To what extent do existing measures of shared decision-making incorporate assessment of deliberation by the patient?

Submitted for the degree of Master of Philosophy

2016

Sara Southall

990360366

Table of Contents

Summary

Introduction

To research and develop robust, effective shared-decision support requires a clear definition of the decision-making process. Elwyn and Miron-Shatz have developed a model incorporating deliberation, but it is uncertain the extent this proposed decision process is addressed in existing measures of decision-making. This review aims to systematically identify and analyse measures considering the process from the patient's perspective, determining the extent items cover a decision process model incorporating deliberation.

Method

A systematic and robust search strategy was developed to identify studies reporting the development and psychometric evaluation of an instrument measuring shared decision-making in the health-care setting, specifically from the patient's perspective of the decision making process. The search strategy incorporated seven electronic databases supplemented by manual searches. As well as descriptive details of the study, scales were mapped against the decision process model, and the methodological quality of study design and the adequacy of instrument measurement properties were appraised using reproducible standards.

Results

Of the 7,563 references identified, ten studies involving the development of measures were included and an additional nine addressed the further evaluation of instruments. The methodological quality was highly variable, both between and within studies and none of the scales scored consistently

well for measurement properties when compared with the criteria of adequacy. One scale was found to map against all stages of deliberation, but none addressed all stages of the decision process model.

Conclusion

The results indicate that current measures of decision-making do not consider all steps of the decision process proposed by Elwyn and Miron-Shatz. The Decision Making Quality Scale (DMQS), developed by Hollen and based on the work of Janis and Mann, addressed aspects of deliberation not covered by other measures. Recommendations for a new instrument specification include methodological improvement and consideration of the work of Janis and Mann.
Word count 297

Declaration/Statements

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed …… (candidate) Date ……

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of MPhil.

Signed …… (candidate) Date ……..

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated.
Other sources are acknowledged by explicit references. The views expressed are my own.

Signed …… (candidate) Date ……

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed …… (candidate) Date ……

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loans after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.

Signed …… (candidate) Date ……

Acknowledgements

I would like to thank:

Dr. Sharon Mayor for her supervision, expertise and support.

Prof. Glyn Elwyn for his expertise and the original concept on which this work is based.

Ms. Mala Mann and Mr. Rowland Summers for their guidance in developing a systematic review search strategy.

Dr. Kate Brain and Dr. Meirion Evans for their guidance and encouragement at each annual appraisal.

Dr. Natalie Joseph-Williams and colleagues in the Decision Laboratory at the Institute of Primary Care and Public Health, Cardiff University for their feedback.

Mrs. Judith Mills, Dr. Mark Limmer, Dr. Richard Jarvis and Dr. Rachel Isba from the North West Public Health specialty training scheme for their encouragement and understanding.

Word Count

39,550 excluding appendices and footnotes

List of figures

List of tables

List of appendices

List of abbreviations

CCT	Classical Test Theory
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
DAS	Decision Attitude Scale
DCS	Decisional Conflict Scale
DECS	Decision Emotional Control Scale
DES	Decision Evaluation Scale
DMQS	Decision-making Quality Scale
DSES	Decision Self-Efficacy Scale
HRT	Hormone replacement therapy
HSR	Health Status Restriction
ICC	Intraclass correlation
IRT	Item Response Theory
MeSH	Medical subject headings
MIC	Minimal important change
MID	Minimal important difference
PrepDM	Preparation for Decision-Making scale
RCT	Randomised control trial
RTI	Respiratory tract infection
SDC	Smallest detectable change
SDM	Shared decision-making
SDM-Q	Shared Decision-Making Questionnaire
SWD	Satisfaction with Decision scale

1. Introduction

This chapter first describes shared decision-making and provides context for the rapid growth in this field, before summarising current definitions of a good decision and decision-making process, with consideration of deliberation as a new approach for evaluation. Measurement instruments are described and the recommended steps in their development summarised, along with the process of systematic review. Finally, all strands are brought together with the research question, project aims and objectives

1.1. Shared decision-making

1.1.1. Defining shared decision-making

Shared decision-making is a rapidly growing field with a similarly evolving definition. (Charles et al., 1997) It is often considered as a middle ground between the extremes of paternalistic health care, where the clinician makes all decisions, and informed choice, which is led by the patient. (Collings and Coulter, 2015) Shared decision-making, in contrast, sees the management plan developed by both the practitioner and the patient, acknowledging the expertise and differing perspectives brought by both parties with subsequent decisions incorporating clinical evidence and the unique context of the individual patient. (Collings and Coulter, 2015, Makoul and Clayman, 2006) It has been described by Al Mulley as an approach that allows patients to get "the care they need and no less, and the care they want and no more." (Collings and Coulter, 2015)

In 1997, Charles and colleagues proposed fixed factors needed for shared decision-making to occur: that the interaction involves at least two participants

with both taking steps to participate, where sharing information is a pre-requisite and a treatment decision is made with which both parties agree. (Charles et al., 1997) This became, for a time, the most cited definition of the SDM concept. (Makoul and Clayman, 2006)

Within ten years, Makoul and Clayman performed a systematic review to identify elements frequently used to define shared decision-making and used these to develop a "clinically sound and conceptually relevant" model. They outlined nine essential components, as listed in the table below. (Makoul and Clayman, 2006) While this definition builds on that of Charles and colleagues by providing greater detail, it arguably also fails to address the more nuanced and challenging features of shared decision-making in healthcare, where decisions may be stressful, weighted with potentially life-changing consequences and requiring both support and, for some, periods of reflection. This limits its use as a framework in practical contexts.

Table 1: Makoul and Clayman, 2006 – defining shared decision-making

Makoul and Clayman's definition of SDM
Define/explain problem
Present options
Discuss pros/cons (benefits/risks/costs)
Patient values/preferences
Discuss patient ability/self-efficacy
Doctor knowledge/recommendations
Check/clarify understanding
Make or explicitly defer decision
Arrange follow-up

In 2012, a model of shared decision-making developed by Elwyn and colleagues reflected a growing appreciation of the complexity of decision-making, grounded in the practicalities of clinical practice. Deliberation is identified as a component of shared decision-making, and defined as a "process of considering information about the pros and cons of [..] options, to assess their implications, and to consider a range of possible futures, practical as well as emotional." (Elwyn et al., 2012) The model acknowledges the "psychological, emotional and social factors" involved, along with the contribution of other parties and reflection away from the healthcare encounter. (Elwyn et al., 2012) Patient-clinician interactions are grouped into option, choice and decision talk, with preferences passing from initial to informed as they are expressed and explored. Deliberation runs through the entirety of the process and, based on identified need, the decision support provided can range from brief to in-depth. (Elwyn et al., 2012)

1.1.2. The growth of shared decision-making

Since the early seventies, interest in shared decision-making has developed with each decade. Mentioned in 6 references in Medline in the 1970s, this number has increased to 1542 since 2010. It is now the subject of training and decision support aids, and has also featured in the policies and practice of governments and professional regulation governing bodies, such as the General Medical Council. (Collings and Coulter, 2015)

A key factor encouraging this growth is the increasing acknowledgment of autonomy, an individual's right to self-determine. (Elwyn et al., 2012) With greater access to information and appraisal of healthcare quality, the balance of power between patient and practitioner has shifted. (Charles et al.) Advances

in healthcare have increased life expectancy, the prevalence of long-term illness and the treatment options available. (Légaré et al., 2010b) The latter, in combination with a growing evidence base, has also introduced "close-call" choices into management plans, where no clear clinical benefit exists between treatments. (O'Connor et al., 2009) Taken in combination, the input of the patient to healthcare decisions is increasingly essential and likely to develop further in future as healthcare options increase and related decisions become more challenging. (Légaré et al., 2010b)

1.1.3. Shared decision-making in current healthcare practice

Shared decision-making is relevant in a broad range of healthcare contexts from acute minor illnesses to major life-changing healthcare decisions, and personal treatment plans for chronic illness to screening and diagnostic tests. (Collings and Coulter, 2015) There may be fewer contexts where SDM is less suitable such as, for example, acute trauma or emergency surgery.

Decision aids have been developed to support different facets of shared-decision making for an extensive number of health conditions. (O'Connor et al., 2009) Aids have been defined by O'Connor and colleagues as "evidence-based tools designed to help clients participate in making specific and deliberated choices among healthcare options in ways they prefer." (O'Connor et al., 2009) Decision aid development has explored ways of communicating information about health conditions, available treatment options, outcome probabilities, risks and uncertainties as well as exploring patient views and preferences. (Collings and Coulter, 2015)

In recent Cochrane Collaboration systematic reviews, decision aids were found to improve patient-practitioner communication and facilitate more informed, values–based choices. (Stacey et al., 2014). While no negative impact on health outcomes or satisfaction was apparent in these reviews, the use of a decision aid was found to increase congruence between treatment choice and the explored values of the patient, with reduced numbers opting for elective surgery, prostate specific antigen screening and menopausal hormone therapy. (O'Connor et al., 2009, Stacey et al., 2014).

Interventions aimed at facilitating shared-decision making may also carry wider implications for population-level healthcare planning, where commissioning aims to deliver healthcare that is appropriate to population needs while also an equitable and effective use of limited resources. (Collings and Coulter, 2015) Collings and Coulter argue that every shared management plan should be considered "micro-commissioning" and, with improved recording of shared decision-making, it could be possible to gather and summarise information to feed back into wider commissioning decisions, producing more responsive strategies. (Collings and Coulter, 2015)

The importance of shared decision-making is likely to grow alongside the increasing evidence-base, treatment options and an ageing population, while the ethical imperative to involve patients in decisions about their lives will remain relevant. (Elwyn et al., 2012, Charles et al., Légaré et al., 2010b) However, there are barriers to the incorporation of shared decision-making as a routine element of healthcare practice, such as time pressure, practitioner access to skills and training, and patient awareness and support. (Agoritsas et al., 2015) In addition, the evidence base must become more established and detailed to ensure that shared decision-making can be defined, measured and

documented validly and reliably, and that interventions aimed at the practitioner-patient relationship are effective and safe. (Collings and Coulter, 2015) As well as evaluating the effect of such interventions, their mechanism of action should also be well understood. As such, what constitutes "good" decisions and decision-making in healthcare must be established. (Makoul and Clayman, 2006)

1.2. Measuring decision-making

1.2.1. Defining a good decision

How best to define a good decision has been widely debated across many fields. In healthcare, a good clinical decision has been described as one that chooses the “best course of action given current scientific evidence, healthcare resources, clinical circumstances and patient preferences”. (Légaré et al., 2010b) While this acknowledges aspects of shared decision-making, such as patient preference and existing evidence, it misses the collaborative aspect of the clinician and patient contributing to the consultation. Several authors have contributed seminal work exploring good decisions and decision-making with relevance to shared decision-making.

In 1997, O’Connor and colleagues asked physicians to select criteria for evaluating decision aids. The actual decision made was the least supported marker of good decision-making. Instead, understanding options and their respective risks and benefits, clarity of decision trade-offs, accuracy of expectations, anxiety, decision commitment, uncertainty and satisfaction were ranked higher. (O'Connor et al., 1997) A further study in 2003 identified clarity about values along with selecting and implementing a choice consistent with these values as essential criteria for a satisfactory decision. (O'Connor et

al., 2003) Similar themes are seen in the Ottawa Decision Support Framework, where decision quality is identified as an informed, values-based enacted decision without regret or blame. (O'Connor, 2006)

O'Connor and O'Brien-Pallas' definition of an effective decision, that "an effective decision is one that is informed, consistent with decision-maker's values and behaviourally implemented", was amended by Marteau and colleagues in 2001. (Marteau et al., 2001) This work, which aimed to develop a measure of informed choice, expanded on the term "informed" in the above definition, suggesting that a good decision is based on relevant and reliable information. (Marteau et al., 2001) In addition, a review team considering decision-making in the context of community healthcare and cancer screening highlighted that, in addition to accuracy of knowledge and value concordance, interventions aimed at decision-making should also facilitate participation at a level desired by the individual. (Briss et al., 2004)

Sepucha and colleagues suggested that to assess decision quality, three sets of information are needed: decision-specific knowledge evaluated with a set of knowledge questions, patient values for relevant outcomes elicited with value-scaling tasks and a correlation between the treatment chosen and the values identified. (Sepucha et al., 2004, Sepucha et al., 2007) It was further noted that patient satisfaction alone is an insufficient marker of decision quality as it is influenced by initial expectations. (Sepucha et al., 2007) As such, in a subsequent study, constructs of decision quality were identified as a choice that is based on realistic expectations as well as relevant knowledge, and demonstrates values-choice agreement. (Sepucha et al., 2013)

The International Patient Decision Aid Standards (IPDAS) Collaboration,

through a Delphi process reported in 2005, suggested aspects of a good decision and decision process. For the decision, this was indicated by a match "between features which matter most to an informed patient and the chosen option". (O'Connor et al., 2005, O'Connor et al., 2009) For the decision process, elements included recognising that a decision needs to be made, knowing the options available and what each involves, appreciating values influencing the decision, being clear which features matter most, discussing values with the health practitioner and involvement in the decision-making being in line with individual preference. (O'Connor et al., 2005, O'Connor et al., 2009)

As noted by Elwyn and Miron-Shatz, patterns emerge in the existing definitions of good decisions and decision-making processes. (Elwyn and Miron-Shatz, 2010) These can be described by the order in which they occur in the decision process. (Scholl et al., 2011) Firstly, there are decision antecedents such as considering the extent a patient wishes to be involved in a decision. Next are elements incorporated in the decision process itself, for example eliciting values and evaluating knowledge sufficiency. Lastly, there is the ability to commit to a decision and subsequent outcomes such as value-concordance, satisfaction or regret. (Scholl et al., 2011, Elwyn and Miron-Shatz, 2010)

1.2.2. Limitations of current approaches

While there are common themes in the definitions of decision and decision-making process quality, there are also variations. A review by Sepucha and colleagues in 2013 noted a lack of consensus in the approaches taken to measuring decision-making, which limits accurate evaluation of decision aids

and other interventions aimed at improving shared decision-making. (Makoul and Clayman, 2006) In addition, this makes comparisons and meta-analyses of intervention studies difficult, if not impossible. (Gopalakrishnan and Ganeshkumar, 2013)

The work of Elwyn and Miron-Shatz further critique the existing definitions. They suggest that even if the decision process was sound, probability influences the subsequent outcome, and how the decision is perceived may change with time. (Elwyn and Miron-Shatz, 2010) As such, the decision-making process, termed deliberation, and decision made or determination, should be considered separately. (Elwyn and Miron-Shatz, 2010) Further, the authors highlight limitations with using knowledge and preferences alone as indicators of a good decision-making process. Knowledge decays and the type of knowledge needed - what and how much - is debatable. In addition, preferences are fluid and change with time. (Elwyn and Miron-Shatz, 2010) They can seem illogical and are influenced by context, experiences and simply the phrasing of questions. (Slovic, 1995, O'Connor et al., 1996)

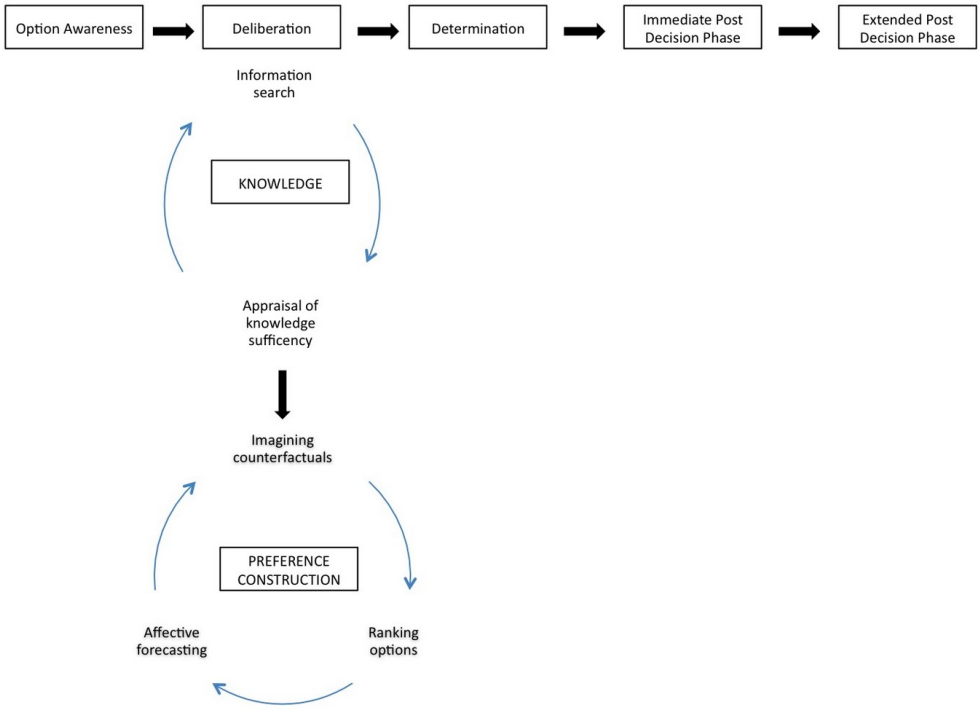
Furthermore, the ideal parameters for knowledge or preferences may be misleading when considering that healthcare decisions are made without infinite time or resources. (Elwyn et al., 2001a) Heuristics, described by Gigerenzer and Gaissmaier as “strategies where some information is ignored in order to make decisions quickly or efficiently in comparison to complex methods”, are at times as effective despite seeming less rational or considered than other decision processes. (Gigerenzer and Gaissmaier, 2011, Goldstein and Gigerenzer, 2009) However, it remains important that decision processes are outlined and understood so that they can be explored further. (Gigerenzer and Gaissmaier, 2011)

1.3. Deliberation and determination

1.3.1. A decisional process map incorporating deliberation

In response to the limitations described above, Elwyn and Miron-Shatz proposed a new decision-process model incorporating deliberation. The model is described as "combining cognitive and emotional contributions" to the decision process, and complements the shared decision-making model developed by Elwyn and colleagues in 2012. (Elwyn and Miron-Shatz, 2010, Elwyn et al., 2012) The stages begin with an awareness that a choice needs to be made. Information searching and knowledge appraisal follow, with preferences constructed through imagining possible outcomes or counterfactuals, and affective forecasting where feelings about different futures are reviewed. The different options are ranked before progressing to determination (Elwyn and Miron-Shatz, 2010) and the model is summarised in Figure 1 below.

Figure 1: Model of decision-making process incorporating deliberation



1.3.2. Practical implications

Elwyn and Edwards described a "black box" of decision-making, where a limited understanding of the decision-making process creates uncertainty over how decision support interventions work, in turn inhibiting further progress. (Edwards and Elwyn, 2006) The deliberation model proposed by Elwyn and Miron-Shatz is a step towards clarifying these elements. The next is to establish whether the patient experience of deliberation is considered in current measures of shared decision-making, as the ability to measure patient deliberation using such tools would facilitate further exploration of decision support interventions and their mechanisms of action. (Scholl et al., 2011, Elwyn and Miron-Shatz, 2010)

1.4. Measurement instruments

1.4.1. What is a measure?

The Scientific Advisory Committee defines measurement scales as a "constellation of items contained in questionnaires and interview schedules, along with their instructions…" while Streiner and Norman consider them an "essential component of scientific research". (Scientific Advisory Committee of the Medical Outcomes Trust, 2002, Streiner and Norman, 2008) While traditionally concerned with outcomes of healthcare such as survival, hospital readmission, laboratory and radiographic tests; due to increasing life expectancy and treatment options, there is growing focus on aspects previously considered immeasurable and subjective, such as quality of life and the patient experience. (Streiner and Norman, 2008, Valderas et al., 2008) Information on the acceptability and appropriateness of healthcare to the patient now shape provision of care from individual management plans through to wider service design. (Valderas et al., 2008, Health Knowledge, 2011)

1.4.2. The importance of quality

A broad range of measures reflecting different aspects of healthcare exist but their method of design and quality have also been noted to vary widely. (Scientific Advisory Committee of the Medical Outcomes Trust, 2002) The quality of an instrument's development process and measurement properties are of considerable importance. Studies have indicated a bias towards interventions in comparison with controls where unpublished scales were used in studies of mental health treatment. (Marshall et al., 2000, Lockwood and Marshall, 1998) Another study highlights how the phrasing of questions may influence the perception of treatment side effects and subsequent work absenteeism. (O'Connor et al., 1996) As such, measures lacking rigor in their

development and with poor measurement property performance become a source of inaccuracy in research with significant consequences in the field of healthcare. (Mokkink et al., 2010b, Terwee et al., 2012)

As such, instruments must be demonstrated to be valid and reliable, measuring what they are intend to measure and doing so in a reproducible manner. (Streiner and Norman, 2008) The development and measurement properties of an instrument should therefore be systematically evaluated in a robust and transparent manner, as done routinely in other areas of healthcare research. (Higgins and Green, 2011, Johnston and Graves, 2008, Mokkink et al., 2010b) However, Johnston and Graves note that the most well-used measure is often selected without consideration of "how valid or how reliable [it is] or in what ways", with the measurement property terms "used as synonymous with good, rather than reflecting the quality of a measure for a particular construct or application." (Johnston and Graves, 2008) As such, before a measure is used in research or clinical practice, its design and measurement properties should be evaluated using clear standards of quality. (Mokkink et al., 2010b)

1.4.3. Developing and evaluating measures

As highlighted by Streiner and Norman, the development of measurement scales is challenging and requires resources such as time and money, along with access to experience and expertise. As such, the recommended first step is to review existing instruments to avoid the unnecessary duplication of efforts in designing another. (Streiner and Norman, 2008) The development process and measurement properties of identified instruments should then be evaluated. (Johnston and Graves, 2008, Mokkink et al., 2010b) Until recently, textbooks and detailed guides were used to develop and review measures, which could be time consuming, inaccessible and hard to synthesise succinctly.

(Streiner and Norman, 2008, Johnston and Graves, 2008, Mokkink et al., 2010b) This, along with the varying terminology used, has arguably contributed to the lack of standardised information available about measures, further impairing appraisal, comparison and choice. (Johnston and Graves, 2008)

1.4.4. Standards for measure development

The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative has developed a checklist of quality standards for measure development. This was produced using a literature review followed by a multidisciplinary, international Delphi process to determine which measurement properties to include and what standards to apply to study design and statistical analysis. (Mokkink et al., 2010b, Mokkink et al., 2012) The checklist evaluates ten measurement properties, each with between five and eighteen items to assess whether a study meets a good methodological standard. Internal consistency, reliability, measurement error, content validity, construct validity, criterion validity and responsiveness are covered, along with interpretability, item response theory and generalisability. (Terwee et al., 2012, Mokkink et al., 2012) A scoring version was subsequently developed to facilitate the comparison of instruments in systematic reviews. This contains four possible responses per item (excellent, good, fair and poor), with an overall score for a measurement property derived from the lowest score obtained for that property. (Terwee et al., 2012, Schellingerhout et al., 2012) A summary of the parameters considered by the COSMIN checklist with their definitions is included in Appendix 1. A sample of the checklist is included in Appendix 2.

