relevant and easy to comprehend. Additionally, the levels must have a scope or range that captures and ensures trade-offs between attributes while still being acceptable to the respondent (Green and Srinivasan 1978). This important because the level range affects the estimates derived from the design. If the scope is inappropriate, respondents might consider the differences to be either unimportant, resulting in dominated levels, or very important, resulting in dominating levels. This will affect the willingness of the respondents to make trade-offs. Levels, therefore, affect the results that should be interpreted when considering the chosen levels. In particular, an attribute with a significant coefficient is significant under the circumstances provided by the levels, and it might not have been significant under different circumstances.

The levels also determine the types of effects possible to consider because a two-level attribute only allows for the estimation of linear effects, while more than two levels can make an estimation of often present non-linear effects. Providing even

spaces between levels can be useful for interpreting the estimated effects (Lancsar and Louviere 2008)

The number of levels assigned to each attribute is also important because an increase in the number of levels assigned to an attribute increases the possible significance of that attribute. This is because respondents tend to give more value to attributes with more levels (Ratcliffe and Longworth 2002). This problem can, however, be minimized by adding the same number of levels to all attributes (if that is sensible).

Most settings in which choice experiments are performed require the researcher to be deeply familiar with the respondents, their personal "evaluation system", and the decision-making situation that they are in. Although there is much written about the theoretical issues to consider when designing a DCE, the literature provides little recognition or guidance for the process of getting the prior knowledge that is accepted as necessary to choose the attributes and levels that fulfil the described criteria and ensure face validity (Hall et al. 2004). Face validity refers to the extent to which a DCE assesses what it is meant to assess without the effects of different biases influencing the result. In other words, it represents the extent to which the DCE includes the factors of significance and thereby ensures the revelation of the true utility of an attribute (Batemann et al. 2002).

A recent review of DCEs published between 1990 and 2008 (n=114) illustrates that the use of qualitative pre-studies to improve the face validity appears to have been rising and is increasingly reported in applied DCEs in the field of health care (De Bekker-Grob et al. 2012). The review shows that the qualitative process has been used to select both attributes and levels and to pre-test whole questionnaires, whereas the reporting of both the qualitative methodology and the impact on the design is generally lacking. In the DCEs published from 1990-2000, 18 percent (n=6) reported that some sort of qualitative work had been applied for the identification of attributes, whereas in the DCEs published from 2001-2008, the percentage had increased to 69 percent (n=79). Similarly, with respect to defining levels of attributes, 18 percent (n=6) of studies published from 1990-2000 reported that qualitative work had been undertaken, and this appeared to have increased to 33 percent (n=38) in 2001-2008. With respect to the qualitative pre-testing of whole questionnaires, the development, in contrast, appears to have been negative: 47 percent (n=16) of studies published from 1990-2000 used qualitative tests, whereas only 38 percent (n=36) of studies published between 2001-2008 used qualitative tests.

Different approaches can be used to perform different tasks in the design process or to conduct pilot tests of designs. The approaches to the qualitative process include the use of focus groups (e.g., Philips et al. 2002, Roux et al. 2004, Kjaer et al. 2006, Ratcliffe et al. 2004, Salkeld et al. 2003, Mark et al. 2003) and interviews (e.g., Seston et al. 2007, Shackley et al. 2001, Telser and Zweifel 2002, Ratcliffe et al. 2002, Sculpher et al. 2004) that help to define the attributes and levels and the framing and lay-out of the questionnaire. Further, debriefing, think aloud exercises and free text commenting have been used to identify misunderstandings and to explore the experience and perception of filling out choice-sets (Lloyd et al. 2007, Mortimor and Segal 2008, Ashcroft et al. 2006). The chosen approach, or indeed the combination of approaches, should depend on the aim of the qualitative work and the context in which the experiment is to be applied.

2 Process of selecting and testing design elements

Establishing the exact decision-context to study was itself an initial step in the process of designing the DCE. Because the field is characterized by minimal and often conflicting evidence, the first part of the process was to concretize the relevant decision-making context in which patient preferences were of interest. Second, potential attributes were identified and tested for relevance. Next, the levels of attributes were proposed and evaluated. Finally, the alternatives and the formulation and framing of the task were designed and tested. The following describes the qualitative methods used in the process.

2.1 Literature review

A systematic literature review of preferences for spine treatment was conducted to screen the literature for possible previous studies and for potential attributes and levels. Relevant databases were systematically searched and supplemented with a hand-search. The search resulted in 137 articles, of which only 6 articles met the inclusion criteria of dealing with preferences. The systematic literature review has been appropriately reported elsewhere but is summarized in this paper with a particular focus on its role for proposing attributes and levels (Kløjgaard et al. forthcoming).

2.2 Observational fieldwork

To get a better understanding of the disease, its treatment and the decision-making context, observational fieldwork was performed in a spine surgical hospital ward in Copenhagen. The fieldwork consisted of a three-day observational stay in an ambulatory centre that received patients who were being reviewed for their treatment indications or who had already received surgery. The patients' questions and thoughts in relation to choosing treatment and their motivational explanations were observed. Furthermore, the doctors' part in the decision-making process was observed and was subsequently discussed with the doctor.