1.4.5. Criteria for measurement properties

Terwee and colleagues suggest "the assessment of methodological quality of a study and the assessment of the quality of an instrument are fundamentally different things, and should be performed separately in systematic reviews." (Terwee et al., 2012) Using the property of internal consistency as an example, the methodological standards consider features such as sample size, checking for scale unidimensionality and the use of appropriate statistical analysis in scale development, while the statistical value for item intercorrelation is considered under measurement property quality criteria. (Higgins and Green, 2011, Terwee et al., 2007)

Several quality criteria for scale measurement properties exist, such as the frequently cited Scientific Advisory Committee of Medical Outcomes Trust criteria. (Terwee et al., 2007) This considers eight attributes of the measure, including its conceptual and measurement model, reliability, validity, responsiveness, interpretability, respondent and administrative burden, alternative forms, cultural and language adaptations, along with specific review criteria. (Scientific Advisory Committee of the Medical Outcomes Trust, 2002) Further studies have clarified such criteria in order to make them more functional and easy to apply in practice. (Valderas et al., 2008, Terwee et al., 2007) These include work by the authors of the COSMIN checklist, whose criteria have subsequently been used in combination with the COSMIN checklist and are detailed in Appendix 3. (Terwee et al., 2007, Schellingerhout et al., 2012)

1.5. Systematic reviews

1.5.1. Overview

A systematic review is a method of finding and summarising all available evidence to answer a clearly defined question. (Liberati et al., 2009, Shea et al., 2007) The findings can be further summarised using statistical processes in meta-analyses. (Gopalakrishnan and Ganeshkumar, 2013) Systematic reviews are used for keeping up to date with the evidence base and guiding treatment decisions, policy development, guidelines and funding allocation. (Liberati et al., 2009)

The first step of a systematic review is to identify a research question with pre-agreed inclusion and exclusion criteria for which evidence is to be consulted. A transparent and reproducible methodology is then developed to systematically search for sources of information meeting the eligibility criteria. The resources found are critically appraised using pre-set criteria to assess the validity of the evidence provided. The findings and strength of evidence are then synthesised to formulate conclusions. (Shea et al., 2007, (Khan, 2005, Higgins and Green, 2011)

Systematic reviews are placed highest in the hierarchy of research evidence due to the increased ability to detect meaningful results gained by combining sources of information and different studies. (Hemingway and Brereton, 2009) This increases the reliability of the findings in comparison with other study designs. (Gopalakrishnan and Ganeshkumar, 2013) However, systematic reviews must use a robust, transparent method that is clearly reported in order to facilitate evaluation of possible sources of bias. (Higgins and Green, 2011) These include publication bias, where studies with positive results are more likely to be found in peer-reviewed journals and so be included in further

reviews. (Gopalakrishnan and Ganeshkumar, 2013) The value of conclusions drawn in a systematic review is also dependent on the quality of the available evidence. (Higgins and Green, 2011)

Multiple guidelines for performing a systematic review are available, including those produced by the Cochrane Collaboration and the Centre for Reviews and Dissemination. (Higgins and Green, 2011, Centre for Reviews and Dissemination, 2009) There are also guidelines for evaluating systematic reviews such as the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement and those produced by the Critical Appraisal Skills Programme. (Liberati et al., 2009, Critical Appraisal Skills Programme (CASP), 2011) However, these resources do not fully address the methodology for a systematic review of measurement instruments, as they are primarily concerned with studies of interventions or diagnostic tests.

1.5.2. Additional considerations for a systematic review of measures.

Systematic reviews can be used to identify existing instruments and evaluate their quality, both in terms of how they were developed and their measurement properties. As such, the systematic review must appraise the design, analysis and interpretations of the development studies and consider any bias influencing the quality of the instrument produced. In addition, the measurement properties of the scale should be evaluated. (Mokkink et al., 2010b, Terwee et al., 2012)

Terwee and colleagues gathered all existing systematic reviews of patient-reported outcome measures to evaluate the processes used. (Mokkink et al., 2009) They found considerable variation in quality of review design and highlighted limited use of standards for measure development and criteria of

adequacy for the instrument measurement properties. (Mokkink et al., 2009) The following recommendations were made for conducting a systematic review of patient-reported outcome measurement instruments:

- The literature search should incorporate terms describing the construct measured by the instrument, the population of interest and any other key characteristics needed.
- Broad methodological search terms should be used due to the inconsistencies in database indexing and the wide terminology in use.
- An additional search including the name of identified instruments is recommended in order that all development and evaluation studies are found.
- The methodological quality of included studies should be appraised using reproducible standards.
- The measurement properties of instruments found should be assessed using reproducible criteria of adequacy.

1.6. Summary with thesis overview

Shared decision-making is an increasingly important facet of modern health-care practice with a growing field of research exploring the methods used and decision support interventions. Evaluating the effectiveness of any changes in clinical practice requires a clear definition of what constitutes a good decision. Elwyn and Miron-Shatz argue that it is the process leading to the decision that should be considered, rather than the decision itself, proposing a decision-process model incorporating deliberation that focuses on the patient's experience. (Elwyn and Miron-Shatz, 2010)

However, it is unclear the extent deliberation is considered in existing measures of shared decision-making. In addition, an instrument's development

process and measurement properties also influence its suitability for use in appraising interventions for use in clinical practice, and should therefore also be considered. As systematic reviews represent the most robust method of appraising the existing evidence field, this approach will be used to address research question: to what extent do existing measures of shared decision-making incorporate assessment of deliberation by the patient?

This review aims to analyse existing shared decision-making measurement scales, determining the extent their items map onto a decision process model incorporating patient deliberation. The research objectives are to perform a systematic search to identify studies concerning the development or evaluation of an instrument measuring decision-making, where the patient's perspective is sought. Once these are identified, the descriptive and measurement properties of the instruments will be assessed along with the extent their items map onto stages in deliberation and determination in a process map of decision making. Based on the review findings, the need for a new instrument will be considered and specifications outlined, if indicated.

The research aims and objectives are further summarised in the next sections. The full methods used in the study, for both the systematic review and instrument content mapping, are reported in Chapter 4. The results for the systematic review are detailed in Chapter 5, and the findings of the instrument content mapping reported in Chapter 6. The results are discussed and placed in the context of the knowledge to date and study method in Chapter 7, with the implications for developing a new measure of the decision-making process considered. Chapter 8 concludes the thesis and summarises the key learning points from the study.

2. Aims

To analyse existing shared decision-making measurement scales where the patient's perspective is sought, determining the extent their items map onto a decision process model incorporating patient deliberation.

3. Objectives

1. Perform a systematic search to identify studies concerning the development or evaluation of an instrument measuring decision-making, where the patient's perspective is sought.
2. To assess the descriptive and measurement properties of the instruments.
3. To determine the extent the existing instruments' items map onto stages in deliberation and determination in a process map of decision making.
4. Outline new instrument specification, if indicated.

4. Method

This chapter begins by detailing the guidelines used to construct a robust systematic review and the additional resources consulted for a review of measurement instruments. Search strategy development is described, along with the eligibility criteria and study selection, data extraction and synthesis processes. The method of mapping instrument items against Elwyn and Miron-Shatz's decision process model is also reported. Finally, protocol amendments are detailed and discussed, along with the steps taken to keep the review findings timely.

4.1. Systematic review

4.1.1. Developing a systematic review

4.1.1.1. Key steps of a systematic review

A systematic review begins with the identification of a research question and eligibility criteria for the evidence drawn upon. A clear and reproducible method is developed to systematically search resources for records matching the criteria. These are then appraised to determine the validity of the evidence presented and their results noted, with the data gathered synthesised to inform a conclusion. (Shea et al., 2007, Khan, 2005, Higgins and Green, 2011)

4.1.1.2. Resources for method development

The method for this systematic review was developed using guidance from the Cochrane Collaboration and the Centre for Reviews and Dissemination. (Lefebvre et al., 2009, Centre for Reviews and Dissemination, 2009) It was then compared against guidelines for evaluating systematic reviews from the

Critical Appraisal Skills Programme and the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement. (Critical Appraisal Skills Programme (CASP), 2011, Moher et al., 2009)

Additional guidance was sought from a healthcare librarian and information specialist, supplemented with resources from the Cardiff University Systematic Review Network (SysNet). (Mann, 2011) Similar systematic reviews of measurement instruments, both in shared decision-making and related fields, were consulted. (Pink et al., 2009, Joseph-Williams et al., 2011, Elwyn et al., 2001b, Dorman, 2005, Scholl et al., 2011, Sepucha and Ozanne, 2010, Simon et al., 2007, Schellingerhout et al., 2012) The protocol was presented to colleagues in the Decision Laboratory at the Institute of Primary Care and Public Health, Cardiff University.

4.1.1.3. Resources specific for a systematic review of measurement instruments

This preliminary exploration and consultation indicated that the guidelines above did not fully address the process of conducting a systematic review of measurement instruments, as they were primarily concerned with studies of interventions or diagnostic tests. To address this, a Medline search was conducted using search terms for systematic reviews, guidelines and measurement instruments. The recommendations made in the identified resources were then incorporated into the systematic review design. (Mokkink et al., 2009) These included use of broad methodological search terms due to indexing variation and conducting an additional search with the names of included scales in order to identify all developmental studies. (Mokkink et al., 2009) Further recommendations included the use of reproducible standards to appraise the methodological quality of included studies and reproducible

criteria of adequacy to evaluate the measurement properties of identified instruments. (Mokkink et al., 2009)

4.1.2. Search strategy development

4.1.2.1 Strategy foundations

Search strategies classically include three sets or strings of search terms. (de Vet et al., 2008) In a systematic review of measurement scales, two search strings concern the construct to be measured and the population of interest, in this case shared decision-making and patients, in particular their perspective of the decision-making process. The final string relates to instrument evaluation and development. (Mokkink et al., 2009) Terms related to these fields were identified and combined, initially with the Boolean operator "OR" for terms within each search string. The three sets were then combined with the Boolean operator "AND" operator, producing a search strategy that identifies records incorporating all three concepts, as shown in Figure 2 below. (de Vet et al., 2008)

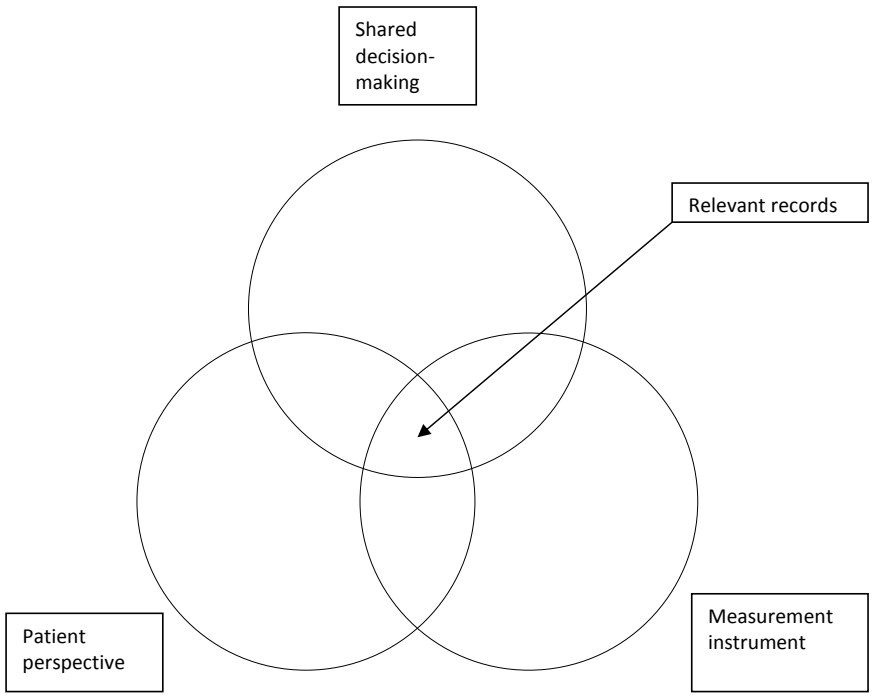
Figure 2: Basic search strategy

#1: construct search (shared decision-making)

#2: population search (patient's perspective on the decision process)

#3: methodology search

#4: #1 AND #2 AND #3



Adapted from: (de Vet et al., 2008)

4.1.2.2 Identifying search terms

As a first step in identifying search terms, key words for the foundation conceptual studies were gathered, such as those for the work of Elwyn and Miron-Shatz. (Elwyn and Miron-Shatz, 2010) These were supplemented with further terms for essential studies that the search strategy should detect, which were chosen following discussion with the review supervisors. Search strategies used in previous reviews of shared decision-making were also

consulted. (Joseph-Williams et al., 2011, Pink et al., 2009, Scholl et al., 2011, Simon et al., 2007) Additional decision-making and participation-related terms were identified from the Cochrane Consumers and Communication Review Group search strategy for the specialised register. (Cochrane Consumer and Communication Review Group)

Search filters have been developed to identify studies reporting instrument development and evaluation. (Terwee et al., 2009) The terms included in the filters were reviewed but the filters themselves were not used due to the risk of inaccuracy associated with variable indexing and other changes to databases over time. (de Vet et al., 2008, Mokkink et al., 2009)

The terms identified were mapped against the standardised subject indexing used in databases. (Lefebvre et al., 2009) In Medline, these are called MeSH terms and found using the "Map to Subject Heading" function. (Cardiff University IT and Library Services, 2012) The scope and tree, or definition and filing location, of each identified MeSH term was reviewed to ensure relevance and to identify other related records that could be incorporated into the search through the "explode" function of the database. (Lefebvre et al., 2009) The use of index terms alone is discouraged due to dependence on the indexers' interpretation of field, the description of the study given by authors and the format of information collated by databases. However, they are an essential tool for identifying records when used in conjunction with free text, as index terms are tailored to a database and have been developed to extract information efficiently from that resource. (Lefebvre et al., 2009)

4.1.2.3 Improving comprehensiveness or sensitivity

A sensitive search strategy detects a greater number of records, with a comprehensive and broad approach missing fewer resources. However, this also means that a greater number of ineligible studies are detected. A sensitive search strategy is recommended for systematic reviews of measurement instrument studies due to varying terminology use and concerns of poor indexing. (de Vet et al., 2008, Mokkink et al., 2009)

To improve the thoroughness of the search, free-text terms along with synonyms and variant spellings were incorporated into the strategy. Truncations, such as "scor*" for scoring, scored and score, were also used to broaden the search. In addition, the "explode" function was used for MeSH terms where the scope and tree of the term suggested this would yield further relevant records. (Lefebvre et al., 2009)

4.1.2.4 Improving precision or specificity

A more specific or precise search strategy yields fewer ineligible studies. (Lefebvre et al., 2009) A balance is needed between the sensitivity and specificity of a strategy to ensure that all relevant records are detected while also producing a manageable volume of results in view of time and resource restrictions. In this review, the proximity operator ADJn (where n is the maximum number of words allowed between the specified search terms) was used to improve the focus of search outputs for otherwise very broad free-text terms. (Cardiff University IT and Library Services, 2012) For example, records containing the free-text term "patient" in their title or abstract were only detected if "decision-making" was also identified at a maximum of three words away. The Boolean operator "NOT" was not used, as it can lead to inadvertent exclusion of results. (Lefebvre et al., 2009)

4.1.2.5 Pilot searches

The search strategy was refined through multiple trial searches. An iterative approach was taken, with the impact of each modification to the strategy evaluated by the relevance of the records identified and the detection of key articles, which were selected during background reading and discussion with supervisors. Examples of the terms excluded are listed in Appendix 4, along with the impact they had on the pilot search outputs. Additional guidance from an information specialist was sought to optimise the strategy. The final search strategy for Medline is shown in Table 2.

The pilot searches also highlighted less of an overlap between two databases than had initially been anticipated. (Lefebvre et al., 2009) The provisional search strategies for Medline and EMBASE were kept sensitive with a target of 5,000 results each. However, few duplicates were detected and the precision of the searches was further developed to reduce the combined search results from over 12,000 to around 3,000. This improved specificity while maintaining the relevance of the search output and its ability to detected key articles.

4.1.2.6 Reducing error

All versions of the search strategies were saved within database accounts in addition to being cut and pasted into Word to ensure that the searches were reproducible and the development process transparent, with all decisions documented in a research diary. (Lefebvre et al., 2009) To reduce the introduction of error, the complete search strategy outputs with full references for each record were downloaded into a reference manager (EndNote versions XV-X7) and notes made of insufficient details in any references and articles. (Lefebvre et al., 2009)

Table 2: Medline via Ovid search strategy

1. ((patient* or client* or consumer*) adj3 decision making).mp.

2. exp patients/
3. exp Patient Participation/
4. exp evaluation studies/ or exp validation studies/
5. psychometrics/
6. reproducibility of results/
7. (valid* or reliab*).mp.
8. ((measur* or scal* or scor* or instrument* or survey* or tool* or question*) adj6 decision*).mp.
9. exp Decision Making/
10. shared decision-making.mp.
11. ((consider* or reflec* or deliberat*) adj3 decision*).mp.
12. (decision* or choice* or prefer* or judg*).tw.
13. or/1-3
14. or/4-8
15. or/9-12
16. and/13-15

/	after an index term (MeSH heading) indicates that all subheadings were selected.
*	before an index term indicates that that term was focused - i.e. limited to records where the term was a major MeSH/Emtree term.
"exp"	before an index term indicates that the term was exploded.
.tw.	indicates a search for a term in title/abstract
.mp.	indicates a free text search for a term
#	retrieves records that contain the search term with substituted character(s) in the specified location.
*	at the end of a term indicates that this term has been truncated.
*n	The limited truncation symbol, $n, Retrieves records that contain the search term and all possible suffix variations of a root word with the maximum number of characters that may follow the root word or phrase, specified by n.
?	in the middle of a term indicates the use of a wildcard.
adj	indicates a search for two terms where they appear adjacent to one another

(Source: Mann, 2011)

4.1.3 Databases and information sources

4.1.3.1 Electronic databases

Guidelines for systematic reviews highlight the need for both a detailed strategy and the use of multiple, overlapping resources when constructing a thorough search. (Lefebvre et al., 2009) Shared decision-making and measurement instrument development are growing fields, with a broad and varying terminology used in both. They are also areas that concern a range of disciplines, such as health, education and psychology. As such, multiple electronic databases were used. (Joseph-Williams et al., 2011)

Medline (including Medline in Process), EMBASE and the Cochrane Library were recommended as key resources and were supplemented with subject specific databases, including PsycINFO, CINAHL and ASSIA. (Lefebvre et al., 2009) The citation index Web of Science was used for citation searching, which also allowed identification of corrections or errata. (Lefebvre et al., 2009) For each database, the service provider, date of search and time period searched were recorded. (Lefebvre et al., 2009)

4.1.3.2 Adapting Medline search strategy to other databases

Each database has differing subject coverage, search processes and standardised search terms. The Cardiff University Information Services guides to each database were consulted in order to adapt the core Medline search strategy to a resource, and the database thesauruses used to identify standardised search term variations for the three main search threads. (Cardiff University IT and Library Services, 2012) For example, PsycINFO database searches required the substitution of “client” for “patient”. The same iterative approach was repeated, with pilot searches run and search terms adapted

according to relevance of results and identification of key articles. The search strategies are included in Appendix 5.

4.1.3.3 Additional resources

Electronic database searches were supplemented with hand searches of the most frequently cited journals, with the reference lists of included reports also reviewed for further relevant studies. Specific follow-up searches using the name of the identified measurement instruments were performed to ensure that all development and evaluation studies for the scales were identified. (Mokkink et al., 2009) The resources used are summarised in Table 3.

4.1.4. Search limits

Systematic reviews should include relevant unpublished studies in order to limit publication bias. However, the reporting and indexing of published instrument development and evaluation reports are acknowledged as being highly variable. Grey literature and unpublished sources introduce further the possibility of incomplete reports. Due to the practical difficulties this would present along with time restrictions, resources for grey literature and unpublished studies were not searched. (Centre for Reviews and Dissemination, 2009)

No restrictions were placed on language or date for the database searches. However, in the event of a translation being unfeasible, this was to be recorded as the cause for exclusion. (Centre for Reviews and Dissemination, 2009)

Table 3: Summary of resources

Bibliographic databases:
Medline (Biomedical)
Medline in Process (as above, with very recent citations and abstracts, and non-indexed citations)
Cochrane Library
EMBASE (Biomedical and pharmaceutical database)
PsycInfo (Psychology and psychological aspects of related disciplines)
ASSIA (Applied social sciences, education and health)
CINAHL (Nursing and allied health)
Web of Science (Social Science Citation Index and Science Citation Index)
Journals: manual searches of those most frequently identified by database searches (for preceding five years)
Health Expectations
Patient Education and Counseling
References:
Examination of the reference lists of included papers

(Adapted from: Mann, 2011)

4.1.5. Eligibility criteria

Pre-specification of eligibility criteria is a key requirement of systematic reviews, in order to maintain transparency and limit the introduction of bias. The inclusion and exclusion criteria for this review are detailed in Table 4 and Table 5. They were developed using the review question and objectives, and kept broad where possible while also giving sufficient detail to appraise eligibility consistently. (Bossuyt and Leeflang, 2008)

To be included, a study had to address the development or further evaluation of a measurement instrument for an aspect of shared decision-making, specifically considering the patient's perspective of the decision-making process. Including a minimum number of the stages of deliberation described in Elwyn and Miron-Shatz's decision process model as an inclusion criterion was considered unfeasible at the title and abstract screening phase, as it was

felt that this level of detail was unlikely to be present in references. (Patrick et al., 2009)

Due to the different disciplines with an interest in shared decision-making, two main limitations were placed on study design. (Joseph-Williams et al., 2011) First, that the decision context related to healthcare and second, that the study must involve the psychometric evaluation of specific measurement. The latter meant that the study reports would include details of the design and measurement properties of the instrument, allowing appraisal of the quality of these aspects. While this meant excluding intervention studies and reviews, effort was made to search for original instrument development and evaluation reports. Broad eligibility criteria were used for the description of the psychometric evaluation of instruments due to the variation in reporting and terminology described by Mokkink and colleagues. (Joseph-Williams et al., 2011, Mokkink et al., 2009)

Table 4: Inclusion criteria

	Criterion	Clarification	Justification
Concept to be measured	Shared decision-making in healthcare setting	Construct measured by instrument is an aspect of shared decision-making Can be for research or clinical purpose	Main construct must be decision-making if deliberation in decision-making to be assessed
Population of interest	Patients involved in healthcare decision-making, for decision regarding their own health or treatment	Must seek to measure patient's perspective of decision-making process i.e. patient-reported, though can be self-completed or via interviewer	Patient deliberation can only be determined by seeking the patient's perspective on their experience
Instrument of interest	Reports on the development and testing of instrument designed to measure shared decision-making	Psychometric evaluation: at least two measurement properties noted If modified version of previously tested scale: further psychometric testing reported. If multidimensional instrument: separate psychometric data for SDM relevant subscales reported.	Information needed to appraise quality of instrument

Table 5: Exclusion criteria

	Criterion	Clarification	Justification
Concept to be measured	Measured construct is not an aspect of shared decision-making in the health-care setting	e.g. decision-making in business or management environment	Not relevant to patient deliberation in shared decision-making
Population of interest	Proxy-reported or physician/ researcher rated Not specific to patient involvement in decision-making regarding their own health or treatment	Completed on behalf of patient by health professional, guardian or carer. E.g. Patient rating of service quality	Not direct recording of patients perspective or recording of their experience of decision-making process Not relevant to decision-making from patients' perspective
Instrument of interest	Paper does not focus on development and evaluation of instrument Instrument or evaluation data unavailable for scrutiny	e.g. intervention study or review. Unlikely to have evaluation information for instrument. Limits mapping of instrument items against decision process map and evaluation of instrument quality.	Information needed to appraise quality of instrument

4.1.6. Study selection

Full references were downloaded into reference management software (Endnote versions XV-X7), with the search outputs of different databases merged and duplicates removed. The titles and abstracts were screened using the eligibility criteria and graded as include, exclude or uncertain. A full text report was then obtained for those graded as include or uncertain. (Higgins and Deeks, 2011, Centre for Reviews and Dissemination, 2009)

One reviewer (SS) screened all titles and abstracts with a second reviewer (SM) independently examining 10% of the database search results. Guidelines recommend that more than one reviewer independently review all results to reduce the possibility of error and increase the reproducibility of results. However, in view of the resources available and high number of records produced by a sensitive search strategy, a second reviewer for a proportion of the results is considered an accepted compromise (Higgins and Deeks, 2011, Centre for Reviews and Dissemination, 2009)

Study selection pilot was performed with 10 references to assess the clarity of the eligibility criteria and improve the consistency of its application between reviewers. A kappa statistic was calculated to assess inter-assessor reliability. Though not recommended as standard by Cochrane Collaboration, it is used to highlight problems at the piloting stage. Values of 0.6 - 0.74 are considered a good level of agreement between reviewers, with equal to or greater than 0.75 defined as excellent. For this pilot, the kappa statistic obtained was 0.78. (Higgins and Deeks, 2011)

There was a low threshold for suggesting a full text review due to concerns over the level of detail available in the titles and abstract of measurement

instrument studies. (Reeves et al., 2009) Following full text retrieval of potentially relevant reports, one reviewer (SS) determined study inclusion based on compliance with the eligibility criteria. At this stage, the individual items of the instruments were evaluated to assess their relevance to the measurement of shared decision-making from the patients' perspective. (Patrick et al., 2009)

In the event of uncertainty over study eligibility not being resolved by review of the full text or discussion with the second reviewer, the study would be referred for arbitration by an agreed third party (GE) with a record made of decision reached. (Higgins and Deeks, 2011) As suggested by the Cochrane Collaboration guidelines, the reviewers had different backgrounds, reducing the potential bias introduced by previous experience in the field under review. (Higgins and Deeks, 2011)

4.1.7. Data extraction

4.1.7.1. Selection of data extraction fields

A key consideration in selecting data extraction fields was relevance to the research question and the practical application of the review findings. In addition, as recommended by the Cochrane Collaboration, systematic reviews of similar constructs were consulted and appraised. (Higgins and Deeks, 2011) Reproducible, pre-determined standards and criteria were applied to the study design and the measurement properties of the instrument developed, to increase transparency and minimise variation and bias. (Mokkink et al., 2009)

4.1.7.1.1 Descriptive features

Study characteristics such as the setting, population and decision context were noted, along with instrument characteristics including language and readability. These details determine how practical the scales are to use and the generalisability of study findings to other populations. (Critical Appraisal Skills Programme (CASP), 2011)

4.1.7.1.2 Methodological quality

The standards applied to the study design were the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist for the methodological quality of studies of measurement instruments. The COSMIN checklist includes standards for validity, reliability, responsiveness, interpretability, IRT statistical methods and generalisability. These standards were used due to the availability of published reports of checklist development and application, along with recommendations made in previous reviews of shared decision-making measurement instruments. (Mokkink et al., 2010a, Scholl et al., 2011) Both the checklist and the 4-point rating scale versions of the standards were used, as the checklist gathered greater detail and was therefore used for data extraction, while the rating scale produced concise, easily comparable summaries more suited for synthesising the review findings. (Terwee et al., 2012) However, caution is needed with the use of rating scales as they can lead to a loss of detail and oversimplification of results. (Moher et al., 2009) A comparison of the two methods is detailed in Appendix 6.

4.1.7.1.3 Instrument quality

Reproducible criteria for instrument measurement properties developed by Terwee and colleagues were used to evaluate the quality of the scales identified. This guidance was selected as it was developed by the same research team as the COSMIN checklist and has also been used in the Institute of Primary Care and Public Health at Cardiff University. (Terwee et al., 2007, Pink et al., 2009)

4.1.7.1.4 Further evaluation studies

Eligible further evaluation studies for included instruments were reviewed and information gathered to supplement data from original development studies. The extraction fields included the study context and design, and measure validity and reliability.

4.1.7.2. Data extraction form

A standardised form was created using Microsoft Excel and the fields used are summarised in Appendix 7. An electronic format was developed instead of paper forms due to the ease of gathering, recording, transferring and combining data. (Higgins and Deeks, 2011)

4.1.7.3. Data extraction process

Due to time and resource limitations, one reviewer (SS) extracted the data with a second (SM) reviewing the data extraction form and any queries. The form was piloted to ensure clarity and consistency of interpretation. The third reviewer (GE) was to be consulted if there was a disagreement. (Higgins and Deeks, 2011)

4.1.8. Narrative synthesis

The final step in a systematic review is to summarise, or synthesise, the results, such that patterns can be identified and conclusions drawn in relation to the research question. (Ryan, 2013) There are different methods for this process, with steps used dependent on the variation, or heterogeneity, of the data gathered and statistical processes applied in the included studies. (Ryan, 2013) As highlighted by Mokkink and colleagues, there is a wide variation both in the methodological approaches used for instrument development and the measurement properties evaluated. (Mokkink et al., 2009) It is therefore unlikely that a statistical meta-analysis will be possible and a narrative synthesis will be performed. As recommended by the Cochrane Collaboration, the findings will be tabulated to provide an overview of the results and comparison of the included studies. (Ryan, 2013)

4.2. Instrument content mapping

In addition to a systematic review of literature to identify and appraise the quality of existing measures, each identified scale was evaluated to determine the extent they address deliberation. This was done by appraising the individual items forming the measures and mapping these against the stages of the decision process model described by Elwyn and Miron-Shatz. (Elwyn and Miron-Shatz, 2010) The mapping results for each scale were then gathered in a summary table to compare their content and performance, and also evaluate whether all stages of the decision process model were met.

4.3. Protocol amendments.

While the electronic database Scopus was originally included as a search resource, it was excluded at the protocol stage due to potential overlap with other databases and resource limitations. Though not a deviation, a key aspect

of the eligibility criteria wording was demonstrated during the initial study selection pilot. The second criterion was that the population of interest in the study was patients involved in healthcare decision-making for a choice regarding their own health or treatment. This was clarified to include that the scale developed must seek to measure the patient's perspective of the decision-making process. This emphasis excludes scales assessing role preference or consultation scoring, which are not concerned with the decision-making process itself. It also excludes scales that focus on testing patient knowledge of a specific condition or treatment. While knowledge appraisal is an aspect of shared decision-making, the ability to answer specific questions regarding screening accuracy, for example, does not address the patient's perspective of their decision-making process.

It was originally proposed that report authors would be contacted for further details or clarification to aid evaluation by the standards and criteria used in the review. However, the varying terminology and detail of the reports made this an unfeasibly lengthy and time intensive process that would be needed for all of the included studies. As such the data extraction was based on the published reports of instrument development and evaluation.

4.3. Updating the review.

The original database searches for the review were conducted between April 13th and April 17th, 2012. These were supplemented with manual searches of the most frequently cited journals, Health Expectations (May 2007 to May 2012) and Patient Education and Counseling (May 2007 to May 2012), and the references of included studies were reviewed. In order that the review findings were kept timely and relevant, the database searches were updated in July 2014. The supplemental search strategies were not repeated in view of the low

yield and duplication of database results noted originally. A more in-depth report of each search output is detailed in Chapter 5, the results of the systematic review.

This section described the systematic review process undertaken, which was in keeping with Cochrane Collaboration and the Centre for Reviews and Dissemination guidelines. Further guidance for systematic reviews of measurement instruments was utilised and the subsequent approach included a broad search strategy, with supplemental database searches performed in order to identify all development or evaluation studies for each instrument. Due to the known methodological heterogeneity in instrument development and evaluation, a narrative synthesis was used to summarise the review findings. In addition to a systematic review to identify and appraise the quality of existing measures, content mapping was performed to determine the extent scale items address the stages of a decision process model incorporating deliberation proposed by Elwyn and Miron-Shatz. (Elwyn and Miron-Shatz, 2010)

5. Results of the systematic review

This chapter describes the output of the search strategies undertaken, details progress through study selection and provides a summary of the included scales. For each of these scales, the descriptive features are outlined before analysis of the methodological quality and measurement properties is undertaken. These findings are then summarised in a narrative synthesis of the review findings.

5.1. Summary of results

5.1.1. Output of search strategies

The final databases searches were performed between April 13th and April 17th, 2012. From the seven databases, 5950 references were retrieved. Table 6 details the references found from each database.

Table 6: Individual database search outputs

Database	Service Provider	Dates covered	Number of references
Medline	Ovid	1945 to April 2012	1878
EMBASE	Ovid	1947 to April 2012	1550
Medline in process	Ovid	As listed on April 13th, 2012	90
Cochrane Library	Wiley	1898 to April 2012	357
PsycINFO	Ovid	1806 to April 2012	776
CINAHL	EBSCO	1937 to April 2012	1208
ASSIA	Proquest	1987 to April 2012	91

The references were exported into EndNote XV and duplicates identified using the reference manager search option. On title and abstract review, further duplicates were identified and removed. The total number of duplicate references was 1034, leaving a final total of 4916. Following title and abstract

review, 159 references progressed to full text review. From these, ten eligible studies concerned the initial development of an instrument and another eight considered further evaluation.

Database searches were supplemented with manual searches of the most frequently cited journals, Health Expectations (May 2012 to May 2007) and Patient Education and Counseling (May 2012 to May 2007), and the references of included studies were reviewed. In addition, citations of the included studies were identified using Web of Science, and additional database searches performed using the names of the identified instruments. The results of these searches are detailed in Table 7. However, the supplemental searches did not add to the final yield of eligible studies.

Table 7: Supplemental search outputs

Source	Detail	Number of potentially relevant studies identified	Number eligible on further review	Number once database duplicates removed
Manual journal searches	Health Expectations (May 2007 to May 2012) and Patient Education and Counseling (May 2007 to 2012)	14	0	0
References of included studies	-	39	0	0
Citations of included studies	Identified via Web of Science (Thomas Reuters), from 1900 to May 2012	827	8	0
Instrument name database searches	Using Medline via Ovid (1945 to May 2012)	136	1	0

The database searches were updated in July 2014. In total, a further 1977 references were identified. The individual database outputs are detailed in Table 8. These were exported into the reference manager EndNote X7 and 346 duplicates removed. Following title and abstract review, 9 articles progressed

to full text review. No new instruments were identified, and one eligible study concerning the further evaluation of a measure was included. The supplemental search strategies were not repeated in view of the low yield and duplication of database results noted originally.

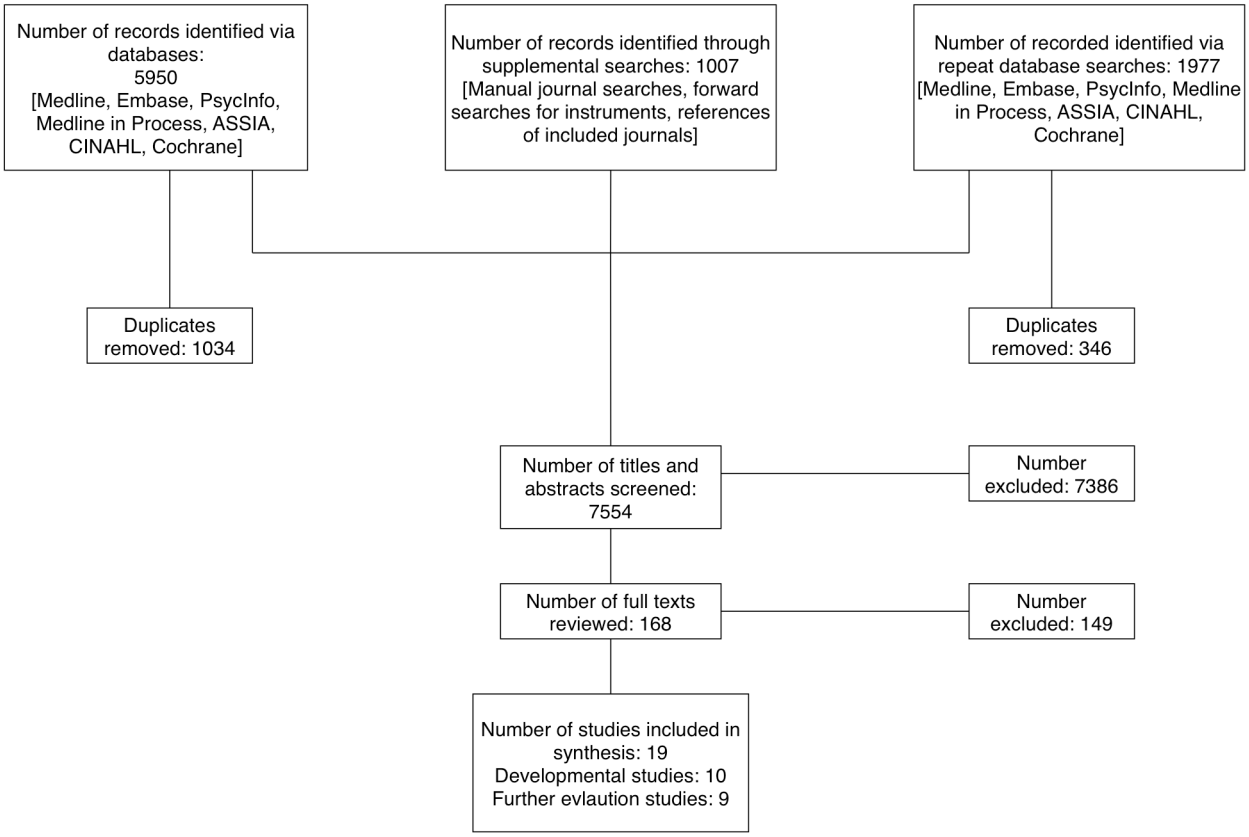
Table 8: Updated individual database search outputs

Database	Service Provider	Dates covered	Number of references
Medline	Ovid	April 2012 to July 2014	490
EMBASE	Ovid	April 2012 to July 2014	684
Medline in process	Ovid	As listed on July 18th, 2014	176
Cochrane Library	Wiley	April 2012 to July 2014	114
PsycINFO	Ovid	April 2012 to July 2014	245
CINAHL	EBSCO	April 2012 to July 2014	260
ASSIA	Proquest	April 2012 to July 2014	8

5.1.2. Progress through systematic review

Figure 3 details the progress through the systematic review. In total, ten studies were identified addressing the development of an instrument and nine concerned the further evaluation of an instrument.

Figure 3: A PRISMA flowchart of progress through the systematic review



5.1.3. Agreement between reviewers

One reviewer (SS) screened all titles and abstracts. A second reviewer (SM) reviewed 10% of the titles and abstract. From the 500 titles and abstracts, 24 were selected for full text review, with no disagreement between reviewers. Of the full texts considered for eligibility, the third reviewer contributed to the consideration of three studies. (Barry et al., 1997, Bunn and O'Connor, 1996, Hollen, 1994)

5.1.4. Excluded scales

A balance was made during the search strategy development to ensure strong sensitivity for detecting eligible studies. The resulting low specificity identified

a high number of references. At the title and abstract study selection stage, the main reasons for exclusion were not addressing shared decision-making or the patient's perspective on their involvement in the decision process. These studies were often concerned with decisions from the perspective of clinicians or patient-reported outcomes for different treatment options. Other themes encountered were evaluations of decision capacity and competency.

At full text review, ineligible studies often focused on patient or clinician preferences for involvement in consultations, and grading of subsequent encounters. Studies were also excluded where decision evaluation focused on testing patient knowledge of specific conditions or tests, as this was felt to not represent evaluation of the decisional process itself. Examples of key studies excluded at the full text stage are included in Table 9. (Barry et al., 1997, Elwyn et al., 2013, Finnell and Lee, 2011, Geiger et al., 2011, Kaltoft et al., 2014, Kremer and Ironson, 2008, Kriston et al., 2010, Melbourne et al., 2010, Miller et al., 2009, Ogden et al., 2008, Sepucha et al., 2011, Sepucha et al., 2012, Sung et al., 2010)

Table 9: Key excluded scales

First Author, Year	Instrument name	Reason for exclusion
Kaltoft, 2014	MyDecisionQuality	Theoretical with no psychometric evaluation
Elwyn, 2013	CollaboRATE	Rating shared decision-making in consultation rather than a measure of the process of decision making from the patient perspective.
Geiger, 2011	SDM-MASS	A compound measure considering three perspectives in one index (patient, clinician, observer) and rating shared decision-making within a consultation.
Sepucha, 2012	Herniated disc-decision quality instrument (HD-DQI)	Condition specific – concerned with knowledge and decisions for herniated discs.
Finnell, 2011	Decisional Balance for Patient Choice in Substance Abuse Treatment	Hypothetical preference for involvement in future decision making for a specific health context.
Sepucha, 2011	HK-DQI	Condition and treatment specific, with focus on testing knowledge rather than decisional process.
Kriston, 2010	SDM-Q-9	A measure of shared decision-making in consultation than evaluation of patient decision process
Melbourne, 2010	Dyadic OPTION	A measure of shared decision-making in the consultation rather than consideration of decision-making process from the patient perspective.
Miller, 2009	Decision Making Control Instrument	Not decisional process – measure of perceived freedom to make own decision
Ogden, 2008	Choice questionnaire	Preference for involvement in shared decision-making than patient evaluation of decisional process
Kremer, 2008	Control Preferences Scale	Preference for involvement in shared decision-making than patient evaluation of decisional process
Sung, 2008	Autonomy Preference Index (API)	Preference for decision involvement and disease specific.
Barry, 1997	BPH Knowledge and Satisfaction Questionnaires	Nested in randomised controlled trial and involving three different scales. Addresses determination, understanding and involvement in decision for specific condition rather than patient perception of decisional process.

5.1.5. Summary of included scales

Ten scales were included, and Table 10 outlines the key details for these. A further nine studies concerned additional evaluation of the Decisional Conflict Scale, the Satisfaction with Decision Scale, SURE, COMRADE and the Decision Evaluation Scale. The full scales are included in Appendix 8.

Table 10: Summary of included scales

Name of scale	First author, year	Country	Language
Decisional Conflict Scale	O'Connor, 1995	Canada	English
Satisfaction with Decision	Holmes-Rovner, 1996	USA	English
Decision Attitude Scale	Sainfort, 2000	USA	English
Prep-DM	Bennett, 2010	Canada	English
SDM-Q	Simon, 2006	Germany	German
SURE	Légaré, 2010	Canada & USA	French & English
COMRADE	Edwards, 2003	UK	English
Decision Evaluation Scales	Stalmeier, 2005	Netherlands	Dutch
Decision-making quality scale	Hollen, 1994	USA	English
Decision Self-efficacy scale	Bunn, 1996	Canada	English

5.2. Evaluation of scales

5.2.1. The Decisional Conflict Scale (DCS)

5.2.1.1 Descriptive features

Developed by O'Connor in 1995, the scale focuses on decisional conflict and perceived contributory factors. (O'Connor, 1995) It contains three subscales: three items for uncertainty, four for effective decision making and nine items assessing factors contributing to uncertainty. The uncertainty subscales can be used during the decision making process, while that for effective decision making is intended for use once a decision is reached. Designed for self-administration or use over the telephone, it is structured in a five point Likert scale, takes five to ten minutes to complete and requires a Grade 8 reading level. A higher score is considered to represent higher decisional conflict.

The scale was intended for application both in clinical and research uses. It was originally evaluated in two healthcare decision scenarios, with the DCS

administered immediately after the decision in all groups. A sample of 45 students was also retested two weeks later.

The first of these scenarios concerned a decision to receive an influenza vaccine. Nursing agency and teaching hospital staff (n=115), cardiorespiratory patients (n=283) and two groups of health science students (n= 45 and 106) were given information about the vaccine and asked regarding their decision or intention to be immunised. For the student groups, this represented a hypothetical choice as they were not eligible for the vaccine.

The second concerned a hypothetical decision to undergo breast cancer screening. A random sample of 360 women aged 50 to 69 was recruited from the community via a telephone survey. As well as determining background knowledge, attitudes and approaches to health screening; the participants were asked what they would do if invited for screening.

5.2.1.2 Methodological quality

Table 11 outlines the measurement properties evaluated in the DCS development study, with the COSMIN checklist for methodological quality then applied to each. (Mokkink et al., 2012)

Table 11: DCS development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in DCS development study
Reliability	
Internal consistency	Yes
Reliability	Yes
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	No
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

Internal consistency, that is the interrelatedness of items such that they measure the same construct without creating redundancy, was evaluated both for the global scale and each subscale. (Mokkink et al., 2012) A Cronbach's alpha was calculated for each. However, there is no documented factor analysis for the scale or subscales. This is required to determine the presence of unidimensionality, an assumption to be met for internal consistency statistics. Its absence raises the query of whether the scale or subscales can truly be said to be measuring only one attribute or factor, and whether other elements could be influencing item endorsement and scoring. (Mokkink et al., 2012) Correlations between the subscales are considered, and noted to be consistent with the predicted relationship between uncertainty and decisional conflict. These results are included under reliability in the study report, although they are relevant for the validity of the scale.

Reliability is defined by COSMIN as "the proportion of total variance [….] due to true differences between patients". (Mokkink et al., 2012) Two sets of measurements are available, collected at specified time intervals, though from a smaller subgroup of participants (n = 45). Limited information is given concerning test conditions and, while the decision is hypothetical, information concerning participant stability between measurements would be beneficial. For example, as students it is possible that they encountered further information about the vaccine between administrations of the scales. In addition, while a Pearson correlation coefficient is calculated, this is considered insufficient by COSMIN due to the potential influence of systematic error. An intraclass correlation coefficient (ICC) is recommended instead. (Mokkink et al., 2012)

Content validity evaluates how thoroughly the included items cover the domain of interest. (Mokkink et al., 2012) Limited information is provided of the early developmental stages of the scale. No details are given of expert opinion being sought, or the involvement of the target population. However, the author considers the theoretical foundation for the scale and whether the included items cover decisional conflict as a construct.

Hypothesis testing, also referred to as construct validity, considers whether scores correlate with other instruments, participant subgroups or other traits in a manner consistent with existing knowledge of the construct of interest. (Mokkink et al., 2012) The proposed relationships can be convergent or divergent, with the latter also referred to as discriminant. (Streiner and Norman, 2008) In this study, the hypotheses are developed a priori, thus limiting post hoc bias. (Mokkink et al., 2012) However, while the expected directions of correlations are described, the magnitude is not. The DCS is not

tested against another instrument, with the hypotheses instead based on whether a decision is made and the level of knowledge demonstrated in a breast cancer knowledge test. Limited information is given concerning the validity and reliability of this test and it is also assumed that increasing knowledge decreases conflict. The results are reported with mean scale scores, standard deviations and p-values for the decision outcome analysis, which limits interpretation of the magnitude of any correlation. A Pearson r correlation coefficient and P value are provided for the breast cancer knowledge test, but the supporting data for these calculations are not provided.

The interpretability of the scores obtained by the scale is limited. Mean scores and their standard deviations are given but no proportions given for those scoring the highest and lowest scores. While scores are given for some subgroups, such as decision outcome, none are given for subgroups such as age or gender. Finally, no minimal important change (MIC) or minimal important difference (MID) is outlined, which makes the clinical relevance of scores hard to determine. For example, results reported as statistically significant correspond to differences in scores of less than one, which may have limited meaning in practice. However, these omissions may also be due to the early stages of scale development.

When considering the generalisability of the study findings, limited demographic details are given for some participant groups, with insufficient disease and treatment characteristics for the cardiorespiratory patient group. The selection criteria are not fully described and, while the proportion of missing responses is reported as 1%, further details, such as which items were incomplete and how such scales were handled in statistical analyses, are not provided.

Further limitations in study design include that different subscales are evaluated in different groups, limiting the interpretability of both the scale scores and measurement properties. The study report also describes varying versions of subscales being used, but provide few details of how and why the changes were made.

5.2.1.3. Instrument quality

Table 12 maps the measurement properties for the DCS against quality criteria developed by Terwee and colleagues. (Terwee et al., 2007)

Table 12: DCS development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	**Evaluated in DCS development study**
Reliability	
Internal consistency	Yes
Reliability	Yes
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Floor and ceiling effects	No
Interpretability	Yes

For internal consistency, the Cronbach's alpha obtained from the different participant groups for the complete scale ranged from 0.70 to 0.95. For individual subscales, the values ranged from 0.58 to 0.70, which fall outside the recommended levels of 0.70 to 0.95. (Terwee et al., 2007) The authors describe reliability in terms of test-retest scores, with a Pearson correlation coefficient of 0.81 and no statistically significant difference observed. However,

the quality criteria developed by Terwee and colleagues recommend the use of ICC or a weighted Kappa statistic. (Terwee et al., 2007)

When evaluating content validity, a clear description is given of the measurement aim but not the target population, item development or of target group or expert involvement in scale development. For construct validity, specific hypothesis were described a priori with at least 75% of the results in accordance. Though mean scores and their standard deviation are reported, interpretation of measure scores beyond the study is limited as no MIC or subgroup score analyses are provided.

5.2.1.4. Further evaluation studies

The Web of Science identified 565 citations of the original development study for the DCS by the end of 2014. Of these, six further evaluation studies matched the eligibility criteria and reported the translation and adaptation of the DCS for use in different health fields. The evaluation studies are summarised in Table 13 below. These studies also highlight variation in measurement property terminology.

The study by Koedoot and colleagues concerns the validation of a Dutch version of the DCS in two samples of oncology patients (N = 29, 141). (Koedoot et al., 2001) Psychometric evaluation included internal consistency, with Cronbach's alpha obtained for the three subscales in both groups (0.61-0.75, 0.77-0.80, 0.83-0.81) and for the overall scale in both groups (0.75-0.82). The authors evaluate construct, criterion and factoral validity. For construct validity, correlations between subscales were examined. These matched predictions in the first group but no relation was found between decision uncertainty and the perceived effectiveness of decision-making in second group

with weak correlation between factors contributing and effective decision-making. Criterion validity was assessed using known-group comparison, with significant support of the hypothesis noted in first group of participants, but the hypothesis was supported for only two of the subscales in the second group. For factoral validity, the confirmatory factor analysis performed did not find a similar structure to that proposed by O'Connor with subsequent exploratory factor analysis finding four rather than three factors. It is possible that one of these factors – uncertainty - reflected response tendencies and was influenced by bias from positive or negative wording of the item rather than its meaning, with the authors also suggesting that the timing of subscale completion influenced the results.