2.3 Interviews with clinicians

To make sure that the proposed attributes truly reflected the decision-making context and were clinically plausible, an interview of doctors working surgically and non-surgically with DDD patients was undertaken based on a preliminary version of the DCE questionnaire. The interviews took place at two hospitals that are highly specialized in surgical treatment and at one hospital specialized in non-surgical treatment. Both interviewees were professors with extensive experience in both research in and the treatment of spine patients but with a difference in focus, one being an expert in surgical treatment and the other in non-surgical treatment. The interviews comprised of both a general introduction to the treatment performed in each hospital and a discussion of the proposed design. The interviews were recorded and transcribed.

2.4 Interviews and qualitative pilot tests aimed at patients

Structured interviews and think aloud pilot tests of the proposed questionnaire were conducted in a hospital ward with admitted patients who were in the diagnostic process. A total of three interviews, each approximately two hours in duration, were conducted. All interviewed patients suffered from lower back pain caused by degenerative disease, and all were both physically and mentally affected. They all had some, but not all, indications for surgery and were awaiting a consultation with the doctor about the choice between surgical or non-surgical treatment. The interviews were recorded and transcribed, and the content was analyzed with manual coding.

The interviews were structured in two parts. In the first part, patients were asked about their disease, the current state of treatment, treatment history and demographic background data. Next, a general discussion of their treatment wishes or preferences and motivations for such was performed. In the second part, a pilot test of 10 choice sets was conducted as a think aloud exercise. The qualitative pilot test included both examples of a labelled (surgery vs. non-surgery) and unlabeled (treatment A vs. treatment B with treatment modality as attribute) design and different orders of attributes.

Next, the following issues were discussed:

- Attributes: Inclusion, formulation, conciseness, dominance, interconnectedness and mutual dependence. Perception of order and the change in the order.
- Levels: Formulation, range and acceptability.
- Labelling: Perception of the labelled and unlabeled part of the qualitative pilot.
- Framing: Perception and understandability of the given text and explanations.
- Opt-out: Perception of the opt-out option of "neither of these".
- Total design: Layout, length, complexity and overall experience with filling out the pilot.

Finally, the respondents were asked to rank the attributes and give an overall evaluation of the questionnaire.

3 Results

3.1 Phase 1: Exploratory start - literature review and observational field work

Exploring the literature for previous DCEs in spine surgery quickly revealed that little work has been done in the field. The few existing studies were, however, helpful in elucidating some of the motivational factors and thought-processes behind choosing treatments.

The literature suggested that patients favour surgical over non-surgical treatment because this modality was perceived to be most likely to reduce pain, which was the main concern. The literature suggested that both the duration and severity of pain and its onset was important to patients. Further literature provided other possible attributes, including walking ability, the ability to pursue leisure activities, the ability to work and neurological deficits.

The literature review also helped to identify the most relevant patient group, namely, those faced with the most difficult treatment choice where a relative, but not absolute, indication for one specific treatment is present (specifically, patients with

DDDs). Hence, the literature review helped formulate the research question for the DCE, focusing on the value for DDD patients of getting better and the motivations behind choosing specific treatment paths.

The fieldwork revealed that patients suffering from back pain were greatly disabled both physically and mentally. Their ability to work both in and outside of their homes was greatly decreased, and their mental ability and social capacity was lowered. Patients who had chosen to receive surgery were not always content with this choice when returning for check-ups and still complained about pain and disabilities. Some were even somewhat regretting their choice. The work also made the doctor's role clearer, as it showed that they are a good source of information for the patient concerning the conflicting evidence on treatment choices. The potential attributes resulting from the literature review and fieldwork are listed in Table 1.

Table 1. Possible attributes

Attributes
Pain – duration, severity and onset
Walking ability
Ability to pursue leisure activities
Ability to work
Neurological deficits
Physical ability
Mental state
Social ability
Exercises to perform
Ability to sleep
Use of pain killers
Time spent on treatment
Time to wait for treatment to begin
Time for effect to occur

3.1.1Conclusions from phase 1 to pursue in phase 2

From the first phase, it became clear that not only are the patients faced with a difficult decision to make, it can also be difficult for doctors to know what treatment to recommend or offer. Given the literature, it seems as if patients have a preference for surgery, but the fieldwork proved that surgery did not always meet the success criteria and wishes of the patients. The patients were often under the impression that surgery was able to free them from all pain, for which there is no evidence. Hence, a surgery leading to no or little pain-reduction left some patients very unsatisfied, even though the doctors did inform them about all possible outcomes prior to surgery.

The fieldwork revealed that patients had many concerns about their future. Their disease had often come on without warning, and patients wanted to regain physical and mental/social ability after treatment.

It also became clear that many patients had long treatment paths, having seen different sorts of health professionals and their general practitioners (GPs) before referral to a hospital.