Mancini and colleagues sought to validate a French version of the scale in a sample of patients (N=342) considering taking a genetic test for breast and ovarian cancer. (Mancini et al., 2006) The DCS is described here as having 16 items that form 5 subscales: uncertainty, uninformed, unclear values, unsupported and ineffective choice. Of the respondents, 294 were patients with controls brining the total to 560. For internal consistency, the Cronbach's alpha obtained for the whole scale and subscales ranged 0.67 to 0.916, with the exception of 0.441 to 0.593 for the unsupported subscale. Exploratory and confirmatory factor analyses were performed in two control groups and the intervention group, with results varying between a four and five factor model. The authors conclude that the factorial structure of the DCS is dependent on the level of conflict involved in the decision. Criterion validity was tested by the relationship between the DCS score and whether a decision was made. For the control groups, the DCS score was significantly lower in those who wanted to have the test, but there was no statistically significant difference in the intervention group. However, as only three of the patients in the intervention

group were categorised as uncertain about the decision, the small numbers may have influenced the statistical significance obtained. Three items were identified as having unacceptable levels of item difficulty – "this is an easy decision to make", "I am clear about how important the advantages are to me in this decision" and "I have the right amount of support from others in making this choice".

The study by Song and colleagues evaluates the use of DCS in a trial of a patient-centred advance care planning intervention. (Song and Sereika, 2006) The DCS is described here as having three subscales (uncertainty, factors contributing to uncertainty and decision effectiveness) with 16 items. The study uses combined data from two trials, where 59 participants were recruited from the 95 approached. For internal consistency, a Cronbach's alpha of 0.81 was obtained for the complete scale, and ranged from 0.50-0.73 for the subscales, with item to total correlation weakest for uncertainty items and one "contributing to uncertainty" factor item. Removing these raised the Cronbach's alpha for the complete scale to 0.84. For construct validity, Spearman correlation coefficients were obtained for the DCS and Quality of Communication questionnaire ($r=-0.44$, $p = 0.001$), and the DCS and anxiety rating ($r=0.47$, $p=0.006$), while the correlations varied for the DCS subscales. For discriminant construct validity, a statistically significant difference was found between the control and intervention groups.

Knapp and colleagues sought to validate the properties of both the DCS and COMRADE for the parents of children with life-limiting illness involved in paediatric palliative care programmes. (Knapp et al., 2009) A sample of 936 was contacted and 266 completed surveys returned. For internal consistency, the Cronbach's alpha was over 0.84 and confirmatory factor analysis indicated

that the original structure of the DCS persisted in this context. Floor and ceiling effects were considered. Item-domain convergent/discriminant validity analysis, used to assess congruence between an item and its domain, was supported for two of the three domains and tests for known-group construct validity were statistically significant ($p<0.001$ and $P<0.05$).

Katapodi and colleagues evaluated a modified DCS for decisions concerning genetic testing for hereditary breast and ovarian cancer. (Katapodi et al., 2011) The scale items were reworded to reflect a genetic testing decision and positive phrasing used. From a sample of 372, 354 scales were completed. Factor analysis for inter-item correlation was preformed, producing two new subscales from the items with one remaining subscale unchanged. A Cronbach's alpha of 0.96 was obtained for internal consistency. Convergent and divergent validity were tested with other scales ($r=-1.3$ to $r=-.31$, $p<0.05$-$.001$) and ($r = -0.30$, $p <0.01$) respectively, predictive validity using risk reduction prophylaxis ($r= 0.24$, $p<.001$) and contrast validity with known-group approach ($p<.001$).

A study by Linder and colleagues evaluated a low literacy version of the DCS (DCS-LL). (Linder et al., 2011) The scale was adapted for ordered category responses, with the effective decision subscale omitted and an item removed from each of the remaining subscales. The reliability and validity of the DCS-LL was evaluated before and after the use of a prostate cancer screening decision aid. The study had 149 participants before the intervention and 89 afterwards. For internal consistency, a Cronbach's alpha of 0.80 was obtained and ICCs ranged from 0.246-0.748. An adequate model fit was found with factor analysis at baseline but not at follow up, with subsequent exploratory factor analysis suggesting a three rather than four factor model. Construct validity was evaluated using subscale correlation, with a good correlation was

noted for two of the three “uncertainty” subscales but poor correlation for the “supported” subscale. Discriminatory validity was based on whether a decision was made, and reported as statistically significant. There is uncertainty over which subscales are used in the study and some subscale data are removed from the statistical analyses.

Table 13: Summary of further evaluation studies for the DCS

First author, year	Reliability	Validity
Koedoot,2001	Internal consistency - Cronbach's alpha obtained for the three subscales in both groups (0.61-0.75, 0.77-0.80, 0.83-0.81) and for the overall scale in both groups (0.75-0.82).	Mixed results from the two participant groups for construct and criterion validity.
Mancini, 2006 (Mancini et al., 2006)	Internal consistency - Cronbach's alpha obtained for the whole scale and subscales ranged 0.67 to 0.916, with the exception of 0.441 to 0.593 for the unsupported subscale.	Criterion validity and factor analysis outcomes varied, depending on whether the groups were patients or controls.
Song, 2006	Internal consistency- Cronbach's alpha of 0.81 for the complete scale, ranging from 0.50-0.73 for the subscales, with item to total correlation weakest for uncertainty items and one contributing factor item. Removing these raised the complete scale Cronbach's alpha to 0.84.	Construct validity: Spearman correlation coefficients were obtained for the DCS and Quality of Communication questionnaire (r=-0.44, p = 0.001), DCS and anxiety rating (r=0.47, p=0.006), with varying correlation between the subscales. Scale able to discriminate between control and intervention group.
Knapp, 2009	Internal consistency: Cronbach's alpha > 0.84.	Known-groups validity statistically significant (p<0.001 and P<0.05).
Katapodi, 2011	Internal consistency: Cronbach's alpha of 0.96	Convergent (r=-1.3 to r=-.31, p<0.05-.001), divergent (r = -0.30, p <0.01), predictive (r= 0.24, p<.001) & contrast (p<.001) validity were tested.
Linder, 2011	Internal consistency: ICCs ranged from 0.246-0.748 and a Cronbach's alpha of 0.80 obtained, although this involved excluding the supported subscale.	Discriminatory validity was based on whether a decision was made, and reported as statistically significant.

5.2.2 The Satisfaction with Decision Scale (SWD)

5.2.2.1 Descriptive features.

The Satisfaction with Decision (SWD) scale was developed to evaluate a decision-support intervention for hormone replacement therapy (HRT), although the scale was kept independent of a specific medical context. (Holmes-Rovner et al., 1996) It was intended both as an outcome measure and to provide insight into influences on treatment compliance.

The scale measures global satisfaction with a decision, and three attributes of effective decision-making based on those outlined by O'Connor in the DCS scale. (O'Connor, 1995) It is intended for use after a decision is made, but before any consequences have occurred. The six items use a five point Likert scale. It is self-administered and requires an 8th grade reading level. A higher score is associated with higher decision satisfaction.

The scale was piloted in a convenience sample of female university staff (n=120). For further evaluation, community volunteers wishing to gain information on menopause management were recruited via the local press into a randomised intervention trial (n=252). Three methods of decision support were included in the trial - brochures, lecture-discussion or individual. Participants were followed for twelve months and were noted by the authors "more likely to be white, college educated and [with] relatively high household income". (Holmes-Rovner et al., 1996) As the intervention concerned was specifically concerned with the menopause, 58% of the participants were still menstruating, while 59% had menopausal symptoms either before or at the time of the study.

5.2.2.2.Methodological quality

Table 14 outlines the measurement properties evaluated in the SWD scale development study.

Table 14: SWD development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in SWD development study
Reliability	
Internal consistency	Yes
Reliability	No
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	No
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

Internal consistency is termed reliability in this study. Though a Cronbach's alpha was calculated for each subscale separately, no factor analysis or other test of unidimensionality was performed. Three items were dropped at pilot stage due to poor internal consistency, with two more eliminated as they were considered condition specific and unsuitable for a generic scale. The sample sizes for both the pilot and main study were adequate in accordance with the COSMIN grading, at over a 100 or the number of items multiplied by seven (7 x 6 = 42). (Mokkink et al., 2012) The number of missing or incomplete items was given for the main study but not the pilot, with no information provided on how such items were processed.

Consideration of measurement error is only mentioned when the attributed variance is removed from a comparison of the correlation between SWD, DCS and Health Status Restriction (HSR) scales.

For content validity, although the theoretical foundations for the scale are addressed, it is not reported whether the items were chosen to reflect all aspects of the underlying construct, the target population or the intended purpose of the instrument. While the use of the scale as a predictive measure for decision uncertainty is evaluated, this is not clearly indicated in the initial aims of the study.

The four of the six hypotheses to be tested were reported a priori, with the predicted direction of any supporting correlation given. Limited information was provided about the comparator scales. In particular, several of the measures were developed within the study, with no detail given as to the development process or of their validity and reliability. While a good response rate is noted for the SWD scale, no information is provided for the comparator measures. No information was provided of the correlation coefficient used, making its interpretation challenging. The use of p-values is also questionable, as the direction and size of any correlation is considered of greater relevance. (Mokkink et al., 2012)

Though the study was concerned with the evaluation of decision support, there was no pre-intervention exploration of decision intention or administration of the scales. As such, the responsiveness of the SWD was not evaluated.

Limited information is provided to give context to the scores obtained by the SWD scale. While the mean and standard deviation of the scores are given, interpretability would be aided by the proportion of respondents achieving the highest and lowest score, subgroup analysis and consideration of MIC or MID. (Mokkink et al., 2012)

The study participants are noted to be largely homogenous in terms of ethnicity, education and income, and were by necessity restricted in age and sex due to the condition of interest. However, this carries consequences for the generalisability of a scale intended for use as a generic measure of decision-making.

5.2.2.3. Instrument quality

Table 15 details the measurement properties evaluated for the SWD scale.

Table 15: SWD development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	**Evaluated in SWD development study**
Reliability	
Internal consistency	Yes
Reliability	No
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Floor and ceiling effects	No
Interpretability	Yes

Principal component analysis is described rather than factor analysis, although it is used to assess discriminant construct validity of the SWD in comparison with the DCS and HSR measures, rather than to evaluate of the unidimensionality of the scale. The Cronbach's alpha of 0.88 for the pilot and 0.86 for the main study suggests an acceptable level of internal consistency. (Terwee et al., 2007)

The quality of the content validity is limited by the lack of clear description concerning item development and target population involvement. Construct

validity is supported by four specific hypotheses with concordant results, although not all of these were clearly outlined a priori. Interpretability is hampered by the absence of a MIC and subgroup analysis.

5.2.2.4. Further evaluation studies

The SWD scale had been cited 8 times by the end of 2014, as identified using Web of Science. One study met the eligibility criteria. Wills and Holmes-Rovner evaluated the use of the SWD scale in a sample of 97 depressed primary care patients, who were undertaking a new decision about antidepressant medication. (Wills and Holmes-Rovner, 2003) The wording of the scale was adapted for a decision concerning anti-depressant medication. The internal consistency was evaluated and a Cronbach's alpha of 0.85 obtained. The presence of significant correlations with other scales for knowledge, conflict and depression were used to assess construct validity. With the exception of the relationship with a scale for depression, all correlations were statistically significant.

5.2.3. Decision Attitude Scale (DAS)

5.2.3.1 Descriptive features

The Decision Attitude Scale (DAS) is intended to measure the perceived quality of both the process and outcome of decision-making. (Sainfort and Booske, 2006) In the development study, the decision of interest concerned healthcare plans, with further trials intended to study the impact of varying amounts of information on the decision-process. A decision about health plans was selected by the authors due to it being “complex, value-laden with uncertain long term outcomes”.

The scale contains ten items, two for each of the five separate subscales listed below:

- Evaluative attitude for decisional process
- Feelings about decisional process
- Behavioural attitude for decisional process
- Evaluative attitude for final choice
- Feelings about final choice

Each item is graded using a five point Likert, with a higher score on any item corresponding to a positive attitude towards the decision process and outcome. The scale is intended for completion immediately after the decision is made. Initially piloted on paper, the scale used in the trial was computerised.

The development trial involved 197 employees of the State of Wisconsin, which provides health insurance with a variety of health plan options. The average age of the participants was 44, 63% were female and most were well educated with a high reported household income. 56.7% had been employed by the State of Wisconsin for over 10 yrs. The group was chosen due to their experience of health plan evaluation.

5.2.3.2 Methodological quality

The following measurement properties were evaluated in the development study:

Table 16: Decision Attitude Scale development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in DAS development study
Reliability	
Internal consistency	Yes
Reliability	No
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	No
Other properties	
Responsiveness	Yes
Interpretability	Yes
Generalisability	Yes

The sample size is considered adequate for calculations addressing internal consistency by the COSMIN standards. (Mokkink et al., 2012) No account is given for missing or incomplete items, nor any detail for how such items would be handled. Factor analysis was performed, with one item subsequently excluded as it was considered to be in opposition to the others. Cronbach's alpha was calculated, but not for the individual subscales - as these contained only two items each, a correlation was felt to provide more information.

For content validity, the authors considered whether the scale sufficiently covered the theoretical foundation, although no detail was given for item development or target population involvement. For construct validity or hypothesis testing, supporting theories were outlined a priori, with the direction of the expected relationship described though not the magnitude. The authors predicted that "decision attitude [would be] better for individuals able to choose than those unable to choose". (Sainfort and Booske, 2006) An f-value was calculated, although this is a measure of statistical significance rather than a true value for correlation. In addition, only values for the

subscales are given rather than for the whole scale, and not all participants were given all subscales.

For responsiveness, two measurements were taken either side of a decision support intervention. However, no clear time interval and consideration of external influences on the participants during this period are described. No comparator instrument or gold standard was used, though a clearly stated hypothesis was given as to the predicted impact of the intervention of the Decision Attitude Scale score. Only analysis results were provided with a focus on subscale scores, with no underlying data or complete scale evaluation.

Interpretability is aided by the subgroup details provided, along with the mean and standard deviation scores for the scale. Highest and lowest scores are referred to but insufficient detail provided, such as the proportions of participants and the overall distribution of scores in the sample. No MIC or MID is given, with the actual difference in subscale scores often very small.

Evaluating the generalisability of scale is aided by the demographic and setting details provided. However, insufficient information is given for the sampling strategy, which is described as being an "appropriate random sample of respondents", but acknowledged as "…..not [being] representative of general population but chosen as [they] had experience in health-plan decisions". (Sainfort and Booske, 2006) In addition, it is unclear how many declined to participate or failed to fully complete the intervention and evaluation.

5.2.3.3. Instrument quality

The following table outlines the measurement properties evaluated for the Decision Attitude Scale in comparison to those outlined by the quality criteria (Terwee et al., 2007):

Table 17: Decision Attitude Scale development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	Evaluated in DAS development study
Reliability	
Internal consistency	Yes
Reliability	No
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	Yes
Floor and ceiling effects	No
Interpretability	Yes

Factor analysis demonstrated a three-factor scale with one overall dimension, and the sample size was adequate by COSMIN standards. The Cronbach's alpha for the complete scale was 0.86, within recommended parameters for internal consistency. For the subscales, that for the first factor, satisfaction with choice, had a Cronbach's alpha of 0.83. The subscale for usability of information had a correlation coefficient (r) of 0.64, while that for adequacy of information measured 0.49. Correlation coefficients were used here instead of a Cronbach's alpha as these subscales contained only two items.

The content validity of the scale is impaired by uncertainty over how the items were developed and who was consulted during this process. For construct validity, while four hypotheses are proposed, results are given for only one subscale, with these providing variable degrees of support for the relationships

expected to exist should the construct tested be accurately represented. (Terwee et al., 2007)

Doubtful design or method is used for the responsiveness, as none of the recommended statistical analyses such as MIC or area under the curve (AUC) calculations are used. (Terwee et al., 2007) In addition, the method is not reproducible and does not address external sources of bias. The absence of MIC also impacts on the interpretability of the results, despite the mean scores and subgroup analysis reported.

5.2.3.4. Further evaluation studies

The Decision Attitude Scale was cited 12 times by the end of 2014, when checked using Web of Science. No further evaluation studies meet the eligibility criteria.

5.2.4. Preparation for Decision-Making scale (PrepDM)

5.2.4.1. Descriptive features.

The PrepDM scale is intended to evaluate the patient's perception of how useful a decision aid is in supporting decision-making and communicating with a health professional. (Bennett et al., 2010) This study focuses specifically on evaluating the PrepDM scale, preceding its use in evaluating support for decisions concerning HRT, prostate cancer and breast cancer prevention options.

The scale contains ten items, and is self-administered. It is completed before and after the intervention. A higher score indicates a higher perceived level of preparedness for decision-making.

The study evaluates the scale in a sample of patients who were referred by orthopaedic healthcare providers for decision support, which took the form of a condition-specific decision aid delivered in a rural academic medical centre. Consecutive consenting patients were recruited from clinics.

A total of 400 people participated from the 966 referred, with a mean age of 57.3 years, and 55% were female. By orthopaedic condition, 127 of the participants were diagnosed with spinal stenosis, 42 knee osteoarthritis (OA), 105 herniated discs, 94 with chronic lower back pain and 32 hip OA.

5.2.4.2. Methodological quality

The Item Response Theory (IRT) approach to scale development was used in this study, differing to the Classical Test Theory (CCT) encountered so far. Both are methods of developing a measure and detailing its psychometric properties. (Streiner and Norman, 2008) CCT provides mainly scale-level information for the test sample. While its use is less dependent on pre-set assumptions, CCT exhibits "circular dependency" as the psychometric values obtained for the scale are dependent on both the test sample and measure as a whole. (Streiner and Norman, 2008) As such, subsequent scale application and interpretation are influenced by the original sample group and perceived item equivalence, where each item is considered to have the same properties. (Streiner and Norman, 2008) In comparison, IRT focuses instead on item-level information. It equates the probability of a specific item response to individual participant performance, and mathematically models the relationship between different levels of the examined trait and item scores. (Streiner and Norman, 2008) As such, the models can be used to explore test performances for particular levels of a trait, and visa versa. (Streiner and Norman, 2008) The use

of IRT is dependent on two main assumptions being met, which are unidimensionality, where the items measure only one construct, and item independence. (Streiner and Norman, 2008)

Additional standards are provided by COSMIN for studies using IRT. (Mokkink et al., 2012) In accordance with these, the model used in PrepDM evaluation was described as partial credit. However, it is unclear which computer software package was used or the estimation method applied. In addition, of the assumptions required for the IRT model, only unidimensionality is clearly accounted for.

The following measurement properties were evaluated in the development study:

Table 18: PrepDM development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in PrepDM development study
Reliability	
Internal consistency	Yes
Reliability	Yes
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	No
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

When considering the methodology for internal consistency, no information was given on the proportion of missing items or on how these were then handled. The sample size was adequate for the internal consistency calculation at 400, although some of the subgroup samples fall below the recommended

standard of 100 or (7x number of participants). (Mokkink et al., 2012) Internal consistency was checked for each subscale, with an alpha correlation calculated. For IRT, COSMIN recommend the calculation of a goodness of fit statistic at a global level, such as a chi-squared statistic, to compare the actual and expected response patterns. This is not evident in this study. Although a test reliability score is calculated, no additional information is provided on how this was produced.

The theoretical grounding for content validity is clearly reported, with items matched to the stages of the International Patient Decision Aid Standards (IPDAS) decision quality criteria and there is consideration of whether all aspects of the construct are addressed. However, the development does not seem to include the target group and the report lacks clarity for how the new items are produced.
For hypothesis testing or construct validity, the majority of the hypotheses were formulated a priori, with the direction but not magnitude of correlations predicted. However, the hypothesis for the DCS is not explicitly stated. Limited details are given for the comparator instrument, the DCS, though this may be due to word count restrictions. In addition, no information is given to verify the stage of decision-making questionnaire used. The statistical methods used to test the hypothesis are correlation coefficients, along with p-values.

Though pre and post intervention measurements were taken there is no clear description of evaluating responsiveness. For interpretation, the distribution of the scores in the sample is described, with values also given for each diagnostic subgroup. The proportions with the highest and lowest scores are not given, focusing instead how well the items discriminate between low and high levels of decision preparedness. The MIC and MID are not determined.

For evaluating generalisability, the mean age and sex of the participants are described, though there is no standard deviation for age, and other demographic details such as socioeconomic group and education level are absent. Limited detail is given for disease characteristics and the full setting of the study. Further information is also needed about the participant recruitment process, as there is potential for bias in the referral for a decision aid. No information is given about non-returned questionnaires and the characteristics of the non-responders, who accounted for more than half of those who gave their consent at clinic.

5.2.4.3. Instrument quality

The measurement properties produced by the study for the PrepDM are mapped against the quality criteria below. (Terwee et al., 2007)

Table 19: PrepDM development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	**Evaluated in PrepDM development study**
Reliability	
Internal consistency	Yes
Reliability	Yes
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Floor and ceiling effects	Yes
Interpretability	Yes

For internal consistency, the diagnostic subgroups had Cronbach's Alpha ranging from 0.92 to 0.96. The higher value falls above the recommended value range, suggesting there may be redundancy among items. (Terwee et al., 2007)

It is unclear which results belong to each diagnostic subgroup. The item-total values range from 0.75 to 0.81 for each subgroup, within the quality criteria range. (Terwee et al., 2007) The sample of 400 is divided into five patient groups numbering from 32 to 127, which diminishes the power for statistical analysis. (Terwee et al., 2007) Little detail is given for the evaluation of reliability, though the total test reliability calculated is high at 0.944

Limited information is provided in support of content validity, such as describing item development and target population involvement. For construct validity, a priori hypotheses are described for only some of the outcomes, though there is full congruence between the PrepDM and the proposed construct markers.

While poor reliability is described at extreme scores there is no clear discussion of floor and ceiling effects, such as proportions and numbers. For interpretability, mean scores with their standard deviation are provided for diagnostic subgroups. The IRT analyses suggest good discrimination values for the items, with the scale measuring well across all but the extremes of readiness for decision-making. No MIC is provided.

5.2.4.4. Further evaluation study

The PrepDM scale had been cited 20 times by the end of 2014, as identified with the Web of Science. No eligible further evaluation studies for the PrepDM were identified.

5.2.5. The Shared Decision-Making Questionnaire (SDM-Q)

5.2.5.1. Descriptive features

This scale was developed following a research consortium on shared decision-making in Germany from 2001 to 2005. (Simon et al., 2006) A total of ten projects explored shared decision-making interventions in different clinical scenarios, including breast cancer and hypertension. Pre-existing measures were found to have poor psychometric rating on translation into German. A new instrument was therefore developed, intended to explore "patient preferences for information and participation as well as the process and outcome of decision-making". (Simon et al., 2006)

The original self-administered scale had 24 items, nine of which were dichotomous and fifteen used a four point Likert scale. The measure was to be completed immediately after a consultation.

Set in both primary and secondary care, 773 participants were recruited with 32 excluded as their responses missed out over 30% of the scale. For the general practice consultations, decisions involved medical management (n = 210) or treatment for depression (n = 230). In secondary care, urology patients (n = 66) completed a score following a consultation concerning non-emergency surgery, the gynaecology clinic patients (n = 111) considered a decision regarding in-patient treatment for breast cancer and the anaesthetic consultations (n = 156) covered preparation and discussion of analgesia for several forms of non-emergency surgery. The participants had a mean age of 51.9 years with a standard deviation of 16, 59.1% were female and 66% had received higher education.

5.2.5.2. Methodological quality

The Item Response Theory (IRT) model was utilised in this study. Applying the COSMIN standards, the model used was clearly described as the partial credit model and the software identified as Winsteps, although the method of estimation used within the model (such as conditional maximum likelihood or marginal maximum likelihood) is not fully described. (Mokkink et al., 2012) The assumptions required for the IRT model including unidimensionality, local independence and item fit were checked.

Table 20 outlines the measurement properties evaluated in the SDM-Q development study, with the COSMIN checklist for methodological quality then applied to each. (Mokkink et al., 2012)

Table 20: SDM-Q development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in SDM-Q development study
Reliability	
Internal consistency	Yes
Reliability	Yes
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	No
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

In assessing the evaluation of internal consistency, a clear description is given of the number of missing items and how these were subsequently dealt with. Participants were excluded if the scale was more than 30% incomplete (n=32).

Of the remaining participants, fewer than 5% of items were incomplete and these were handled using multiple imputation. The total sample size was adequate for testing internal consistency, but not in the subgroups as one contained only 66 participants. (Mokkink et al., 2012) A goodness of fit statistic, in-fit mean square, was calculated to assess for items disrupting the unidimensionality of the scale, and reliability values for person and item parameters.

For content validity, a clear account is given of the theoretical foundation and expert involvement in item development. Consideration was given to construct coverage and the target population were involved. There is a lack of clear, a priori hypotheses for the construct validity, with little detail given for the comparator instruments. Correlation coefficients are used to test the hypotheses.

Interpretability of the scale is aided by the use of IRT to evaluate score distributions, person fit and differential item functioning among different patient subgroups. Additional information concerning the MIC would have added further clarity. Applicability of the scale in other populations, or generalisability, can be judged using the clear description of the study setting and participant demographics provided. However, the IRT model revealed poor scale parameters in the urology subgroup, which was subsequently removed from further analysis. This limits the generalisability of the scale, a concern further exacerbated by the limited detail given for participant sampling, recruitment, and the characteristics of non-respondents.

5.2.5.3. Instrument quality

Table 21 maps the measurement properties for the SDM-Q against quality criteria developed by Terwee and colleagues. (Terwee et al., 2007)

Table 21: SDM-Q development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	Evaluated in SDM-Q development study
Reliability	
Internal consistency	Yes
Reliability	Yes
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Floor and ceiling effects	IRT equivalent
Interpretability	Yes

Four items were discarded due to poor fit with the overall scale and underlying construct, the remaining items having item-fit measures of 0.78-1.14. As described above, IRT methods mathematically model the relationship between different levels of the examined trait and item scores. (Streiner and Norman, 2008) A reliability score of 0.77 was obtained for the person parameters and 0.95 for item parameters, with Winsteps software guidance indicating that the person reliability is equivalent to the traditional test reliability. (Winsteps Software, 2014)

Content validity is well covered, with clear descriptions of measurement aims, consideration of construct coverage and involvement of expert and target population in the development of the scale. Construct validity is less well covered, with a clear a priori hypothesis generation lacking and low correlation values between SDM-Q and the comparator scales.

While interpretability is aided by subgroup analysis, further detail such as the MIC would have further enhanced this measurement property. (Terwee et al., 2007) The IRT analysis for participant responses indicates only 29.8% of the participants completed the scale in a pattern consistent with its proposed model. Differential item functioning suggests a high variation between subgroups in the interpretation and answering of items. The distribution of person and item fit parameters do not complement each other, with a high ceiling effect noted for person parameters. The latter suggests that items of higher difficulty are needed to differentiate between individuals reporting greater shared decision-making.

Overall, the authors conclude that the scale was designed to cover too much, in view of the broad aims of measuring patient preference as well as the process and outcome of decision-making. (Simon et al., 2006)

5.2.5.4. Further evaluation studies

The SDM-Q study had been cited 34 times by the end of 2014, as checked with the Web of Science. One study was identified for the further development of the SDM-Q. This produced a new scale, SDM-9. (Kriston et al., 2010) However, this was not eligible for inclusion as it was a measure of the extent of SDM in a consultation, rather than decision-making process from the patient perspective.

5.2.6. The SURE scale

5.2.6.1. Descriptive features

The SURE scale was developed as a rapid screening test for decisional conflict, styled on the CAGE alcohol dependency tool. (Légaré et al., 2010a) The latter

uses four questions to determine the extent of alcohol consumption. The SURE scale includes four dichotomous items based on the core concepts of the Ottawa Decision Support Framework, with the exception of decision effectiveness as this was not considered applicable to all stages of decision-making. The test is self-administered in a paper format. A score of less than four indicates decisional conflict.

The scale was developed in French and English. For the French version, the participants were pregnant women recruited from four family medicine groups in Quebec City. The scale was administered after the first prenatal consultation, during which screening for Down's Syndrome was discussed. Those at high risk of a Downs' syndrome were excluded from the study. From the 180 women approached, 148 were recruited with 21 found to be ineligible and 11 declining. The participants did not reflect a broad range of educational backgrounds, as 96% had some college education, a high school diploma or higher. Response rates for the included scales varied, with 141 participants completing the DCS and 123 completing the SURE scale.

For the English version, participants were recruited from a rural academic medical institution. The scale was administered after the use of a decision aid, which concerned decisions relating to a broad variety of medical conditions such as musculoskeletal conditions, prostate and breast cancer. While 1474 patients completed the scale, there was a poor response rate as only 34% completed and returned the questionnaire. Of these 1474 participants, 52% were women and 93% had received some college education, a high school diploma or higher.

5.2.6.2. Methodological quality

Table 22 outlines the measurement properties evaluated in the SURE development study, with the COSMIN checklist for methodological quality then applied to each. (Mokkink et al., 2012)

Table 22: SURE development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in SURE development study
Reliability	
Internal consistency	Yes
Reliability	No
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	Yes
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

For the French version of the scale, 25 of 148 questionnaires were incomplete and therefore excluded. Although only 34% of the English version of the scale were completed and returned, it is unclear how many of the scales were incomplete. There is also limited information about which items were incomplete and whether this varied between the versions of the scale. For internal consistency, the unidimensionality of the scale was tested with factor analysis, and a Cronbach's Alpha calculated. Intra-rater reliability was not evaluated as the authors felt that decisional conflict represented a "state rather than a trait", rendering this measurement property inappropriate. (Légaré et al., 2010a) In addition, the self-administered nature was felt to preclude inter-rater reliability.

When evaluating the content validity, limited information is given concerning how the items were developed. The scale is based on four core aspects of the Ottawa Decision Support Framework, and was tested on experts and graduate students taking a clinical course in decision support. However, there is no discussion of the involvement of the target population or of the adequacy of construct coverage.

For the hypothesis testing, the authors believed that the score would discriminate between those who were able to make a decision about treatment options and those who could not. However, a clear a priori hypothesis is lacking, as is a predicted magnitude and direction for the relationship. A t-test was used rather than the recommended correlation coefficient. While the authors specify the comparison of SURE with the DCS as a measure of criterion validity, little evidence is given to support the use of the DCS as a gold standard. (Mokkink et al., 2012) The COSMIN standards also state that only the original version of shortened scales can be used as a gold standard for patient reported outcomes. (Mokkink et al., 2012) Only the French version of the scale is tested against the DCS, limiting the extrapolation of these findings to the English version of the scale. In addition, the sample size for this property is small, with only 148 participants in comparison with 1474 for the English scale. There was a higher response rate for the DCS in comparison with the SURE scale (141 versus 123 respectively), despite the latter being shorter.

Facets of interpretability are clearly described, with details given for score distribution, proportions achieving the highest and lowest scores and a threshold level for clinical significance. Generalisability of the score can be evaluated, as clear demographic details are provided along with the context for

the decision considered and a description of the convenience sampling method used. This could have been further improved by including details such as socioeconomic status and ethnicity. In addition, the response rates of 34% for the English SURE and 82% for the French SURE may indicate a responder bias, as it is feasible that the non-responders might be experiencing greater decisional conflict. Limited information is given about the non-responders.

5.2.6.3. Instrument Quality

Table 23 maps the measurement properties for the SURE scale against quality criteria developed by Terwee and colleagues. (Terwee et al., 2007)

Table 23: SURE development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	**Evaluated in SURE development study**
Reliability	
Internal consistency	Yes
Reliability	No
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	Yes
Other properties	
Responsiveness	No
Floor and ceiling effects	Yes
Interpretability	Yes

The factor analysis for the French version of the SURE scale indicated that the items loaded on two factors or constructs rather than being unidimensional. The first factor was knowledge, value and certainty, with the second described as support. Furthermore, the support item negatively correlated with the other factor and poorly correlated with the overall score. This version of the scale had a Cronbach's alpha of 0.54, which lies outside the recommended range of 0.70

and 0.95. (Terwee et al., 2007) The factor analysis for the English SURE found that the scale was unidimensional, with all items loading onto one factor. For this version of the scale, one factor accounted for 49% of the variance. A Cronbach's alpha of 0.65 was obtained, which is also below the recommended range. (Terwee et al., 2007)

Content validity is impeded by the lack of target population involvement in the development of the items. For construct validity, only a partial hypothesis was constructed a priori and it was difficult to judge whether 75% of the results were in concordance with this, as recommended by the quality criteria. (Terwee et al., 2007) There was little evidence provided to support the use of the DCS as a gold standard, and the correlation between the DCS and SURE was moderate at r = -0.46. (Terwee et al., 2007)

For interpretability, a threshold score indicating decisional conflict was suggested, though not a minimal important change (MIC) or grading of the scores in relation to degrees of conflict. For the French SURE, 85% of respondents had the highest score while only 1% scored the lowest possible. For the English SURE, 67% had the highest score while 1% had the lowest. There is a heavy bias evident for scoring highly on this scale, with few of the respondents reporting scores consistent with decisional conflict.

5.2.6.4. Further evaluation studies

This scale has been cited 9 times, as identified using a Web of Science search at the end of 2014. The only eligible further evaluation study was by Ferron Parayre and colleagues in 2014. (Ferron Parayre et al., 2014) Further evaluation of the SURE scale is performed using data from a clustered randomised trial of a decision aid for the use of antibiotics in treating acute

respiratory tract infections (RTI). Patients diagnosed with an acute RTI, where either the patient or clinician had considered the use of antibiotics, were recruited from family practice teaching units. Of the 712 patients recruited, 654 completed both SURE & DCS. The patients recruited were mostly women (64%), adult (71%), in employment (72%) and had received higher education (60%).

Internal consistency of the SURE scale was evaluated using a Kuder-Richardson 20 coefficient, which was calculated as 0.70 and considered adequate by authors. The DCS and SURE scores were compared using the Spearman correlation coefficient. A moderate negative association was suggested with a correction coefficient of -0.45, with a statistically significant p-value of less than 0.0001. Using the SURE scale, decisional conflict was defined as a score of 3 or less. The sensitivity and specificity of the scale in detecting decisional conflict was then evaluated using DCS scores. A sensitivity of 94.3% was obtained, with a 95% confidence interval of 78.9 to 99.0% and specificity of 89.8%, with a 95% confidence interval of 87.1 to 92.0%. Two weeks later, the participants were evaluated for decisional regret, adherence to decision and attendance for further consultation. While a correlation was noted between decision regret and decision conflict as measured using the SURE scale, adherence to decision and re-consultation rates did not correlate with SURE scores. Sensitivity did not change with gender, clinician exposure to decision aid and did not differ with decision outcome. Overall, the evaluation approach taken for this study concerned screening more than the psychometric properties of a measure.

5.2.7. The COMRADE scale

5.2.7.1 Descriptive features

The COMRADE scale was developed to measure both risk communication and the effectiveness of decisions made during consultations. (Edwards et al., 2003) The scale is self-administered and intended for completion after the consultation. It contained 43 items at first testing, with higher scores indicating a better outcome from the consultations.

A pilot was carried out in five South Wales general practices, with the subsequent trial recruiting from twenty practices in urban, semi-rural and rural settings. The study formed part of a larger trial of risk communication and shared decision-making interventions for doctors in the UK, and the decisions concerned prostatism, atrial fibrillation, menorrhagia and menopausal symptoms. The pilot had 72 participants, with a mean age of 45.9 years with 51% female participants. For the trial, consent was obtained from 1135 patients, 43.9% of those approached. Of these, 960 were randomly selected and invited to attend an appointment. The participants had a mean age of 59 years and 58.8% were female. There were 335 non-attenders for the allocated appointments. This group had a younger mean age and a higher proportion of women. The trial was divided into three data collection points: baseline, first intervention and second intervention phases. Weekly reminders were sent out to encourage return of the questionnaires. In total, 715 (95.7%) of questionnaires were returned. From the baseline stage of the trial, 133 of 197 (67.5%) of the full 43 item questionnaires were complete.

5.2.7.2. Methodological quality

The following section outlines the measurement properties evaluated for the COMRADE scale, with the COSMIN checklist for methodological quality applied to each. (Mokkink et al., 2012)

Table 24: COMRADE development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in COMRADE development study
Reliability	
Internal consistency	Yes
Reliability	No
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	No
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

In evaluating internal consistency, clear descriptions are given of the number of missing items, and how these were handled during statistical analysis. Only complete responses from the first 197 questionnaires returned were used for the factor analysis. Of these baseline responses, 67.5% had full data. There was a significant difference noted between those who completed the questionnaire and those who did not, in terms of age and the physical condition diagnosed. Factor analysis was used to evaluate unidimensionality, and Cronbach's alpha calculated for internal consistency.

For content validity, the items were generated following a systematic literature review and semi-structured focus group interviews with consumers and patients. Interviews were also conducted with general practitioners identified as key informants. The same groups also reviewed the questionnaire, with the

wording subsequently amended before it was piloted in a group similar to the target population. The authors highlight that this was an iterative process, with different groups involved at different stages of review in order to avoid overfamiliarity with the scale. Twenty semi-structured interviews were then performed with participants from the pilot to explore their views on the questionnaire and the consultation, and so assess congruence with their initial questionnaire responses.

Clear hypotheses were described a priori for construct validity, with the expected direction and magnitude of the correlations. Detailed descriptions, including psychometric properties, were provided for the comparator instrument measuring enablement, though less information was provided for that evaluating anxiety.

When considering the interpretability of the scale, the distributions of the scores are reported. These were also illustrated in a box plot with consideration of floor-ceiling effects, although the raw numbers and proportions involved are not clearly reported. No MIC or MID are given, though this may be due to the early stages of scale development.

A comparison is made between non-responders and responders, and those returning complete and incomplete questionnaires. The authors note a significant difference between the groups, which impacts on the generalisability of the study findings. In addition, while details are given for the study setting, patient demographics and diagnoses, more information concerning disease characteristics, such as severity and duration, would be beneficial. In terms of recruitment, patients were invited to participate in a trial if they had one of the conditions of interest. A random 960 people were then

selected and invited to an appointment. No further information is given for inclusion criteria or the randomisation process.

5.2.7.3. Instrument Quality

Table 25 maps the measurement properties for the COMRADE against quality criteria developed by Terwee and colleagues. (Terwee et al., 2007)

Table 25: COMRADE development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	**Evaluated in COMRADE development study**
Reliability	
Internal consistency	Yes
Reliability	No
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Floor and ceiling effects	Yes
Interpretability	Yes

For the initial factor analysis, 197 responses were used. This is too small a sample size for 43 items if the quality criterion of number of items multiplied by seven is used. (Terwee et al., 2007) Three factors were identified, with twenty items that failed to load on to a factor eliminated. The third factor, support, was transient and its presence varied in the different trial phases. The Cronbach's alpha for the whole scale is 0.92, which falls within the recommended levels for internal consistency. However, as subscales were identified, statistical analysis should also have been performed on them. (Terwee et al., 2007)

The instrument scores strongly on content validity, with target group involvement and consideration of construct coverage although there is limited expert involvement evident for item selection. Five hypotheses were tested, with clear a priori predictions. Four of these had supporting results. The correlation between patient concerns about treatment with risk communication was not reported.

For interpretation, the mean and distribution of scores are provided and illustrated with a box plot, which show that the confidence in decision subscale was scored highly. In addition, clear proportions are not given and therefore cannot be said to comply with the quality criteria of no more than 15% achieving the highest or lowest scores. No MIC is defined, though this is arguably less relevant at an early stage of instrument development.

5.2.7.4. Further evaluation studies

The COMRADE scale development study has been cited 53 times by the end of 2014, as identified by Web of Science. One eligible study further evaluating the scale was identified. Knapp and colleagues, 2009, validated the properties of both DCS and COMRADE for children with life-limiting illness, specifically for the parents involved in paediatric palliative care programmes. (Knapp et al., 2009) A sample of 936 was contacted with 266 completed surveys returned. For internal consistency, the Cronbach's alpha was over 0.93. Confirmatory factor analysis did not replicate the original structure of COMRADE and mixed results were obtained for the known groups construct validity. The authors conclude that the DCS was better suited for use in the context.

5.2.8. The Decision Evaluation Scale (DES)

5.2.8.1 Descriptive features

The Decision Evaluation Scale is intended to measure factors influencing the patient's appraisal of treatment options. (Stalmeier et al., 2005) Measure development was set within a decision support intervention trial for women with or without breast or ovarian cancer who had chosen to undergo genetic testing. Possible decision outcomes included prophylactic mastectomy, breast cancer screening or remaining undecided. The self-administered scale contained 36 items at first testing, each with a five point Likert score.

The trial was performed in family centre clinics, where 453 eligible patients were identified, of which 390 consented to participate. Half of the women received a brochure and video with information about the decision. Questionnaires were sent out at baseline (T1), the time of the blood test used for screening, then at 4 weeks after the test (T2), 2 weeks after a positive results (T3) and 3 months after a positive result (T4) and onwards (T5). 368 women were still present in the study at the T2 stage. 22 women were ineligible to continue in the trial as they underwent bilateral mastectomy and 3 participants were excluded due to incomplete data. As such, 343 participants remained for the measure evaluation. 91 participants were positive for the BRCA1 or 2 genes, three of which then withdrew from the trial. At T4 stage of follow up, 88 participants remained, with 87 continuing beyond this to the next evaluation stage (T5).

5.2.8.2. Methodological quality

The following section outlines measurement properties evaluated for the DES scale, with the COSMIN checklist for methodological quality applied to each. (Mokkink et al., 2012)

Table 26: DES development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in DES development study
Reliability	
Internal consistency	Yes
Reliability	No
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

For internal consistency evaluation, a clear account was given of the number of missing items and how these were handled. One item was deleted as it was missed in too many of the responses. An average 1.5% of the item responses were incomplete after these were eliminated, with 299 (87%) of the participants completing all 35 remaining items. Missing data analyses were performed on decision items. For incomplete items from multiple–item scales, an imputed mean was calculated when at least half of the items were completed. The sample size was adequate for assessing internal consistency and for evaluating unidimensionality, despite the large number of items involved (7*35= 245). (Mokkink et al., 2012) Factor analysis was performed, unidimensionality checked and a Cronbach's alpha calculated for each identified subscale.

Experts were involved in item development, with consideration of construct coverage by the scale. However, content validity is impaired by no evident involvement of target group in measure development or any evaluation of the relevance of the items to the population of interest. For construct validity, hypotheses were developed after the scales were identified with factor analysis but before the relationship between the DES and other measures was tested. The direction but not the magnitude of expected correlation was outlined. Limited detail was provided for the comparator measures and it is also unclear which correlation coefficient is used.

The scale was originally developed in Dutch before 15 items were translated into English by the first author, with an independent translation undertaken by a professional translator. Any discrepancies were resolved "by consensus". (Stalmeier et al., 2005) There are limitations in the approach taken, as the proficiency of both translators for the medical context and languages are not explicitly outlined. There is no "forward and backward" translation to confirm that meaning of the items has been preserved through the translation process. (Mokkink et al., 2012) In addition, there is neither review of the translated scale by the original committee nor re-testing of the factor analysis and other measurement properties.

The interpretability is limited by the provision of mean scores and their standard deviations only, grouped by treatment choice. There is no evaluation for floor-ceiling effects or statement of a MIC. For evaluating generalisability, the trial setting and participant characteristics, such as demographic and medical condition details, are provided. However, limited detail is given for the sampling strategy and it is difficult to keep track of the number of respondents involved in each stage of the measure evaluation.

5.2.8.3. Instrument Quality

Table 27 maps the measurement properties for the DES against quality criteria developed by Terwee and colleagues. (Terwee et al., 2007)

Table 27: DES development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	Evaluated in DES development study
Reliability	
Internal consistency	Yes
Reliability	No
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Floor and ceiling effects	No
Interpretability	Yes

Factor analysis was performed, indicating the presence of three subscales with each containing five items. The remaining items were discarded. The subscales concerned satisfaction-uncertainty, informed choice and decision control. Correlations between the subscales were moderate. For internal consistency, the subscale Cronbach's alphas are 0.79, 0.85 and 0.75 respectively which are within the acceptable range of 0.70 to 0.95. (Terwee et al., 2007)

Content validity is limited by the absence of target group involvement. For construct validity, there is doubtful design in terms of hypothesis generation timing and the majority of correlations obtained are weak. The DES is tested against a broad number of comparator constructs. These include anxiety, strength of treatment preference, depression, avoidance, subjective knowledge, amount of information, satisfaction with quality of information, negative emotions and partner agreement. All but two have correlations statistically significant at $P < 0.001$. However, the correlations obtained are weak except

between strength of preference and satisfaction-uncertainty (0.64), and between informed choice and the three information/knowledge-related scores (0.52, 0.61 and 0.58). DES scores were significantly worse for undecided participants.

The participant demographic details provided allow evaluation of the scale generalisability to different contexts. However, these details indicate that women who are married, employed and with a high school level of education are most represented by this sample. The interpretability of the scale can be considered using the mean scores and standard deviations provided, but this is limited by the absence of MIC or consideration of floor-ceiling effects. (Terwee et al., 2007)

5.2.8.4. Further evaluation studies

The DES development study had been cited 19 times by the end of 2014, as identified by Web of Science. The only eligible study addressing further evaluation of this scale was by Erci and colleagues, 2008 who adapted the DES for cancer patients in Turkey. (Erci and Özdemir, 2008) The trial used a convenience sample of 199 patients. Back translation was used in adapting the scale, but there was no target group involvement. For content validity, a panel of specialists reviewed the scale, with additional pre-testing in 30 patients from medical oncology groups. Factor analysis identified three factors: satisfaction-uncertainty, informed choice and decision control. For internal consistency and homogeneity, Cronbach's alphas for the three factors were 0.74, 0.75 & 0.71, with item-total correlations ranging from 0.36 to 0.71. Evaluation of test-retest reliability indicated Pearson correlations of 0.74 (complete DES), 0.71 (satisfaction-uncertainty), 0.78 (informed choice) and 0.70 (decision control).

The authors noted the need to test further in different regions and populations in Turkey.

5.2.9. The Decision-Making Quality Scale (DMQS)

5.2.9.1. Descriptive features

The DMQS is intended to evaluate the decision process used by the respondent, and allow the targeted development of what is defined by Janis and Mann as a quality decision style. (Hollen, 1994, Janis and Mann, 1977) Its anticipated use was both as a measurement and counselling tool. In the study, respondents were asked to consider their approach to consequential decisions, described as "important choices, not everyday ones". (Hollen, 1994)

Two forms were developed, Form S for self-administration and Form O for completion by observers. The readability of the scale was evaluated, and the completion time estimated at 3 to 5 minutes. The seven items were graded using a four point Likert scale. Scores were evaluated using two indices: a Total Adherence Index, where a higher score indicated greater adherence to the desired quality decision style, and Quality Index, which had a binary outcome of a quality or non-quality decision.

Decisions were evaluated in two settings. The first two groups were healthy high school students attending a health course (n=147 and 374). The other two groups were children attending childhood cancer clinics with their parents for follow up care. Of the patients, 36 were adolescents aged 14 to 19 years, 19 were young adults aged 20 to 26 years. All were five years post-diagnosis and two years clear of treatment. 67 parents of adolescent patients participated, completing a mailed DMQS. Two clinic nurses who were familiar with their

decision-making during the acute illness stage also graded the parents. The participants were recruited at the health course or clinic.

5.2.9.2. Methodological quality

The following section outlines the measurement properties evaluated for the DMQS scale, with the COSMIN checklist for methodological quality applied to each. (Mokkink et al., 2012)

Table 28: DMQS development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in DMQS development study
Reliability	
Internal consistency	Yes
Reliability	Yes
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	No
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

No information is given about the number of missing items or how these were handled, while the sample size included was adequate by the COSMIN standards. No details are given to suggest that the unidimensionality of the scale was assessed by factor analysis. However, a Cronbach's alpha was calculated for internal consistency. Rather than test-retest reliability for the participant reported measures, the observer scale was tested for intra-rater and inter-rater reliability using Kendall Tau and Kappa statistics. While the latter is supported for an ordinal scale, there is no detail to suggest it was weighted as recommended in the standards. Two measurements were made one month

apart, though it is unclear whether the sample and test conditions were stable during this time.