The observational study also showed that it took a while for patients to experience the total treatment effect, regardless of their treatment choice, and that all patients had the possibility of experiencing a relapse.

It seemed clear that the list of possible attributes contained many overlapping and mutually dependent attributes and could easily be reduced to more general categories. Patients seemed to be concerned with their abilities and their quality of life and were

very concerned with the pain they were in and the medication that they took to limit it. Furthermore, there were some time issues to explore.

The knowledge gained also pointed to possible levels or outcomes of the attributes. It became clear that all of the desired effects could actually be gained from treatment, but there was a possibility of no effect or an actual worsening of the disease state. This led to a proposed level definition comprising three possible outcomes of the chosen treatment modality.

The first phase led to the identification of seven overall attributes, seen in Table 2, that were perceived to be inclusive of most of the issues derived from the literature and fieldwork. All attributes could be meaningfully scaled using three levels, thereby preventing a differing number of levels from influencing attribute significance.

Table 2. First proposal for attributes and levels based on literature review and observational fieldwork

Attribute	Levels
Pain at the end of treatment	More than now, same as today, less than now
Physical ability at the end of treatment	More than now, same as today, less than now
Social/mental ability at the end of treatment	More than now, same as today, less than now
Use of pain-killers	Increased, same as today, decreased
Time spent on treatment (your own effort)	10 hours/week in 1/3/5 months
Waiting time for effect to occur	1/6/12 months
Risk of relapse to present state	None, 10, 20

3.2 Phase 2: First design proposal – interviews with doctors

The interviews with the doctors helped to test the knowledge gained in the first phase. It ensured that the attributes were reflective of their experience and knowledge dealing with the target group, and it ensured clinical plausibility. Furthermore, it helped to create and formulate levels for each attribute.

The doctors pointed towards a range of things to consider, mostly concerning the proposed levels.

The levels proposed for the risk of relapse proved to be clinically unrealistic. The doctors both agreed that a relatively large number of patients experience relapse, and they therefore suggested an increase in the probability.

Furthermore, the suggested 10 hours per week of training was too high according to both doctors. Patients rarely participated in more than 2-4 hours of planned training per week, regardless of whether the training was in the form of exercise or rehabilitation after surgery. They further suggested that the levels for waiting time for effect were too broad because the effects from treatment could be seen faster than suggested. A more narrow range of that level seemed more realistic.

The doctors both pointed to a relationship between pain and medication and a difference in medication-strategy depending on treatment modality. After surgical treatment, the medication level was regulated differently than with non-surgical treatment. With both treatment modalities, the use of painkillers was, after a while, up to patients to choose. This suggested that the attribute had a causal relationship with pain or that it could be considered to be dependent on the level of pain. Hence, the attribute was omitted from the list.

Both doctors pointed to the fact that, regardless of treatment modalities, different approaches could be taken for different patients, depending on what the doctors involved in the treatment thought would have the best effects. Surgical treatment can

be varied in terms of the degree of how radical the intervention is, and non-surgical treatment can comprise different exercises and group sessions. This suggests that an unlabeled experiment that includes an attribute of treatment modality might reflect the choice situation better.

The process also resulted in a draft of the DCE because the interviews helped to enlighten the doctors' understanding and formulation of the patients' choice task.

3.2.2. Conclusions from phase 2 to pursue in phase 3

This phase helped to ensure the realism of the questionnaire. The doctors pointed to changes in both attributes, and the levels and the decision situation for the patients, as seen from the clinical point of view, were described and qualified. However, it was important to ensure that the clinically possible attributes and levels did not overshadow the purpose of the DCE, which was to make people reveal trade-offs. Therefore, it was important to test whether patients thought that the levels had an appropriate range of being acceptable yet were still seen as trade-offs. It would not be feasible to have a design that did not provide enough information due to levels that are too narrow. Also, clinical relevancy is variable as research evolves and practices change. DCEs are hypothetical and can include hypothetical attributes/levels if they are important to the respondents, yet it was important to ensure that the attributes and levels were approved by the clinicians. This was to ensure that we did not present to patients who are participating in a "real life" treatment with overly hypothetical scenarios that could potentially make them doubt the treatment that they are in. These issues were tested during patient interviews.

The second phase led to the qualification of attributes as seen in Table 3.

The proposed attributes and levels were used to construct 10 choice sets for discussion and testing in the next phase. Five sets were labelled, and, after filling them out, an extended framing based on the doctors' explanations was given. This was followed by five unlabeled choice sets, including alternatives to the choices of only "types" of surgical and non-surgical treatment.