For content validity, a panel of three experts was consulted and there was a high rate of agreement for coverage of theory, domain representation and individual item coverage. However, there is no reference to tailoring the scale to a target group, nor of target population involvement. The hypotheses tested for construct validity were not explicitly outlined a priori, with correlations and p-values used to evaluate the strength of the relationships observed.

Evaluation of scale generalisation beyond this study is impacted by the limited demographic details provided for the participants. In addition, while clear details are given for the setting of the trial, little is given for the sampling strategy or the response rate. Limited information is provided to evaluate the interpretability of the scale, such as mean scores, score distributions, subgroup results or numbers achieving the highest and lowest scores.

Other limitations include that the participants are asked to consider their decision-making generally, rather than for a decision they were actively working through. This is open to recall bias and may be more marked when considering past decisions of consequence. Limited information is given for the processes used in the observer assessments, which involved only two people. These observers were also involved in the care of the family and bias may be introduced both from emotional connection and knowledge of the health outcome.

5.2.9.3. Instrument quality

Table 29 maps the measurement properties for the DMQS against quality criteria developed by Terwee and colleagues. (Terwee et al., 2007)

Table 29: DMQS development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	Evaluated in DMQS development study
Reliability	
Internal consistency	Yes
Reliability	Yes
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Floor and ceiling effects	No
Interpretability	Yes

Internal consistency is hampered by the absence of a factor analysis, though Cronbach's alphas for the five groups involved span from 0.71 to 0.90, all within the quality criteria. (Terwee et al., 2007) Insufficient details for statistical analysis also impacts on reliability, as it is uncertain whether a weighted kappa statistic was used. Intra-rater reliability varied between the two clinic nurses, from 0.57 to 0.92. For inter-rater reliability, an average Kappa statistic of 0.28 was obtained, suggesting poor agreement. (Terwee et al., 2007) The study author also notes that the observers tended to score the participants highly for adherence with the decision-making quality criteria.

No target group involvement reduces the instrument content validity, while construct validity meets the quality criteria with specific hypothesis formulated with the majority of the results in support with these. However, it is unclear to what extent the hypotheses were formulated a priori and limited information is provided for the comparison instruments. The DMQS scores for each group

were compared with a self-reported risk behaviour scale, with higher quality decision-making criteria expected to correlate with lower risk behaviour. The correlations for the two groups of high school students were – 0.23 ($p < 0.01$) and – 0.17 ($p < 0.1$) respectively, and -0.52 ($p < 0.001$) for the clinical group. The correlations are therefore weak to moderately negative, supporting the hypothesis.

5.2.9.4. Further evaluation studies

The DMQS development study has been cited 24 times by the end of 2014, as identified by Web of Science. No further development or evaluation studies matched the eligibility criteria.

5.2.10. The Decision Self-Efficacy Scale (DSES)

5.2.10.1. Descriptive features

The DSES is intended to measure the respondent's perception of their ability to complete varying stages of the decision-making process, though the authors focused on the social component of engaging with a healthcare team. (Bunn and O'Connor, 1996) The scale has 11 items, using a five point Likert scale, though this was reduced to three points after initial development work.

Originally developed in the context of hormone replacement therapy (HRT) decisions, the scale was evaluated over a ten-week period at a clinic specialising in the treatment of schizophrenia at a large Ottawa hospital. Another scale, the Decision Emotional Control Scale (DECS), was also assessed. The sample of 94 comprised of 68 men and 26 women, with an age range of 27 to 68 and mean age of 41 years. All were able to speak English and were currently taking long-

acting anti-psychotic medications, 86% doing so for over 5 years with 2% for less than one year. Only those with stable schizophrenia were included, with exclusion of those experiencing acute psychosis. The participants were given information about long-term medication injections in a one-to-one setting with a research assistant, and then asked to make a choice – to accept the treatment, decline it or delay the decision. The DCS and DSES were then completed. 4% of the sample described difficulty concentrating on the entire scale during the testing process.

5.2.10.2. Methodological quality

The following section outlines the measurement properties evaluated for the DSES scale, with the COSMIN checklist for methodological quality applied to each. (Mokkink et al., 2012)

Table 30: DSES development study mapped against the COSMIN checklist for methodological quality

COSMIN methodological quality	Evaluated in DSES development study
Reliability	
Internal consistency	Yes
Reliability	No
Measurement error	No
Validity	
Content validity	Yes
Hypothesis testing	Yes
Cross-cultural validity	No
Criterion validity	No
Other properties	
Responsiveness	No
Interpretability	Yes
Generalisability	Yes

While the total percentage of missing items was reported, there was insufficient detail about these items and how they were handled during analysis. In addition, the sample size was insufficient by the COSMIN

standards at less than a hundred participants and there was no evaluation of the unidimensionality of the scale. (Mokkink et al., 2012) A Cronbach's alpha was calculated to evaluate internal consistency.

Content validity was assessed as face validity, with a panel of experts and a client pilot. As such, the target population was included in the development of the items and there was consideration of the theoretical, domain and item coverage. However, the pilot was small with only four participants. The known groups approach was taken for hypothesis testing, using the treatment choice made as a proxy for decision self-efficacy. Hypotheses were formulated a priori, with the direction but not the magnitude of the relationships predicted. However, no correlations were calculated, only mean differences in test scores, standard deviations and p-values.

Limited information was given to support the interpretability of the scores beyond the mean score obtained and the standard deviation. Details of score distribution in the group, subgroup results and the proportions achieving the highest and lowest scores would have improved this measurement property. Demographic and setting information were given but not in enough detail for the sampling strategy, response rate or impact of the background condition on cognition, as most of the clients accepted the support of a research assistant in reading the scales. These aspects limit the generalisability of the scale.

Other aspects to consider include that the scale asks people to consider their perceived ability to make a decision rather than the decision process itself, although the former will influence the latter. In addition, as eligibility to participate was dependent on clinical stability, the decisional context is relatively low in conflict, which may influence the results obtained.

5.2.10.3. Instrument Quality

Table 31 maps the measurement properties for the DSES against quality criteria developed by Terwee and colleagues. (Terwee et al., 2007)

Table 31: DSES development study mapped against the quality criteria

Terwee et al Instrument Quality Criteria	Evaluated in DSES development study
Reliability	
Internal consistency	Yes
Reliability	No
Agreement	No
Validity	
Content validity	Yes
Construct validity	Yes
Criterion validity	No
Other properties	
Responsiveness	No
Floor and ceiling effects	No
Interpretability	Yes

Both content validity and construct validity match the quality criteria outlined by Terwee and colleagues. (Terwee et al., 2007) Internal consistency is limited by the absence of factor analysis, though a Cronbach's alpha of 0.84 is obtained. No further consideration of reliability is reported. Interpretability is limited by the absence of subgroup analysis and consideration of floor-ceiling effects.

The small number of participants in the pilot study limits content validity, with all other criterions covered. For construct validity, clients who were unsure or wanted to delay their decision scored higher on the DSES, indicating difficulty with decision self-efficacy. Comparing the mean scores and standard deviations between those who continued treatment and those who were uncertain found a difference statistically significant at the value of 0.05 ($p = 0.037$), which supports the hypothesis but no correlation statistics were calculated.

5.2.10.4. Further evaluation studies

The DSES development study has been cited 48 times by the end of 2014, as identified by Web of Science. No further eligible development or evaluation studies were identified.

5.3. Narrative synthesis of methodological and instrument quality

Table 32 summarises the methodological quality behind the development of the included scales, with the adequacy of the measurement properties of each included in Table 33.

A four-point scale of the COSMIN checklist was used to summarise the methodological quality of the included studies. For each facet of scale development methodology, a "worse score counts" approach is taken. (Terwee et al., 2012) For example, if one constituent element of the internal consistency methodology is graded as poor, then internal consistency is graded as poor overall. The summary focuses on the ten developmental studies as the further evaluation studies consider adapted versions of the scales or their suitability in new population groups. The adequacy of scale measurement properties are summarised as recommended by criteria authors. (Terwee et al., 2007)

The methodological quality is highly variable, both between and within studies. None of the scales consistently score well for measurement properties when compared with the criteria of adequacy. (Terwee et al., 2007)

Recurrent themes were identified in the summarising process:

- Explanations for choice of sample size, along with descriptions of recruitment processes and how missing items on the scales were handled, were often absent.
- Factor analysis to confirm the unidimensional nature of the scale under development was frequently omitted.
- Few of the studies described target population involvement in the measure development.
- For hypothesis or construct validity, few hypotheses were clearly described as formulated a priori and the choice of statistical analysis varied. For the latter, there was often a reliance on p-values alone rather than using correlations.
- There was often a lack of clarity concerning the exact processes used, especially for hypothesis or construct validity and reliability.
- There was limited consideration of the meaningfulness of the scores, including assessment of how useful scores would be in practice and minimally important changes. This becomes more important should the scales be applied in intervention studies for decision aids where changes in scores are to be measured.
- Related to this, measurement error was routinely not considered. The degree of background variation in repeated measurements and how this relates to the smallest detectable change (SDC) or minimal important change (MIC) in the overall score is of relevance if the tool is used for assessing the effectiveness of an intervention.
- There was also limited consideration of the interpretability of scores, such as evaluating the overall spread of results, subgroup results, and floor-ceiling effects.

- Where measure development was performed alongside an intervention, the scale responsiveness to changes in the construct of interest was not evaluated.
- Further evaluation studies illustrated the varying nature of scale psychometric properties, as they are heavily dependent on the population involved in the testing process and the study context.

This chapter detailed the results of a systematic review, which sought to identify and evaluate the quality of existing measures of the patient's perspective of the decision-making process.

Table 32: Summary of the methodological quality of the included scales

		Measurement Properties							
Instrument name	First author/year	Internal consistency	Reliability	Measurement error	Content validity	Hypothesis testing	Cross-cultural validity	Responsiveness	Interpretability
Decisional Conflict Scale	O'Connor, 1995	Poor	Fair	-	Poor	Fair	-	-	Poor
Satisfaction with Decision Scale	Holmes-Rovner, 1996	Poor	-	-	Poor	Fair	-	-	Poor
Decision Attitude Scale	Sainfort, 2000	Poor	-	-	Poor	Fair	-	Poor	Fair
PrepDM	Bennett, 2010	Poor	Poor	-	Poor	Fair	-	-	Fair
SDM-Q	Simon, 2006	Good	Fair	-	Excellent	Fair	-	-	Good
SURE	Legare, 2010	Good	-	-	Poor	Poor	-	-	Good
COMRADE	Edwards, 2003	Poor	-	-	Good	Good	-	-	Fair
Decision Evaluation Scales	Stalmeier, 2005	Excellent	-	-	Poor	Fair	Fair	-	Poor
Decision-Making Quality Scale	Hollen, 1994	Poor	Fair	-	Poor	Good	-	-	Poor
Decision Self-efficacy Scale	Bunn, 1996	Poor	-	-	Fair	Fair	-	-	Poor

Table 33: Summary of the measurement properties of the included scales

Where: + = positive, 0 = intermediate, - = negative, ? = no data

		Measurement Properties								
Instrument name	First author/year	Content validity	Internal consistency	Criterion validity	Construct validity	Reliability		Responsiveness	Floor & ceiling effect	Interpretability
						Agreement	Reproducibility			
Decisional Conflict Scale	O'Connor, 1995	0	0	?	+	?	0	?	?	0
Satisfaction with Decision Scale	Holmes-Rovner, 1996	0	0	?	+	?	?	?	?	0
Decision Attitude Scale	Sainfort, 2000	0	+	?	0	?	?	0	?	0
PrepDM	Bennett, 2010	0	-	?	0	?	0	?	0	0
SDM-Q	Simon, 2006	+	0	?	-	?	+	?	-	-
SURE	Legare, 2010	-	-	0	0	?	?	?	-	-
COMRADE	Edwards, 2003	+	0	?	+	?	?	?	0	0
Decision Evaluation Scales	Stalmeier, 2005	-	+	?	0	?	?	?	?	0
Decision-Making Quality Scale	Hollen, 1994	-	0	?	+	?	-	?	?	?
Decision Self-efficacy Scale	Bunn, 1996	+	0	?	0	?	?	?	?	?

6. Results of the instrument content mapping

This chapter describes findings from the instrument content mapping, which evaluates the extent existing measures address stages of the decision process model incorporating deliberation, as proposed by Elwyn and Miron-Shatz. (Elwyn and Miron-Shatz, 2010) Items from each individual scale are mapped against the model and the findings synthesised in a summary table, allowing comparison of individual scale performance and evaluation of collective coverage of the model's stages.

6.1. Mapping results for individual scales

6.1.1. The Decisional Conflict Scale (DCS)

As shown in Table 34, individual items from the DCS were mapped against the decision process map incorporating deliberation and determination, as described by Elwyn and Miron-Shatz. (Elwyn and Miron-Shatz, 2010) No items mapped against the imagining counterfactuals and affective forecasting stages of preference construction, though the other stages were addressed by the DCS.

Table 34: DCS items mapped onto decision process map

Decision process stage	Item mapping
Option awareness	
Options exist and these options need to be understood and considered	"I am aware of the choices I have"
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	"I feel I know the benefits of ***" "I feel I know the risks and side effects of ***"
Appraisal of knowledge sufficiency	"I need more advice and information about the choices" "I'm unsure what to do in this decision"
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	" I know how important the risks and side effects are to me" "I know how important the benefits are to me in this decision" "It's hard to decide if the benefits are more important to me than the risks, or the risks are more important than the benefits"
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	" I expect to stick to my decision" "I feel I have made an informed choice" " My decision shows what is most important to me" "I am satisfied with my decision" "It is clear which choice is best for me" "This decision is hard for me to take"
Items that did not map against process stages	
"I feel pressure from others in making this decision" "I have the right amount of support from others in making this choice"	

6.1.2. The Satisfaction with Decision Scale (SWD)

When the constituent items of the SWD scale are mapped against the decision process map, none match the option awareness, information search or preference construction stages, as shown in Table 35. (Elwyn and Miron-Shatz, 2010)

Table 35: SWD items mapped against decision process map

Decision process stage	**Item mapping**
Option awareness	
Options exist and these options need to be understood and considered	
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	
Appraisal of knowledge sufficiency	"I am satisfied that I am adequately informed about the issues important to my decision"
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	"I am satisfied with my decision" "I expect to successfully carry out the decision I made" "I am satisfied that my decision was consistent with my personal values" "The decision I made was the best possible decision for me personally"
Items that did not map against process stages	
"I am satisfied that this was my decision to make"	

6.1.3. Decision Attitude Scale (DAS)

As shown in Table 36, the Decision Attitude Scale items map against appraisal of knowledge sufficiency and determination, but not the option awareness, information search and preference construction stages of the proposed model. (Elwyn and Miron-Shatz, 2010)

Table 36: Decision Attitude Scale items mapped onto decision process map

Decision process stage	**Item mapping**
Option awareness	
Options exist and these options need to be understood and considered	
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	
Appraisal of knowledge sufficiency	"More information would help" "Consulting someone else would have been useful"
• Preference construction	
Imagining counterfactuals *"What if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	"My decision is sound" "I am comfortable with my decision" "My decision is the right one for the situation" "I am satisfied with my decision" "It was difficult to make a choice"
Items that did not map against process stages	
"I had no problem using the information" "The information was easy to understand"	

6.1.4. Preparation for Decision-Making scale (PrepDM)

Mapped against the decision process map incorporating deliberation and determination, the only stages not addressed are the imagining counterfactuals and affective forecasting within preference construction, as shown in Table 37. (Elwyn and Miron-Shatz, 2010)

Table 37: PrepDM items mapped onto decision process map

Decision process stage	**Item mapping**
Option awareness	
Options exist and these options need to be understood and considered	"Help you recognise a decision needs to be made"
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	"Help you think about the pros and cons of each option"
Appraisal of knowledge sufficiency	"Identify questions you want to ask your doctor"
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	" Help you organise your thoughts re the decision" "Help you think about which pros and cons are most important" "Know that the decision depends on what matters most to you"
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	"Prepare you to talk to your doctor regarding what matters most to you" "Prepare you to make a better decision"
Items that did not map against process stages	
"Help you think regarding how involved you want to be in this decision" "Prepare you for follow up meeting with your doctor"	

6.1.5. The Shared Decision-Making Questionnaire (SDM-Q)

As shown in Table 38, individual items from the SDM-Q were mapped against the decision process map incorporating deliberation and determination, as described by Elwyn and Miron-Shatz. (Elwyn and Miron-Shatz, 2010) No items mapped against the appraisal of knowledge sufficiency or preference construction stages. Eight items did not map against decision process, highlighting the additional purpose of this measure in evaluating shared decision-making within a consultation and eliciting patient preference for involvement. In addition, items for equipoise and eliciting preferences were omitted from the scale following the IRT analysis due to poor item-fit.

Table 38: SDM-Q items mapped onto decision process map

Decision process stage	**Item mapping**
Option awareness	
Options exist and these options need to be understood and considered	
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	"I now know the advantages of the individual treatment options"
Appraisal of knowledge sufficiency	
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	"my doctor and I weighed up different treatment options thoroughly"
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	" I now know which treatment option is best for me"
Items that did not map against process stages	
"I was able to discuss the different treatment options with my doctor in detail" "In the selection of treatment method, my thoughts were taken into account just as much as the considerations of the doctor" "There was enough time to ask questions" "My doctor and I selected a treatment option together" "During the consultation, I felt included in the treatment decision" "Through the consultation with the doctor I felt jointly responsible for my further treatment" "My doctor and I discussed the next few steps of the treatment plan in detail" "My doctor and I reached an agreement as to how we will proceed"	

6.1.6. The SURE scale

In mapping the SURE items against the decision process map, option awareness and the imagining counterfactual and affective forecasting stages of preference construction were not covered by this scale. (Elwyn and Miron-Shatz, 2010) These findings are shown in Table 39 below.

Table 39: SURE items mapped onto decision process map

Decision process stage	**Item mapping**
Option awareness	
Options exist and these options need to be understood and considered	
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	"Do you know the benefits and risks of each option?
Appraisal of knowledge sufficiency	"Do you have enough support and advice to make a choice?"
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	"Are you clear about which benefits and risks matter the most to you?"
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	"Do you feel sure about the best choice for you?"
Items that did not map against process stages	

6.1.7. The COMRADE scale

As shown in Table 40 individual items from the COMRADE were mapped against the decision process map incorporating deliberation and determination. (Elwyn and Miron-Shatz, 2010) No items mapped against the imagining counterfactuals and affective forecasting stages of preference construction, although the other stages were addressed. Eight items did not map against the decision process stages. These items concerned evaluating the extent of shared decision-making in the consultation.

Table 40: COMRADE items mapped onto decision process map

Decision process stage	Item mapping
Option awareness	
Options exist and these options need to be understood and considered	"The doctor made me aware of different treatments available" "I am aware of the treatment options I have"
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	"I know the advantages of treatment or not having treatment" "I know the disadvantages of treatment or not having treatment"
Appraisal of knowledge sufficiency	"The doctor gave me enough information regarding treatment choices available" "The doctor gave me enough explanation of the information about the treatment choices" "I am satisfied that I am adequately informed about issues important to the decision" "Overall I am satisfied with the information I was given"
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	"It is clear which choice is best for me"
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	"I am sure the decision made was the right one for me personally" "The decision shows what is important to me" "I feel an informed choice had been made" "I am satisfied with the way the decision was made in the consultation"
Items that did not map against process stages	
"The doctor gave me the chance to express my opinions regarding different treatments available" "The doctor gave me the chance to ask for as much information as I needed for the different treatment options" "The information was easy to understand" "The doctor gave me a chance to decide which treatment I thought was best for me" "The doctor gave me a chance to be involved in the decisions during the consultation" "The doctor and I agreed about which treatment (or no treatment) was best for me" "I can easily discuss my condition again with my doctor"	

6.1.8. The Decision Evaluation Scale (DES)

As shown in Table 41, no items mapped against the preference construction stages on the decision process map. (Elwyn and Miron-Shatz, 2010) Four items did not map against the decision process stages. These items mostly addressed the patient's ownership of the decision and desire to participate.

Table 41: DES items mapped onto decision process map

Decision process stage	**Item mapping**
Option awareness	
Options exist and these options need to be understood and considered	
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	"I know the pros and cons of the treatments"
Appraisal of knowledge sufficiency	"I am satisfied with the information I received" "I want more information about this decision" "I want clearer advice"
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	"I expect to stick to my decision" "I am satisfied with my decision" "I am still doubtful about my choice" "I find it hard to make this choice" "I made a well informed choice" "My decision frightens me" "I regret my decision"
Items that did not map against process stages	
"This is my own decision" "This decision is made without me" "I feel pressure from others in making this decision" "I wish someone else would make this decision for me"	

6.1.9. The Decision-Making Quality Scale (DMQS)

As shown in Table 42, the DMQS items map against all stages with the exception of determination. (Elwyn and Miron-Shatz, 2010) Due to the wording, one item maps against two stages - information search and appraisal of knowledge sufficiency - as it describes seeking additional information if needed.

Table 42: DMQS items mapped onto decision process map

Decision process stage	**Item mapping**
Option awareness	
Options exist and these options need to be understood and considered	Searches for three or more choices
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	*Finds out more information about the pros and cons* when needed
Appraisal of knowledge sufficiency	Finds out more information about the pros and cons *when needed*
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	Weighs pros and cons of consequences Makes detailed plans with back up plans
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	Takes into account values and all goals desired
Ranking options	Thinks about new information and what experts say, even if against first choice Reviews carefully before making a final choice
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	
Items that did not map against process stages	

6.1.10. The Decision Self-Efficacy Scale (DSES)

The DSES asks respondents to rate their confidence for each item. These match against each stage of the decision process, as outlined in Table 43, with the exception of imagining counterfactuals and affective forecasting. (Elwyn and Miron-Shatz, 2010)

Table 43: DSES items mapped onto decision process map

Decision process stage	Item mapping
Option awareness	
Options exist and these options need to be understood and considered	"get the facts about the medication choices available to me"
Deliberation	
• Knowledge	
Information search *Gain information about the attributes of options (characteristics, process and outcomes probabilities)*	"get the facts about the benefits of each choice" "get the facts about the risks and side effects of each choice"
Appraisal of knowledge sufficiency	"understand the information enough to be able to make a choice" "ask for advice" "delay my decision if I feel I need more time"
• Preference construction	
Imagining counterfactuals *"what if" scenarios: provide insight into possible consequences of options by imagining how different futures could play out*	
Affective forecasting *Forecasted feeling towards different counterfactual futures.*	
Ranking options	"figure out the choice that best suits me" "express my concerns about each choice"
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	"let the clinic team know what is best for me"
Items that did not map against process stages	
"ask questions without feeling dumb"	

6.2. Narrative synthesis of content mapping

The extent each scale maps against the decision process map developed by Elwyn and Miron-Shatz, is summarised in Table 44. (Elwyn and Miron-Shatz, 2010) None of the scales map against all of the proposed stages. Appraisal of knowledge sufficiency and determination are the most frequently addressed stages. Only one scale, the Decision Making Quality Scale (DMQS), addresses the imagining counterfactuals and affective forecasting elements of the preference construction stage. This scale also maps against most stages, although one of its items maps against both information search and appraisal of knowledge sufficiency due to the wording of the item.

The words and phrases forming each item were then mapped against stages of the decision-making process, as shown in Table 45. (Elwyn and Miron-Shatz, 2010) This shows that the exact phrasing of items and the timing of their completion in relation to decision-making are important in the stage allocation.