Table 3. Second proposal for attributes and levels based on doctors' comments

Attribute	Levels
Pain at the end of treatment	More than now, same as today, less than now
Physical ability at the end of treatment	More than now, same as today, less than now
Social/mental ability at the end of treatment	More than now, same as today, less than now
Time spent on treatment (your own effort)	3 hours/week in 1/3/5 months
Waiting time for effect to occur	1/3/5 months
Risk of relapse to present state	10, 20, 30
Treatment modality	Surgical, non surgical

3.3. Phase 3: Qualitative pilot tests

Interviewing patients made their motivations for treatment preferences clearer. All of the patients have had a longer period of chronic pain and inability to work. Over this time period, various health professionals had been consulted, resulting in the prescription of painkillers and exercises. During this period, the preference for treatment was already beginning to form before consultation with a specialist,

resulting in much frustration about the need for hospital admission because this led to (more) waiting time, (further) diagnostic tests and (new) discussion of treatment options to arrive at a final indication. Although this process was not understandable to the patients, it appears to have led to much thinking about their preferences. The patient interviews thus resulted in a range of changes and modifications as described below.

3.3.1. Attributes

Although the respondents all agreed that the chosen attributes reflected their view of wishes and fears included in the decision making process, it became very clear that the included six attributes were too many. The respondents had problems remembering the first attributes when looking at the later attributes, and two respondents ended up answering lexicographically or using heuristics with emphasis on the last or the two last attributes, despite not ranking these the highest.

Furthermore, two respondents argued that if their pain were moderated, then their physical and mental well-being would automatically be bettered, hinting at a perceived causal relationship and a mutual dependence or interaction of the attributes of pain and physical and mental/social ability.

Furthermore, two respondents read the word "ability" as "aim" because those two words are close in Danish. This fact proved to make the attributes difficult to understand and pointed to a need for reframing them.

Special attention was given to the included risk attribute, but none of the respondents had difficulty understanding or interpreting the attribute. They all considered it to be relevant and important, and for two respondents it was ranked rather high and was claimed to impact their choices. All respondents seemed to be aware that relapse was a possibility, and some had experience with it.

"Waiting time for effect to occur" or "time spent on treatment" were meant to be the "payment-vehicle" understood as the negative attribute that they had to be willing to pay with to gain improvement. None of the respondents looked upon these attributes as being of great importance. They were happy just to be promised that the treatment would indeed be performed, and they were seemingly ready to wait and give whatever time was needed. All respondents answered that they perceived the time spent on treatment to be positive and that they would enjoy being able to perform exercises and be physically active if that was what was required of them. This implied a need for rethinking those attributes. One respondent considered the "waiting time for effect to occur" as a period in which the stated effects on pain and physical and mental ability could be changed. This perception made the respondent less obsessed with the stated treatment effects because she felt certain that she would be able to change these during the stated time period. The respondent almost perceived a longer waiting time to be a rather positive thing if she was not content with the stated effects.

3.3.2 Levels

Generally the respondents found the levels feasible and with a range not unacceptable but rather as one forcing them to make trade-offs. In more detail, the respondents reacted well to the assigned levels to the attributes of pain and physical and mental ability, providing directions of development with their status quo as the basis. The intuitiveness of taking a starting point in the current situation and imagining changes from there seemed to work well, and the respondents had good understanding. It was also clear that the number of assigned levels should be limited because the cognitive

ability of the respondents was limited, and the assignment of making choices had already proven to be difficult to grasp. Even though it was clear from the literature, the fieldwork and the interviews with clinicians that a certain treatment effect could be promised and that worsening of the condition could happen, regardless of treatment modality, the respondents did find it unrealistic that the levels included a worsening from their present state. A design challenge of how to balance realism with what was perceived as meaningful for the respondents thus became apparent. All respondents agreed that their present state was extremely unpleasant and that the outlook of nothing changing from that state was considered to be very unsatisfying. Along with the perception of unrealism that resulted in dominating levels, and although clinically relevant, the levels reflecting a worsening of an attribute affected the willingness of the patients to make trade-offs in a way that pointed towards excluding them.

3.3.3 Labelling

The qualitative pilot testing included both a labelled and an unlabeled version of the proposed DCE. All respondents understood and accepted both the labelled and unlabeled part of the questionnaire. The respondents all agreed that the two versions represented similar choices after having responded to both.

However, the respondents, and one respondent in particular, had already formed strong preferences for a specific treatment modality (surgery) and appeared unwilling to make trade-offs unless a dominant alternative was presented. The respondent noted that she had answered the unlabeled part more truthfully because she was willing to engage in any treatment that secured her effect but was blinded and heavily guided by her mind focusing only on surgery. This suggested that the treatment modality could be treated as an attribute instead of a label.

The respondents all mistakenly perceived the choice of non-surgical treatment as being old-fashioned, implying that an adjunct to the framing of the questionnaire would be useful.

3.3.4 Framing

All respondents expressed that the framing of the questionnaire was clear and concise. They were also able to explain the task ahead from reading the text only. Nevertheless, none of the respondents were able to perform the task when it was given to them, and they needed much more explanation and help to get started.

Another issue became apparent when evaluating the task because the respondents implied that some of the information about the alternatives was new to them and that they should have already been given the explanation from a doctor. One responded that she was surprised that either treatment modality could result in the worsening of her pain; she had perceived surgery to be a more lasting treatment with less possibility of relapse. The text, therefore, left her with uncertainty and questions. This was of ethical concern and pointed to a need to change the explicitness of the text explaining the attributes to make it not resemble patient information in any way and to underline the "gaming" nature of a DCE and the fact that the scenarios are hypothetical, although plausible.

It also seems that the when implementing the DCE on a larger scale, the timing will be of great importance to ensure that the experiment is not mingling with the work of and information from the doctors but is still given before a choice of treatment modality has already been made.

3.3.5 Opt-out option

The proposed design included an option to opt out, in other words, to not choose any of the proposed treatments. None of the respondents chose this option regardless of the included levels or alternatives.

The respondents described that they perceived the opt-out option as a choice of another treatment than the ones proposed. However, the crucial wish to start treatment right away, almost regardless of results, made the questionnaire resemble a forced choice and led us to propose a reframing of the opt-out option.

3.3.6 Conclusions from phase 3 to pursue in quantitative pilot testing

Overall, the proposed design proved to be too complicated and in need of a much more thorough framing and explanation of each attribute and choice task. It included too many attributes, of which some were related. It had levels that were unrealistic to the patient, although clinically relevant, and it had the appearance of a forced choice, despite including an opt-out option. The thorough qualitative work and testing pointed to a need for a final revision of the design before its application in a quantitative pilot test. The resulting proposal is illustrated in Table 4.

Table 4. Third proposal for attributes and levels based on the think-aloud pilot test with patients

Attribute	Levels
Pain at the end of treatment	Same as today, less than now, no pain
Other symptoms at the end of treatment	Same as today, decreased, none
Waiting time for the given treatment effect to occur	1/3/5 months
Risk of relapse to present state	10 %, 20 %, 30 %
Treatment modality	Surgical, non surgical

4 Discussion

The qualitative process of the present work underlines the difference between approaches (observational field work, interviews, qualitative pilot tests, etc.) and, in particular, how they are individually useful for different steps in the design process. The exploratory phase, including the literature review and observational fieldwork, contributed to the knowledge about what to look for and what to ask the doctors and patients in part two. The qualitative pilot test contributed more to the design and helped ensure the plausibility and formulations.

Generally, there is agreement that attributes and levels are best determined through some sort of qualitative work and that this work is of great importance to the validity and quality of design. However, while almost everyone uses this kind of work to determine attributes, level selection and pre-testing of whole questionnaires, the qualitative process is only reported in one third of the applied literature (De Bekker-Grob et al. 2012). It remains unclear whether this affects the validity of the experiments.

Also, it is noteworthy that few applied DCEs describe in a detailed manner how the attributes and levels were retrieved (an exception is Yi et al. 2011 and Mangham, 2007). Either this is not perceived to be of importance to report, or the work was not done in a systematic way. Consequently, some have pointed to a need for better

reporting (Coast and Horrocks 2007, Lancsar and Louviere 2008 and Coast et al. 2012).

To our knowledge, only two articles in the health economics literature have been published ahead of this one and document the effects of the qualitative process leading to a DCE design (Coast and Horrocks 2007 and Coast et al. 2012). In the paper from 2007, the authors conducted two iterations of interview series to determine relevant attributes and levels in a study of treatment choice for patients with dermatological problems. They further discussed their experienced divergence between aiming for reducing a complex situation into a limited number of attributes and using a methodology for gaining in-depth information. The approach described in this article was rather deductive, narrowing down the possible attributes while gaining more and more knowledge about the field. The gained knowledge was channelled into ensuring the quality of the questionnaires and helped to identify issues to change, consider, explain better or qualify. It became clear that the more knowledge gained, the better the possibility of choosing exactly the few correct attributes. Coast and Horrocks concluded that the qualitative process was very useful and that it ensured the quality and validity of the questionnaire produced in the end. This conclusion is perhaps not surprising, but it underlines the significance of the qualitative process and is in consensus with theory (Ryan 1999).

Recently, the same authors, with others, published a more guidance-oriented article pointing to necessary qualitative steps taken in the process of designing DCEs (Coast et al. 2012). In this work, the authors report their experiences from having conducted eight different DCEs and drew a series of recommendations for future research. However, these recommendations were mainly based on interviews rather than the full range of qualitative techniques. They suggest the stepwise design process of DCEs to be expanded to include an extra step in the process of selecting attributes. Defining attributes is argued to be a two-step process involving both a conceptual framework in which attributes are described in academic (e.g., clinical) terms and a refinement of wording makes the attributes suitable and understandable to the respondents of the experiment. The suggested two-step process is recognizable from the qualitative work described in this paper because the attributes were firstly formulated and tested based on the literature and among doctors. The authors further underlined that qualitative work is especially useful during the (re)wording process because respondents often prove to be of a very different character than academics. That was very clear in the work described in this paper, and a substantial amount of rewording - both of the actual attributes but also in the explanation of the task - was done based on the performed interviews with the patients. The authors also point to a discrepancy between the depth and complexity gained from qualitative work and the reductive use of it when informing attribute selection. The experience from the process described in this paper did not point to the same concern. On the contrary, the deeper the knowledge of the respondents became, the more trustworthy and relevant the selected attributes felt. Any excess of knowledge that might not go into the actual DCE may be relevant for designing other questions to put into a questionnaire. Finally, the paper suggests a checklist that needs to be thought of when evaluating the quality of performed interviews as a part of qualitative work.