Table 44: Summary of the instrument content, mapped against the decision process model

		Decision process steps						
Instrument name	First author/year	Option awareness	Deliberation					Determination
			Knowledge		Preference construction			
			Information search	Appraisal of knowledge sufficiency	Imagining counterfactuals	Affective forecasting	Ranking options	
Decisional Conflict Scale	O'Connor, 1995	✔	✔	✔	✖	✖	✔	✔
Satisfaction with Decision Scale	Holmes-Rovner, 1996	✖	✖	✔	✖	✖	✖	✔
Decision Attitude Scale	Sainfort, 2000	✖	✖	✔	✖	✖	✖	✔
PrepDM	Bennett, 2010	✔	✔	✔	✖	✖	✔	✔
SDM-Q	Simon, 2006	✖	✔	✖	✖	✖	✔	✔
SURE	Legare, 2010	✖	✔	✔	✖	✖	✔	✔
COMRADE	Edwards, 2003	✔	✔	✔	✖	✖	✔	✔
Decision Evaluation Scales	Stalmeier, 2005	✖	✔	✔	✖	✖	✖	✔
Decision-Making Quality Scale	Hollen, 1994	✔	✔	✔	✔	✔	✔	✖
Decision Self-efficacy Scale	Bunn, 1996	✔	✔	✔	✖	✖	✔	✔

Table 45: Mapping of words from scale items against stages of decision-making process

Decision process stage	Terms encountered during mapping
Option awareness	
Options exist and these options need to be understood and considered	Awareness, options, choices, recognise, different, search, more, available,
Deliberation	
• Knowledge	
Information search	Benefits, risks, pro, con, side effects, advantages, disadvantages, think, option, facts, know.
Appraisal of knowledge sufficiency	More, advice, satisfied informed, need, enough, unsure, other person, questions want to ask, support, adequate, clearer, understand, delay.
• Preference construction	
Imagining counterfactuals	Consequences, detailed plans and back-up plans
Affective forecasting	Takes into account all goals desired
Ranking options	Best for me, important to me, matters most to me, hard to decide, organise thought, weighed thoroughly, clear, considers new information against initial choice, review with care, each choice, ”I know… “, concern.
Determination	
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	Decision, timing/tense, satisfied, sure, hard to do, intention to act, informed choice made, prepared to talk, best possible, sound, comfortable, right for context, difficult, better, know now, right, doubt, frightened, regret, choice, clear.

This chapter reported the results of a content mapping process, which determined the extent existing measures addressed the stages of a decision process model incorporating deliberation.

7. Discussion

In this chapter, the systematic review and content mapping findings are summarised and placed in the context of knowledge to date, with the implications for developing a new measure of the decision-making process considered. The study design is outlined and its strengths and limitations discussed.

7.1. Summary

This systematic review answers the research question: to what extent do existing measures of decision-making consider the processes of decision making such as deliberation?

The aim was to systematically identify and analyse existing measures of shared-decision making that consider the process from the patient's perspective, determining the extent the items map onto a decision process map incorporating deliberation. As part of the analysis, the rigour of measure development process was considered along with the measurement properties of the instruments themselves. (Mokkink et al., 2012, Terwee et al., 2007)

The results indicate that current measures of decision-making do not consider all steps of the decision process. There is also marked variation in the methodological approach for instrument development and the quality of existing instruments' measurement properties. The Decision Making Quality Scale (DMQS), developed by Hollen and based on Janis and Mann's work on quality decision-making, addresses aspects of deliberation not covered by other measures. The recommendations for a new instrument will therefore consider the work of Hollen, Janis and Mann. (Hollen, 1994, Janis and Mann, 1977)

7.2. Discussion of findings

7.2.1. Key findings

Ten eligible measures were identified, along with nine further evaluation studies for these scales.

The methodological quality was highly variable, both between studies and when considering the approach to different measurement properties within studies. None of the scales were found to score consistently well across all measurement properties when compared to the quality criteria. (Terwee et al., 2007)

When mapped against the model of deliberation and determination proposed by Elwyn and Miron-Shatz, no scale addressed all stages. (Elwyn and Miron-Shatz, 2010) The most frequently addressed stages were appraisal of knowledge sufficiency and determination. The Decision Making Quality Scale (DMQS) developed by Hollen was the only scale to address the imagining counterfactuals and affective forecasting elements of the preference construction stage. It also mapped against most stages of the decision process, with determination being the only one missing. (Hollen, 1994)

7.2.2. Findings in the context of knowledge to date

7.2.2.1. The methodological approach in existing studies concerning measures of shared decision-making

The development and measurement properties of a scale have been shown to influence study outcomes. (Marshall et al., 2000, O'Connor et al., 1996) As such, the use of a poorly developed scale carries consequences for research findings and clinical practice. Scholl and colleagues have previously

highlighted the variable methodological quality in the development of shared decision-making scales and also identified the COSMIN scale as possible guidance for future studies. (Scholl et al., 2011) This current work is the first application of such reproducible standards to a systematic review in this field. It confirms shortcomings in the methodological approaches taken to date for measures considering the patient perspective. That a measure reported as recently as 2014 was excluded due to the absence of any psychometric evaluation suggests this is an ongoing issue. (Kaltoft et al., 2014) However, with the development of COSMIN and a growing focus on measurement instruments and shared decision-making, future studies can access guidance and checklists from the early stages of scale development, leading to greater rigor and consistency in both methodology and reporting of measure development. (Johnston and Graves, 2008)

7.2.2.2. The measurement properties of existing instruments

Shared decision-making is a rapidly evolving field, which is likely to further increase with growing evidence-based medicine, healthcare options, patient autonomy and access to information. (Elwyn et al., 2012, Légaré et al., 2010b, Collings and Coulter, 2015) The impact of decision aids has been demonstrated in diverse fields and yet varying measurement property performances were found when a criteria of adequacy was applied to existing measures of shared decision-making. (Stacey et al., 2014) This needs to be taken into consideration when using these existing measures in studies of shared decision-making and interpreting findings based on their use in previous research.

Measurement properties are not fixed characteristics of instruments and are instead influenced by the context in which they are used. As such, scales

should be retested when used in new settings or groups. (Streiner and Norman, 2008) This issue was emphasised by the shifting performance of instruments in further evaluation studies. For example, key features such as the number of subscales identified in a measure altered depending on the study context. (O'Connor, 1995, Koedoot et al., 2001)

Scales measuring shared decision-making are increasingly used in intervention studies, such as those evaluating the impact of decision aids. The importance of reliability and responsiveness increases in contexts where scores obtained with an instrument are expected to change. (Mokkink et al., 2010b) However, these measurement properties are currently among the least often examined. To accurately measure the impact of decision support on the decision-making process in future, information is needed on these aspects of scale performance.

It is understandable that in rapidly growing fields such as measure development and shared decision-making, initial studies were pragmatic, clinically-based investigations but a more robust approach is required for future findings to be meaningful. Johnson and Graves note that scales are often chosen due to tradition or popularity rather than consideration of "the quality of a measure for a particular construct or application." (Johnston and Graves, 2008) As such, researchers and clinicians should call for relevant, valid and reliable tools, encouraging the development of scales that meet quality standards.

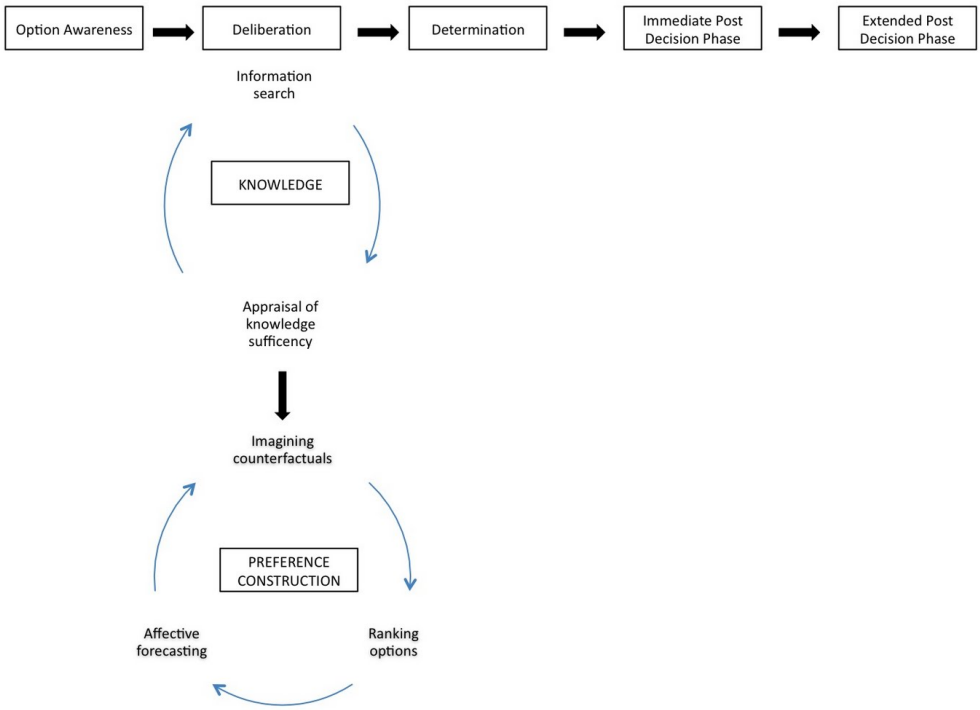
7.2.2.3. What is already known about shared decision-making and deliberation?

Scholl and colleagues arranged existing measures of shared decision-making according to whether they addressed antecedents, processes or outcomes of

decision-making. (Scholl et al., 2011) The included measures considered the decision process from the perspective of patients, clinicians or observers. (Scholl et al., 2011) Of those they identified as measuring decision processes, such as the OPTION scale or the 9-item Shared Decision-Making Questionnaire, none were found to focus on the process of deliberation from the patient perspective in this review. The scales instead focused on appraising the shared decision-making process either from a patient, clinician or observer viewpoint. (Edwards et al., 2003, Simon et al., 2006) Yet, as Edwards and Elwyn suggest, without understanding the decision-making process, interventions are influencing outcomes without insight into the mechanism of action. (Edwards and Elwyn, 2006) In further work by these authors, decision processes such as deliberation are highlighted as a point where support may be needed to facilitate shared decision-making, as the patient may otherwise feel abandoned, wary of participating, surprised, unsettled, and uncertain about their decision and capacity. (Elwyn et al., 2012)

Elwyn and Miron-Shatz highlighted the problem of using knowledge, preferences and outcomes to measure the decision-making process, suggesting instead the analysis of deliberation as an indicator of decision-making quality. (Elwyn and Miron-Shatz, 2010) The decision process model incorporating deliberation outlined by these authors has the following facets: option awareness, information search, knowledge gain, appraisal of knowledge sufficiency, imagining counterfactuals, affective forecasting and preference construction. This is illustrated in Figure 4, revisiting the diagram shown earlier in Chapter 1. (Elwyn and Miron-Shatz, 2010)

Figure 4: Model of decision-making process incorporating deliberation (revisited)



Only one instrument was found to consider the facets of deliberation as described by Elwyn and Miron-Shatz. The Decision Making Quality Scale (DMQS) developed by Hollen considers each aspect of deliberation detailed above. (Hollen, 1994) This scale is based on Janis and Mann's conflict model of decision-making, which considers decision-making in high stress contexts with potential losses regardless of the choice made. (Hollen, 1994) This model, adapted by Hollen, has six non-linear stages: situation appraisal, option assessment, weighing choices, deliberation, evaluation and adherence to choice in the face of challenge. (Hollen, 1994) Janis and Mann also describe styles of decision-making, both positive and negative, based on behaviour at each of these stages. (Janis and Mann, 1977) A feedback loop occurs between each stage of the decision process, with the decision maker appraising options available as well as their perceived hope and resources for finding a better

alternative. (Janis and Mann, 1977) Similar to the "bounded rationality" of decision heuristics, restrictions such as time and resources are considered both as contributing to the decision made and also influencing the decision style. (Janis and Mann, 1977, Elwyn et al., 2001a) This work also informed the development of O'Connor's Decisional Conflict Scale, although this focuses on conflict as a marker of the decision-process, rather Janis and Mann's criteria for quality decision-making, which are outlined in Table 46. (O'Connor, 1995)

Table 46: Janis and Mann's criteria for quality decision-making

Janis and Mann's criteria for quality decision-making
Thorough canvassing of alternatives
Thorough canvassing of objectives
Careful evaluation of consequences
Thorough search for information
Unbiased assimilation of new information
Careful re-evaluation of consequences
Thorough planning for implementation and contingencies

In focusing on deliberation due to concerns over the potential bias of post-hoc evaluation of decision quality, Hollen describes similar reasoning as Elwyn and Miron-Shatz. (Hollen, 1994, Elwyn and Miron-Shatz, 2010) That is, once a decision is made, it is difficult to evaluate the decision process without the influence of outcomes and opportunities lost. As such, the DMQS is the only included scale that did not measure determination. It is not driving the individual to a decision, instead reflecting on the decisional process and explicitly exploring desired goals and the potential for things not going to plan. (Hollen, 1994) In addition, the scale is generic rather than tailored to a specific decision context. Streiner and Norman highlight the strength of such scales due to their generalisability and breadth of perspective, concluding that they yield comparable results to disease or context specific scales. (Streiner and Norman, 2008)

However, there is some conflict between the work of Hollen, based on Janis and Mann, and that proposed by Elwyn and Miron-Shatz. (Elwyn and Miron-Shatz, 2010, Hollen, 1994, Janis and Mann, 1977) The former explicitly emphasises the non-linear nature of decision processes and the recognition of external factors such as time pressure and stress, in contrast to the decision process model of Elwyn and Miron-Shatz. (Elwyn and Miron-Shatz, 2010, Hollen, 1994) There is also the influence of those around the decision maker, as described in Bandura's Social Learning theory and the interlinked autonomy referenced by Elwyn and colleagues. (Bandura, 1977, Elwyn et al., 2012) These, along with past experiences, directly impact on an individual's ability to gather and appraise information – influences that should be considered in the decision process.

Elwyn and colleagues have noted these limitations in subsequent work. In 2012, the clinical model of shared decision-making incorporating deliberation acknowledged the "psychological, emotional and social factors" involved. It also noted the contribution of others in the decision-making process. (Elwyn et al., 2012) In 2014, an expert-led discussion addressed the collective nature of decisions, developing a collaborative deliberation model that considered the role of people in addition to the individual at the centre of the process. (Elwyn et al., 2014) The decision process is seen as a social act, drawing on the support of others and influenced by socially constructed norms, rather than a choice made by an individual in isolation. This work also acknowledges the emotional context and resource constraints impacting on the decisional process. (Elwyn et al., 2014)

The work of Elwyn and Miron-Shatz, along with the subsequent collaborations, are expert-led recommendations and as such can be argued to be opinion-based

and subjective. (Elwyn and Miron-Shatz, 2010, Elwyn et al., 2014, Elwyn et al., 2012) Yet Janis and Mann made similar observations in 1970s and Hollen in the 1990s, suggesting a consistent theme that warrants further investigation. Despite this, Hollen's work is among the least often cited of the included scales, with 24 citations on Web of Science in comparison with the 565 of O'Connor's Decisional Conflict Scale. (Hollen, 1994, O'Connor, 1995) While this may reflect the challenge of measuring what seems hard to define – Edwards and Elwyn's proposed "black box" of the decision-making process - it is important not to neglect further. (Edwards and Elwyn, 2006) With the exception of the DMQS, current instruments are neglecting the personal, challenging aspects of decision making such as affective forecasting and imagining counterfactuals, the omission of which in health-related decisions could have a profound effect on the choices made.

In addition, the varied approaches to defining and measuring decision-making suggest the need to move from a reductionist approach. With varying timescales, contexts, stresses and decision-making styles, it is unlikely that one measure will suit all needs. The recent collaborative deliberation model notes this, emphasising that preference construction can be achieved through a range of processes, both analytical and otherwise. (Elwyn et al., 2014) Perhaps what is needed is a decision evaluation toolkit, as also suggested by Scholl and colleagues, with measures for different circumstances just as there are different methodologies for different research, different investigations for different symptoms. (Scholl et al., 2011) Whatever is done, it must be robustly developed and evaluated to ensure validity and reliability in view of the impact of misleading measures in healthcare.

7.3. Implications for developing a measure of patient deliberation

7.3.1. Specifications for a new instrument

Streiner and Norman identify devising items as the first step in developing a measure, but also recommend avoiding duplication of efforts and unnecessary resource use by appraising the suitability of pre-existing scales and research, and building on these where possible. (Streiner and Norman, 2008) The DMQS was the only scale that covered all steps of deliberation in the decision-making process proposed by Elwyn and Miron-Shatz. However, the methodological design and measurement properties for the scale performed poorly, limiting its further application. (Streiner and Norman, 2008, Johnston and Graves, 2008)

The DMQS drew on the work of Janis and Mann, and therefore the new scale should be structured to cover each domain of Elwyn and Miron-Shatz's decision process model and also Janis and Mann's criteria for quality decision-making. (Elwyn and Miron-Shatz, 2010, Hollen, 1994, Janis and Mann, 1977) In addition, the new scale could draw on items identified in pre-existing scales. However, it is also important to develop the affective forecasting and imagining counterfactuals items, as these were identified least often in existing measures. These foundation elements for a new scale are summarised in the first three columns of Table 47 below.

Two versions of the scale could be developed. A quantitative measure would allow evaluation of scores and comparisons of repeated measurements, such as before and after an intervention. A more qualitative, open question style based on Janis and Mann's decision quality criteria, as outlined in the fourth column of Table 47, would allow flexibility and individual reflection, which could then be used to guide more in-depth decision support counselling.

A further aspect to be considered is the timing of scale administration. Based on the findings of this work, the scale should be designed for completion before a decision is made, especially for consequential decisions, as post-hoc evaluation of the decisional process outcomes is open to the influence of outcomes and bias. (Hollen, 1994, Elwyn and Miron-Shatz, 2010)

In practice, a measure considering the decisional process from the patient's perspective would allow tailored, targeted support to areas of the decision-making process. It would facilitate a more nuanced understanding of what makes a good decision for an individual, moving away from the rigidity of externally applied criteria such as those for knowledge, preference and satisfaction alone.

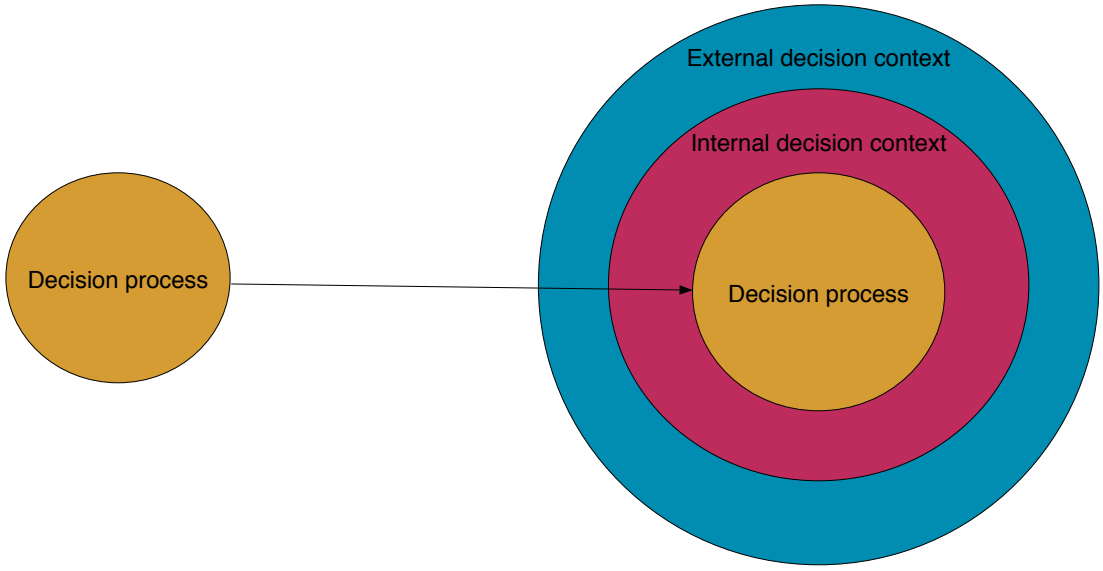
Table 47: Foundations for a new scale

Elwyn and Miron-Shatz's decision process model stages	**Pre-existing terms encountered during item mapping**	**Janis and Mann's criteria for quality decision-making**	**More discursive items to gather reflections**
Option awareness			
Options exist and these options need to be understood and considered	Awareness, options, choices, recognise, different, search, more, available.		
Deliberation			
• Knowledge			
Information search	Benefits, risks, pro, con, side effects, advantages, disadvantages, think, option, facts, know.	Thorough search for information Thorough canvassing of alternatives Unbiased assimilation of new information	What am I mostly basing my decision on? Other than the information given, what extra things are influencing my choice?
Appraisal of knowledge sufficiency	More, advice, satisfied informed, need, enough, unsure, other person, questions want to ask, support, adequate, clearer, understand, delay.		
• Preference construction			
Imagining counterfactuals	Consequences, detailed plans and back-up plans	Careful evaluation of consequences Thorough canvassing of objectives Careful re-evaluation of consequences Thorough planning for implementation and contingencies	What problems/goals can I see coming with each choice? How would I handle each of them? Has looking into the choices more changed what I think might happen? If so, how? Has it changed how I feel?
Affective forecasting	Takes into account all goals desired		
Ranking options	Best for me, important to me, matters most to me, hard to decide, organise thought, weighed thoroughly, clear, considers new information against initial choice, review with care, each choice, "I know… ", concern.		
Determination			
Integrating deliberation input and making a choice, prior to enacting a decision. *Intention, certainty and evaluation of decision made.*	Decision, timing/tense, satisfied, sure, hard to do, intention to act, informed choice made, prepared to talk, best possible, sound, comfortable, right for context, difficult, better, know now, right, doubt, frightened, regret, choice, clear.		

7.3.2. Assessing the decision context

In addition to a scale measuring patient deliberation, limitations identified with the Elwyn and Miron-Shatz model could be further addressed with the development of complementary tools to assess the internal and external context in which the decision takes place, as illustrated in Figure 5 below. This would draw on Janis and Mann's decisional conflict model, along with collaborative decision-making and shared decision-making for clinical practice models. (Janis and Mann, 1977, Elwyn et al., 2014, Elwyn et al., 2012)

Figure 5: Assessing the decision context



For the internal or individual decision context, potential facets to consider include the patient's narrative of the situation; their emotional state, how they perceive their role in the decision and previous decision-making experiences and style.

External or collective decision contexts to consider include the alternatives and resources available, the experiences and perspectives of other people involved in the situation, and any identifiable wider influences or patterns.

Evaluation of internal and external decision contexts would allow consideration of all interconnected elements that might influence the decision, placing it both within the unique context of the individual and the wider setting. (Williams and Hummelbrunner, 2010) This offers learning relevant not just to the decision in question, but also for future decisions. Both the patient and others involved, such as the healthcare team, could benefit, with the external and collective factors likely to be of relevance to other patients and wider healthcare practice. For example, continued issues with staffing, treatment options or other resources may lead to a change in the provision of care or commissioning.

These areas may be more suitable to collaborative exploration, for example with the patient, their support network and the healthcare team. In addition, further work is needed to explore how such elements could be assessed efficiently in often time-pressured healthcare settings.

7.3.3. Developing a new measure

Methods for developing new scales include involving the population of interest, seeking expert opinion and basing items on research, clinical observation or theory. (Streiner and Norman, 2008) The work of Janis and Mann, which underpinned the DMQS, along with more recent decision-process models proposed by experts in shared decision-making, together provide a strong foundation for the development of a new measure and should be utilised in the development of a new scale. (Elwyn and Miron-Shatz, 2010, Hollen, 1994, Janis and Mann, 1977, Elwyn et al., 2012, Elwyn et al., 2014) However, the voices of the patients themselves are currently marked in their absence. As such, the next steps in scale development should involve the people for whom the scale is intended.

A key element to address is to further develop understanding of the decision-making process from the patient perspective. Recent work by Barr and Elwyn has also supported the importance of target group involvement, highlighting the potential impact on scale interpretation and validity, and therefore on research findings. The authors suggest cognitive interviews as an alternative method, where item interpretation and participant responses are explored in depth. (Barr and Elwyn, 2015) Other possible methods include using focus groups, targeted interviews or observational studies to further explore and test the relevance existing theories and models. (Elwyn and Miron-Shatz, 2010, Streiner and Norman, 2008)

While these methods move closer to incorporating patient's views, they are more limited in their exploration of the patient's experience of the decision-making process and also of collaborating with healthcare professionals. In addition, they remain researcher-led, with researchers ultimately determining interview topics and provisional scale items. As such, another option is to use a more participative approach, with the study design developed and delivered by patients working with research and clinical teams. (Olshansky et al., 2005) This collaborative exploration of the decision-making process would be more in keeping with the underlying principles of shared decision-making, while the process itself would also provide insight into the mechanics of collaboration and co-production in healthcare.

Another consideration is the sampling strategy and recruitment process used in future studies. The recent study by Barr and Elwyn highlighted the need for a representative or generalisable sample to be included in instrument development, as the participants involved were mostly university educated,

with no other demographic details described except age and gender. (Barr and Elwyn, 2015) Similar patterns were encountered in the development studies of instruments included in this review. There is a risk of contributing to health inequalities by omitting members of the population from such research and failing to gather insights from groups that may already find accessing shared decision-making challenging (Elwyn et al., 2014), and further research should specifically address this in the study design.