As this paper suggests, various sources of information can be useful to inform the design process, and all steps proved useful. Hence, it would not be recommendable to focus too narrowly on one qualitative method and exclude the others.

Other research fields using DCEs may have a different tradition of reporting qualitative work ahead of designing DCEs. However, some researchers point to a

similar lack of reporting, transparency of methodology and consistency of the use of qualitative methods.

Some even suggest a development of methods to handle this by automatically constructing attributes from the wording of user reviews in marketing, arguing that the reviews made by costumers reveal both the important attributes of the good and the appropriate wording of the attributes (Lee et al. 2007). Others point to a possible use of qualitative work for debriefing respondents (Philips et al. 2001).

It is surely context-specific how much or little qualitative work is needed before designing a DCE. However, it seems clear that applying DCEs in a difficult setting, with little knowledge to gain from the literature, where the decision-making situation is complicated, and in which vulnerable persons are participating, requires thorough pre-testing and qualitative work. The process described in this paper also showed that the usability of the knowledge gained could reach a point of exhaustion. The effort should mirror the gains, and researchers should only continue the qualitative process to a degree that gives results and until the knowledge gained seems sufficient to trust in the validity of the design. It is certain that some of the qualitative tools are very time consuming, especially observational work. The various qualitative methods should be used when appropriate, and it might not be feasible to set a standard of what to include in a qualitative process because it will depend on the setting and prior knowledge.

On a more practical level, the qualitative work process can also affect the feasibility of a DCE. In a context like the one described in this paper and in similar contexts in which it can be difficult to perform choice experiments, the qualitative process can work both as an essential source of valuable information but also as a mean of ensuring the inclusion of stakeholders and key-persons in an early stage (Mangham et al. 2009 and Baltussen et al. 2006). The qualitative process included some of the gatekeepers of the relevant hospital departments very early on, and that helped to ensure the success of further collaboration on a quantitative data collection with the DCE. The qualitative testing also worked as a pilot for the patient inclusion process. Since doors often need to be opened and agreements be made in advance of a DCE, it seems obvious to take advantage of this process to inform the experiment in a structured way.

Researchers performing DCEs in health care settings are not used to always having their results directly affect policy making because this is still mainly based on cost-utility measures, whereas applied DCEs in many other research fields are directly influencing policies (Goldman et al. 2010). Making sure that the end users (in this case, hospital departments) are incorporated into the entire process and that the experiment truly reflects the real decision task for patients and doctors might help to make the way for this type of analysis.

4.1 Limitations

Our process did have some limitations. A relatively small number of patient interviews were performed in phase three, which could have led to important attributes not being identified or to the levelling not being defined in a meaningful way for all respondents. The respondents did represent different ages, both genders and had different experiences with treatment paths and, hence, a range of views; despite these differences, the respondents pointed to similar issues with the design during interviews. Mis-specifying attributes based on the qualitative work is possible; however, the thoroughness and thoughtfulness that went into the collected process,

including all the described steps, leaves little doubt that the chance of mis-specifying attributes or levels is smaller than if the process had been limited or based on expert-opinions or researchers opinions. Furthermore, it should be noted that the qualitative design process should lead to a qualified design proposal but cannot supersede a quantitative pilot testing, which provides the final test of the exhaustiveness of the attributes and the appropriateness of levelling. The process seems to have provided enough knowledge to trust in the choice of the attributes and levels because they have been accepted by both doctors and patients within this setting. Since every step contributes to the development of the design, having only done one or two of the steps may have resulted in a less valid design. The thorough qualitative work might even influence response rates or the meticulousness of responses because the understanding of the experiment for respondents has been tested.

5 Conclusion

Systematic and thorough qualitative investigation of the decision context, relevant attributes and levels, and the appropriate framing appear valuable in the process of designing a DCE for quantitative pilot testing. In the present work, we demonstrated how each step of a design process, including different qualitative techniques, significantly impacted and guided the design. The initial phase was based on a literature review and observational fieldwork and led to the identification of seven overall attributes that were perceived to be important to patient preferences for the treatment of DDD. The second phase included interviews with doctors and helped to ensure the realism of the questionnaire. The doctors pointed out changes in both attributes and, especially, levels, and the decision situation in which the patients found themselves was described and qualified. The third phase was based on in-depth interviews with patients and pointed to several changes. Overall, the proposed design proved to be too complicated and complex and in need of a much more thorough framing and explanation of each attribute and choice task. It included too many attributes, of which some were related. It had levels that were unrealistic to the patient, despite being clinically relevant, and it had the appearance of a forced choice, despite including an opt-out option.

Collectively, all the steps or phases provided change and optimization of the design, and, although a small number of patient interviews were performed, the concurrent nature of their responses and the effect of the entire process seem to have provided the basis for the ability to trust in the choice of the attributes and levels.

Most important, this study demonstrated that important understanding - with direct influence on the final design proposal - would have been missed if the qualitative process had been restricted to just one technique, e.g., literature review or patient interviews. The different techniques appear to each have their own rationale in the qualitative design process. It could be useful for other researchers to follow some or all of the applied steps, depending on their pre-knowledge and the context of implementation. As the qualitative process directly influences the validity of the design, an improved reporting of the process seems to be warranted in future studies.

Acknowledgements

Thanks should be directed to Benny Dahl of the Orthopaedic department, Rigshospitalet and Tom Bendix and staff at Rheumatology, Glostrup and to patients who volunteered to be either observed or interviewed. The research was financed by CeSpine. The study was part of the interdisciplinary research project CESpine

(www.cespine.dk), which is financed by the Danish Strategic Research Council (grant 2142-08-0017)

References

Allen, T., J. F., Rihn, S. D., Glassman, B. L., Currier , T. J., Albert and F. M., Phillips, 2009. An Evidence-Based Approach to Spine Surgery. *American Journal of Medical Quality*, 24(6 Suppl.), 15S-24S.

Ashcroft, D.M., E., Seston C. E., Griffiths, 2006. Trade-offs between the benefits and risks of drug treatment for psoriasis: a discrete choice experiment with U.K. dermatologists. *British Journal of Dermatology*, 155(6), 1236–1241.

Baltussen, R. And L., Niessen, 2006. Priority setting of health interventions: the need for multi-criteria decision analysis. *Cost Effectiveness and Resource Allocation*, 4(14).

Bateman, I. J., R. T., Carson, B., Day, W. M., Hanemann, N., Hanley, T., Hett, M.J., Lee, G., Loomes, S., Mourato, E., Özdemiroglu and D. W., Pearce, 2002. Economic Evaluation with Stated Preference Techniques - A Manual. 1st edition, Edward Elgar Publishing Limited, Cheltenham.

Bennett, J. and R. K., Blamey, 2001. The Choice Modelling approach to Environmental Valuation. Edward Elgar Publishing Limited, UK.

Blamey, R. K., J. W., Bennett, J. J., Louviere, M. D., Morrison, and J. C., Rolfe, 2002. Attribute causality in environmental choice modelling. *Environmental and Resource Economics*, 23(2), 167-186.

Chou, R., J., Baisden, E. J., Carragee, D. K., Resnick, W. O., Shaffer, J. D., Loeser, 2009. Surgery for Low Back Pain. A Review of the Evidence for an American Pain Society Clinical Practice Guideline. *Spine*, 34(10), 1094-1109.

Coast, J. and S., Horrocks, 2007. Developing attributes and levels for discrete choice experiments using qualitative methods. *Journal of Health Services Research and Policy*, 12(1): 25-30.

Coast, J., H., Al-Janabi, E. J., Sutton, S., Horrocks, J., Vosper, D. R., Swancutt and T., Flynn, 2012. Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations. *Health Economics*, 21(6), 730-741.

De Bekker-Grob, E. W., M., Ryan and K., Gerard, 2012. Discrete Choice Experiments in Health Economics: A Review of the Literature. *Health Economics*, 21(2), 145-172.

Goldman, D., Lakdawalla, D., Philipson, T. J. and Yin, W. (2010). Valuing health technologies at NICE: recommendations for improved incorporation of treatment value in HTA. Health Economics, 19: 1109–1116.

Green, P. E. and V., Srinivasan, 1978. Conjoint Analysis in Consumer Research - Issues and Outlook 2. *Journal of consumer research*, 5(2), 103-123.

Hall, J., R., Viney, M., Haas and J. J., Louviere, 2004. Using stated preference discrete choice modeling to evaluate health care programs. *Journal of Business Research*, 57(9), 1026-1032.

Hensher, D. A., J. M., Rose and W. H., Greene, 2005. Applied choice analysis: a primer. Cambridge: Cambridge University Press.

Irwin, Z. N., A., Hilibrand and M., Gustavel, 2005. Variation in surgical decision making for degenerative spinal disorders. Spine.1;30(19):2208-13.

Kjaer T., D, Gyrd-Hansen and I., Willaing, 2006. Investigating patients' preferences for cardiac rehabilitation in Denmark. *International Journal of Technology Assessment in Health Care*, 22(2), 211–218.

Kløjgaard, M., M., Bech, B., Dahl and R., Søgaard, *forthcoming*. Preferences for Spine Surgery –a Systematic Literature Review.

Kuper, A., L., Lingard and W. Levinson, 2008. Critically appraising qualitative research. *British Medical Journal*, 337(7671).

Lancsar, E. And J., Louviere, 2006. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? *Health Economics*, 15(8), 797-811.

Lancsar, E. and J., Louviere, 2008. Conducting discrete choice experiments to inform health care decision making: a user's guide. *Pharmacoeconomics*, 26(8), 661-677.

Lee, Y. T. and E. T., Bradlow, 2007. Automatic construction of conjoint attributes and levels from online consumers reviews. University of Pennsylvania. The Warton School.

Mangham, L., 2007. Addressing the Human Resource Crisis in Malawi's Health Sector: Employment preferences of public sector registered nurses, ESAU Working Paper 18.

Lloyd, A., E., McIntosh, K. F., Rabe, A., Williams, 2007. Patient preferences for asthma therapy: a discrete choice experiment. *Primary Care Respiratory Journal*, 16(4), 241–248.

Louviere, J., D. A., Hensher and J., Swait, 2000. Stated Choice Methods, analysis and application. Cambridge University Press, U.K.

Mangham, L. J., K., Hanson and B., McPake, 2009. How to do (or not to do) … Designing a discrete choice experiment for application in a low-income country. *Health Policy and Planning*, 24(2), 151-158.

Mark, T. L. And J., Swait, 2003. Using stated preference modeling to forecast the effect of medication attributes on prescriptions of alcoholism medications. *Value in Health*, 6(4), 474–482.

Mays, N. And C., Pope, 2000. Assessing quality in qualitative research. *British Medical Journal*, 320(7226), 50-52.

Mortimer, D. and L., Segal, 2008. Is the value of a life or life-year saved context specific? Further evidence from a discrete choice experiment. *Cost Effectiveness and Resource Allocation*, 6(8).

Peters, E., D., Vastfjall, P., Slovic, C. K., Mertz, K., Mazzocco and S., Dickert, 2006.Numeracy and decision making. *Psychological Science*, 17(5), 407-413.

Philips, K. A., F. R., Johnson and T., Maddala 2001. Measuring what people value: A comparison of "attitude" and "preference" surveys. *Health Services Research*, 37(6), 1659–1679.

Phillips, K.A., T., Maddala and F. R., Johnson, 2002. Measuring preferences for health care interventions using conjoint analysis: an application to HIV testing. *Health Services Research*, 37(6), 1681–1705.

Ratcliffe, J., M., Buxton, T., McGarry, R, Sheldon and J. Chancellor, 2004. Patients' preferences for characteristics associated with treatments for osteoarthritis. *Rheumatology* (*Oxford*), 43(3), 337–345.

Ratcliffe, J., R., Van Haselen, M., Buxton, K., Hardy, J., Colehan and M., Partridge, 2002. Assessing patients' preferences for characteristics associated with homeopathic and conventional treatment of asthma: a conjoint analysis study. *Thorax*, 57(6), 503–508.

Ratcliffe, J. and L., Longworth, 2002. Investigating the structural reliability of a discrete choice experiment within health technology assessment. *International Journal of Technology Assessment in Health Care*, 18(1), 139-144.

Roux, L., C., Ubach, C., Donaldson, and M., Ryan, 2004. Valuing the benefits of weight loss programs: an application of the discrete choice experiment. *Obesity Research*, 12(8), 1342–1351.

Ryan, M. and K., Gerard, 2003. Using discrete choice experiments to value health care: current practice and future prospects. *Applied Health Economics and Health Policy*, 2(1), 55–64.

Ryan, M., 1999. A role for conjoint analysis in technology assessment in health care? *International Journal of Technology Assessment in Health Care*, 15(3), 443-457.

Salkeld, G., M., Solomon, L., Short, M., Ryan and J. E., Ward 2003. Evidence-based consumer choice: a case study in colorectal cancer screening. *Australian and New Zealand Journal of Public Health*, 27(4), 449–455.

Sculpher, M., S., Bryan, P., Fry, P., de Winter, H., Payne and M.. Emberton, 2004. Patients' preferences for the management of non-metastatic prostate cancer: discrete choice experiment. *British Medical Journal*, 328(7436): 382.

Seston, E.M., R. A., Elliott, P. R., Noyce and K., Payne, 2007. Women's preferences for the provision of emergency hormonal contraception services. *Pharmacy World and Science*, 29(3), 183–189.

Shackley P, R., Slack and J., Michaels, 2001. Vascular patients' preferences for local treatment: an application of conjoint analysis. *Journal of Health Services Research and Policy*, 6(3), 151–157.

Smith H.E., J. F., Rihn, D. S., Brodke, R., Guyer, D., Coric, B., Lonner, A.P., Shelokov, B. L., Currier, L., Riley, F. M., Phillips and T. J., Albtert, 2009. Spine Care: Evaluation of the Efficacy and Cost of Emerging Technology. *American Journal of Medical Quality*. 24(6 Suppl), 25S-31S.

Telser, H. and P., Zweifel, 2002. Measuring willingness-to-pay for risk reduction: an application of conjoint analysis. *Health Economics*, 11(2), 129–139.

Yi. D., M., Ryan, S., Campbell, A., Elliott, N., Torrance, A., Chambers, M., Johnston, P., Hannaford and B. H., Smith, 2011. Using discrete choice experiments to inform randomised controlled trials: an application to chronic low back pain management in primary care. *European Journal of Pain*, 15(5), 531.e1-531.e10.