The findings from a participative exploration of the decision-making process should then be used to review the recommendations made in this study, including the specifications for a new measure of patient deliberation and the proposed complementary tools for internal and external decision contexts.

Once the foundations of the scale have been further established and the proposed scale revisited, robust measure development and evaluation is needed. This should include development work with patient groups, along with consideration of the accessibility of the tools produced. The COSMIN checklist should be used to guide which measurement properties to use and how to best evaluate them. (Mokkink et al., 2012) There must also be clear performance criteria for the scale's measurement properties in keeping with guidance, such as those developed by Terwee and colleagues, and used in this review. (Terwee et al., 2007)

The process of developing a measure of deliberation from the patient's perspective will not be without challenges. Scale development is a resource intensive process. (Streiner and Norman, 2008) The nuance of terms and phrases as used in the scales identified in this review re-iterates the importance of target group involvement. Elwyn and colleagues confirm this in recent work

where the terms options, decisions and preferences were identified as unfamiliar and barrier forming by people attending a medical centre. (Elwyn et al., 2013) The steps of forecasting the future and imagining counterfactuals are potentially emotionally laden and psychologically challenging in some medical contexts, particularly where choices may alter long-term goals and plans. This would be more complicated in patient journeys that involve a change of perceived identity or life narrative by illness. (Hydén, 1997) In addition, the scale developed needs to be practical to use in time-pressured clinical settings. Such challenges underline the need for such measures to be robustly developed and evaluated.

7.3.4. Next steps

In summary, the recommended next steps for developing a scale to measure the patient's perspective of decision-making processes are:

1. Improve understanding of decision-making processes and collaboration in healthcare from the patient's perspective using a participative research approach.
2. Use the findings to revisit the proposed scale outlined in this study, along with the suggested use of recent models of shared decision-making and the work of Janis and Mann as foundations for the scale.
3. Amend the proposed scale and develop complementary tools for the internal and external decision contexts, if still indicated.
4. Design a robust and transparent instrument development study using the COSMIN standards, and evaluate the new scale's psychometric properties using criteria such as those developed by Terwee and colleagues.
5. Ensure an inclusive sampling and recruitment strategy at each step, along with evaluation of the accessibility and functionality of the scale produced.

7.4. Methodology of the review

7.4.1. Overview of method

The systematic review search strategy was developed in accordance with guidelines, with additional advice from an information specialist and medical librarian. (Critical Appraisal Skills Programme (CASP), 2011, Higgins and Green, 2011, Mann, 2011, Moher et al., 2009)Seven electronic databases (Medline, EMBASE, Cochrane Library, PsycInfo, Assia, CINAHL and Medline in Process) were searched with no restrictions placed on language or date. Additional searches included manual searches of the most frequently cited journals, the references of included reports and additional database searches using the name of each included scale. The searches were repeated in July 2014.

Studies were included if they reported on the development and psychometric evaluation of an instrument designed to measure shared decision-making in the health-care setting, specifically the patient's perspective of the decision making process. A second reviewer reviewed ten per cent of the references independently.

A total of 7,927 references were identified from the electronic databases. 1,380 duplicates were removed. Following title and abstract review, 168 progressed to full text review. Of these, ten studies involving the development of measures were included and a further nine addressed the further evaluation of instruments. No further instruments or evaluation studies were identified through the additional searches.

Data extraction fields were: i) the descriptive features of the study and scale ii) assessment of study methodological quality using the COnsensus-based

Standards for the selection of health Measurement INstruments (COSMIN) checklist as reproducible standards iii) assessment of the scale measurement properties using predefined criteria of adequacy and iv) content evaluation of the scale by mapping the items against decisional process map incorporating deliberation. (Elwyn and Miron-Shatz, 2010, Mokkink et al., 2012, Terwee et al., 2007)The results were summarised using narrative synthesis due to the heterogeneity of the included studies. (Higgins and Green, 2011) However, a COSMIN scoring system and summaries of the measurement properties and content mapping for included scales were used to improve the clarity of the results. (Terwee et al., 2012)

7.4.2. Strengths

Systematic reviews are rigorous, transparent summaries of research evidence where information is gathered and analysed according to a pre-agreed, reproducible process that reduces sources of bias where possible. (Hemingway and Brereton, 2009)

A robust systematic review was developed using recognised guidance from the Cochrane Collaboration and the Centre for Reviews and Dissemination. (Higgins and Green, 2011, Centre for Reviews and Dissemination, 2009, de Vet et al., 2008) However, these guidelines were mostly concerned with systematic reviews of interventions or diagnostic tests and addressed different study designs to those used in the development of measures of shared decision-making. Additional guidance was sought by searching the literature available on the development of measures and patient-reported outcomes. This led to the identification of the work by Mokkink and colleagues, which includes the development of the COSMIN checklists. (Mokkink et al., 2009)

The use of these resources contributed to a more robust, transparent systematic review. The protocol was presented to colleagues in the Decision Laboratory at the Institute of Primary Care and Public Health, Cardiff University and checked against the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement. (Liberati et al., 2009)

A thorough search strategy was developed, incorporating several diverse electronic databases and supplemented by manual searches, review of the references of included reports and additional measure-specific database searches. A sensitive strategy was used due to the limitations and challenges in searching for measures of patient reported decision-making processes, such as variation in the terms used by authors and indexing on databases. (Mokkink et al., 2009) While this produced a high number of references, it ensured greater detection of potentially eligible studies. Once the duration of the project became apparent, the searches were repeated in order to keep the findings relevant and up-to-date.

Two independent reviewers were involved in the study selection, with a third reviewer available for consultation in the event of differing opinions. Both first and second reviewers are from fields separate to that of shared decision-making and decision support, thus providing independent perspectives. The third reviewer has extensive experience in these fields and therefore provided expertise and guidance as needed without overly influencing the decisions made. (Lefebvre et al., 2009)

The study selection and data extraction processes were piloted, allowing identification of any issues and assessment of inter-reviewer consistency.

Electronic data extraction forms were used to facilitate record keeping and sharing. (Higgins and Green, 2011)

A previously identified issue in existing systematic reviews of measurement instruments is the limited use of reproducible standards for methodological quality and criteria for scale measurement properties. (Mokkink et al., 2009) These factors influence scale quality and also their suitability for use in research and clinical practice. As such, the COSMIN standards for study design and measurement property criteria produced by Terwee and colleagues were applied in this review. (Mokkink et al., 2012, Terwee et al., 2007)

7.4.3. Limitations

When developing the search strategy, a grey literature search was not included as the variations in indexing and terminology noted for published resources would be further compounded in unpublished work, placing it beyond the scope of this review's resources. However, this does open the findings to publication bias. (Hemingway and Brereton, 2009) In addition, omitting grey literature meant that online resources about the scales, such as those for the Decisional Conflict Scale (DCS), were not identified. (O'Connor, 2010) In addition, the database Scopus was not searched as it was new at the time of project development and there was uncertainty over which databases would be covered by its use. The searches were repeated in July 2014 due to the duration of the project, but an alternative approach would have been to create search alerts on databases. (Higgins and Green, 2011)

Resource limitations influenced what was possible during the project. For example, contacting authors for further information about scale development and leading experts to suggest measures in development was not possible with

the resources available. (Higgins and Green, 2011) However, this could have been anticipated due to known variation in terminology and reporting in the field of instrument development. (Mokkink et al., 2009) As such, future systematic reviews should allow sufficient time and resources for this step.

A second reviewer participated in the study selection in order to reduce the potential selection bias introduced with only one reviewer. However, the sensitive search strategy produced a high number of results and time limitations meant that both reviewers screened only ten per cent of the references. This is considered an acceptable compromise by the Cochrane Collaboration guidance and ten per cent of the references still equated to 500 titles and abstracts. (Higgins and Green, 2011)

During the initial title and abstract review, a key aspect of the eligibility criteria wording was demonstrated. The second inclusion criterion was that the population involved in the study be patients involved in healthcare decision-making for a decision regarding their own health or treatment. As such, the study and instrument had to measure the patient's perspective of a decision-making process. This emphasis excludes scales assessing role preference or consultation scoring and those testing patient knowledge of a specific condition or treatment. While the latter is an aspect of shared decision-making, answering specific questions regarding, for example, treatment side effect frequency does not reflect the patient's perspective of the decision-making process.

There were limitations in using the COSMIN checklist for methodological quality and the criteria for the measurement properties. At time of review, the checklist was only recently developed and still under evaluation. A Delphi process had been used during its production and aspects of the design are

debatable, such as exclusion of participants who were not published experts and the lack of justification for a 67% agreement threshold. (Mokkink et al., 2010b) Similarly, for the four-point scoring version of the checklist, most of the development was led by one author, although a steering group oversaw the process and further evaluation planned. (Terwee et al., 2012) In addition, to date the COSMIN checklist is more frequently used for traditional clinical measures such as the evaluation of pain rather than shared decision-making. (Schellingerhout et al., 2012)

The use of a checklist such as the COSMIN standards involves the application of a framework. While this creates transparency and reproducibility in the data extraction and evaluation, it can also be reductive, compressing study findings into a more standardised form with the potential loss of information. (Liberati et al., 2009) This is especially relevant for the four-point scoring version as, for example, the COMRADE scale was graded as poor for internal consistency, despite meeting all but one of the methodological standards for this property. (Edwards et al., 2003, Terwee et al., 2012)

Systematic review and COSMIN guidelines recommend two independent reviewers for data extraction. (Higgins and Green, 2011, Centre for Reviews and Dissemination, 2009, Mokkink et al., 2009) Due to time restrictions, a second reviewer checked data extractions forms but did not independently perform data extraction. In addition, a suggested limitation of the COSMIN checklist is the intra-rater reliability, largely due to variation in item interpretation and study report terminology. (Mokkink et al., 2010a, Schellingerhout et al., 2012) Seeking additional reviewers to check data extraction would have addressed this issue, an option restricted by limited resources. However, in accordance with recommendations, the reviewers

discussed the data extraction fields before their use, reducing subjectivity introduced by individual independent interpretations. (Mokkink et al., 2010a, Schellingerhout et al., 2012

Variation in the terminology was noted even between the COSMIN checklist and the quality criteria for instrument measurement properties, despite both having authors in common. In addition, there was repetition and variation in how the measurement properties were covered by the checklist and the quality criteria. For example, in the COSMIN checklist, the items for structural validity were already covered in other domains and this property was therefore omitted for clarity. Floor-ceiling effects are considered alone in the quality criteria but pooled with interpretability in the COSMIN checklist, while coverage of item response theory (IRT) was weak in the measurement property criteria in comparison with COSMIN. (Mokkink et al., 2012, Terwee et al., 2007)

The variation in terminology and methods used were also notable between the studies identified. These are well illustrated by the further evaluation studies identified for the Decisional Conflict Scale (DCS), where the property identified as hypothesis testing or construct validity by the standards and criteria was termed criterion, contrast, discriminatory, convergent and divergent validity, and investigated using a broad range of approaches. (Katapodi et al., 2011, Knapp et al., 2009, Koedoot et al., 2001, Linder et al., 2011, Mancini et al., 2006, Song and Sereika, 2006) As such, no meta-analysis was possible due to the variation in methods and statistical processes used. (Gopalakrishnan and Ganeshkumar, 2013) In addition, face validity was considered too subjective by the authors of COSMIN and was omitted from the standards, despite being an approach encountered often in measure development. (Mokkink et al., 2012)

When study authors were not explicit about the measurement properties assessed, this information was extrapolated from the reported approach taken as COSMIN advises, although this introduces the subjective analysis of the reviewer. (Mokkink et al., 2012) COSMIN also suggests that the measurement properties evaluated should depend on the perceived use of the scale. (Mokkink et al., 2012) In shared decision-making, for example, validity may be of greater relevance when determining decision readiness with reliability and responsiveness more so in a scale intended for decision aid evaluation. (Mokkink et al., 2012) In addition, for score interpretation, consideration of the minimally important change or difference (MID and MID) would aid application. (Mokkink et al., 2012, Terwee et al., 2007) However, such selectivity was rarely covered explicitly in study reports. It is also worth noting that the intended use of a scale may change over time especially with the increasing development of decision aids and interventions.

The context specific nature of instrument measurement properties was illustrated by the further evaluation studies identified in this review, with the results of psychometric analysis differing between groups tested. For this review, the synthesis was therefore kept focused on the original development studies but this also highlights the need to re-evaluate measures when used in a new context or population. (Streiner and Norman, 2008)

Items from the measures identified were compared to a decision process map incorporating deliberation. Elwyn and Miron-Shatz, drawing on their experience in the field of decision-making, developed the underlying model, which is described as “an agreed process map of decision-making”. (Elwyn and Miron-Shatz, 2010) However, the authors note its theoretical nature, with

further verification currently lacking. (Elwyn and Miron-Shatz, 2010) In addition, the item mapping performed in this review was subjective and based on the opinions of the reviewers involved. Alternative approaches would have been to include field experts or patient representatives in the process, while another option would have been to use a different review method, such as a realist synthesis. (Rycroft-Malone et al., 2012) However, these approaches would have been difficult with the resources available.

8. Conclusion

This systematic review, the first in shared decision-making to utilise the COSMIN methodological standards, found that current measures of decision-making do not consider all steps of the decisional process. (Mokkink et al., 2012) Only the DMQS by Hollen, which is based on the work of Janis and Mann, considers the steps attributed to deliberation by Elwyn and Shatz. (Elwyn and Miron-Shatz, 2010, Hollen, 1994, Janis and Mann, 1977) In addition, the methodological approach and measurement properties for existing instruments are highly variable. As such, a suitable measure is needed to further explore the decision-making process and mechanism of action of decision support interventions. An instrument with robust methodological development and measurement properties, based on decision-process models incorporating deliberation and the work of Janis and Mann, is recommended.

9. Appendices

Appendix 1: COSMIN measurement property definitions

Term			Definition
Domain	**Measurement property**	**Aspect of a measurement property**	
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same health related-patient reported outcomes (HR- PRO) (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true' differences between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which an HR-PRO instrument measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured

	Construct validity		The degree to which the scores of an HR-PRO instrument are consistent with hypotheses *(for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups)* based on the assumption that the HR-PRO instrument validly measures the construct to be measured
		Structural validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Item construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument
	Criterion validity		The degree to which the scores of an HR-PRO instrument are an adequate reflection of a 'gold standard'
Responsiveness			The ability of an HR-PRO instrument to detect change over time in the construct to be measured
	Responsiveness		Item responsiveness
Interpretability			Interpretability is the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores.

Reproduced with permission (Mokkink et al., 2010c)

Appendix 2: COSMIN checklist sample

Box F. Hypotheses testing

Design requirements

1 Was the percentage of missing items given?

2 Was there a description of how missing items were handled?

3 Was the sample size included in the analysis adequate?

4 Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?

5 Was the expected *direction* of correlations or mean differences included in the hypotheses?

6 Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?

7 for convergent validity: Was an adequate description provided of the comparator instrument(s)?

8 for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?

9 Were there any important flaws in the design or methods of the study?

Statistical methods

10 Were design and statistical methods adequate for the hypotheses to be tested?

Source: COSMIN manual (Mokkink et al., 2012)

Appendix 3: Quality criteria for instrument measurement properties

Reproduced with permission (Terwee et al., 2007)

Property	Definition	Quality criteria
1. Content validity	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire	+ A clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection AND target population and (investigators OR experts) were involved in item selection; ? A clear description of above-mentioned aspects is lacking OR only target population involved OR doubtful design or method; - No target population involvement 0 No information found on target population involvement.
2. Internal consistency	The extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct	+ Factor analyses performed on adequate sample size (7 * # items and ≥ 100) AND Cronbach's alpha(s) calculated per dimension AND Cronbach's alpha(s) between 0.70 and 0.95; ?No factor analysis OR doubtful design or method - Cronbach's alpha(s) < 0.70 or > 0.95, despite adequate design and method 0 No information found on internal consistency.

Key: MIC = minimal important change; SDC = smallest detectable change; LOA = limits of agreement; ICC = Intraclass correlation; SD, standard deviation. + = positive rating; ? = indeterminate rating; - = negative rating; 0 = no information available. Doubtful design or method = lacking of a clear description of the design or methods of the study, sample size smaller than 50 subjects (should be at least 50 in every (subgroup) analysis), or any important methodological weakness in the design or execution of the study.

3. Criterion validity	The extent to which scores on a particular questionnaire relate to a gold standard	+ Convincing arguments that gold standard is ‘‘gold’’ AND correlation with gold standard ≥0.70; ? No convincing arguments that gold standard is ‘‘gold’’ OR doubtful design or method; - Correlation with gold standard <0.70, despite adequate design and method; 0 No information found on criterion validity.
4. Construct validity	The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured	+ Specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses; ? Doubtful design or method (e.g., no hypotheses); - Less than 75% of hypotheses were confirmed, despite adequate design and methods; 0 No information found on construct validity.
5. Reproducibility		

5.1. Agreement	The extent to which the scores on repeated measures are close to each other (absolute measurement error)	+ MIC < SDC OR MIC outside the LOAOR convincing arguments that agreement is acceptable; ? Doubtful design or method OR (MIC not defined AND no convincing arguments that agreement is acceptable); - MIC ≥ SDC OR MIC equals or inside LOA, despite adequate design and method; 0 No information found on agreement.
5.2. Reliability	The extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error)	+ ICC or weighted Kappa ≥ 0.70; ? Doubtful design or method (e.g., time interval not mentioned); - ICC or weighted Kappa < 0.70, despite adequate design and method; 0 No information found on reliability.

Responsiveness	The ability of a questionnaire to detect clinically important changes over time	+ SDC or SDC < MIC OR MIC outside the LOA OR RR > 1.96 OR AUC ≥ 0.70; ? Doubtful design or method; - SDC or SDC ≥ MIC OR MIC equals or inside LOA OR RR ≤ 1.96 OR AUC < 0.70, despite adequate design and methods; 0 No information found on responsiveness.
Floor & ceiling effects	The number of respondents who achieved the lowest or highest possible score	+ ≤15% of the respondents achieved the highest or lowest possible scores; ? Doubtful design or method; - > 15% of the respondents achieved the highest or lowest possible scores, despite adequate design and methods; 0 No information found on interpretation.
Interpretability	The degree to which one can assign qualitative meaning to quantitative scores	+ Mean and SD scores presented of at least four relevant subgroups of patients and MIC defined; ?Doubtful design or method OR less than four subgroups OR no MIC defined; 0 No information found on interpretation.

Appendix 4: Terms excluded from search strategy

Concept	Term	Impact on search results
Shared decision-making	Choice behaviour Decision support techniques Decision theory	Too broad: focus not on decision-making process
	Patient education Comprehension, understanding and knowledge Health attitudes, knowledge and practice	Focus on education and knowledge rather than decision-making
	Co-operative behaviour Physician-patient relationship	Aspects of relationship rather than decision-making
Population	Person, subject Family, carer	Too broad Not from patients' perspective
	Patient-centred care Focused or orientated	Shift to consultation approach
	Advocacy or empowerment Satisfaction, opinion or evaluation Empower, involve, promote or support	Shift to service provision
	Healthcare	Incorporates clinician decision-making
Instrument	Data collection or research design Interviews or self-report	Too broad Shift to qualitative or diagnostic

Appendix 5: Search strategies

5.1. Medline via Ovid

1. ((patient* or client* or consumer*) adj3 decision making).mp.
2. exp patients/
3. exp Patient Participation/
4. exp evaluation studies/ or exp validation studies/
5. psychometrics/
6. reproducibility of results/
7. (valid* or reliab*).mp.
8. ((measur* or scal* or scor* or instrument* or survey* or tool* or question*) adj6 decision*).mp.
9. exp Decision Making/
10. shared decision-making.mp.
11. ((consider* or reflec* or deliberat*) adj3 decision*).mp.
12. (decision* or choice* or prefer* or judg*).tw.
13. or/1-3
14. or/4-8

5.2 EMBASE via Ovid

1. ((measur* or survey* or scal* or scor* or instrument* or tool* or question*) adj6 decision*).mp.
2. reproducibility/
3. evaluation/
4. validation study/
5. (valid* or reliab*).mp.
6. or/1-5
7. patient participation/
8. ((patient* or client* or consumer*) adj6 decision-making).mp.
9. or/7-8
10. patient decision making/
11. medical decision making/
12. shared decision-making.mp.
13. (delib* adj3 decision-making).mp.
14. ((consider* or reflec*) adj3 decision-making).tw.
15. or/10-14
16. 6 and 9 and 15

5.3. The Cochrane Library via Wiley Online Library

ID	Search
#1	((patient* or client* or consumer*) near/3 "decision making"):ti,ab,kw
#2	MeSH descriptor: [Patient Participation] 1 tree(s) exploded
#3	#1 or #2
#4	MeSH descriptor: [Evaluation Studies as Topic] explode all trees
#5	MeSH descriptor: [Psychometrics] this term only
#6	(measur* or scal* or scor* or survey* or instrument* or tool* or question*):ti,ab,kw
#7	(valid* or reliab*):ti,ab,kw
#8	#4 or #5 or #6 or #7
#9	MeSH descriptor: [Decision Making] explode all trees
#10	"shared decision making":ti,ab,kw
#11	((deliberat* or choice* or reflec*) near/3 "decision making"):ti,ab,kw
#12	(decision* or choice* or prefer*):ti,ab,kw
#13	#9 or #10 or #11 or #12
#14	#3 and #8 and #13

5.4. PsycINFO via Ovid

1. exp Decision Making/
2. shared decision-making.mp.
3. (decision* or choice* or prefer*).tw.
4. ((consider* or reflec* or deliberat*) adj3 decision making).mp.
5. or/1-4
6. exp client participation/
7. ((patient* or client* or consumer*) adj6 decision making).mp.
8. or/6-7
9. ((measur* or survey* or scal* or scor* or instrument* or tool* or question*) adj6 decision*).mp.
10. (valid* or reliab*).mp.
11. exp Psychometrics/
12. or/9-11
13. 5 and 8 and 12

5.5 . Medline in Process via Ovid

1. ("evaluation studies" or "validation studies").mp.
2. psychometric*.mp.
3. "reproducibility of results".mp.
4. ((measur* or survey* or scal* or scor* or instrument* or tool* or question*) adj6 decision*).mp.
5. (valid* or reliab*).mp.
6. or/1-5
7. shared decision-making.mp.
8. decision making.mp.
9. (deliberat* adj3 decision making).mp.
10. ((consider* or reflec*) adj3 decision making).mp.
11. (decision* or choice* or prefer*).tw.
12. or/7-11
13. (patient adj (perspective or opinion)).tw.
14. ((patient* or client* or consumer*) adj3 decision*).mp.
15. or/13-14
16. 6 and 12 and 15

5.6. CINAHL via EBSCOhost

Search ID number	Search terms	Search options
S6	S3 and S4 and S5	Search modes – Boolean/Phrase
S5	(MH "Decision Making+") OR (MH "Decision-Making Support (Iowa NIC)") OR "deliberat$" OR "reflec$" OR "deliberat$" OR "shared decision making"	Search modes – Boolean/Phrase
S4	(MH "Patients+") OR (MH "Decision Making, Patient+")	Search modes – Boolean/Phrase
S3	S1 or S2	Search modes – Boolean/Phrase
S2	(MH "Validation Studies") OR (MH "Evaluation Research") OR "valid$" OR "reliab$" OR (MH "Psychometrics")	Search modes – Boolean/Phrase
S1	"measur$" OR "survey$" OR "scal$" OR "scor$" OR "instrument$" OR "tool$" OR "question$"	Search modes – Boolean/Phrase

5.7: ASSIA via ProQuest

(SU.EXACT("Patient participation") OR SU.EXACT("Patient control") OR SU.EXACT("Health professional-Patient interactions") AND SU.EXACT("Patients")) AND (SU.EXACT("Psychometric properties") OR all((valid* OR reliab*)) OR all((measur* OR survey* OR scal* OR scor* OR instrument* OR tool* OR question*))) AND (SU.EXACT("Decision making") OR all("shared decision making") OR all((deliber* OR consider* OR reflect*)))

Appendix 6: Comparison of COSMIN 4-point scale and checklist

COSMIN 4-point scale		
	Pro	**Con**
Purpose	Rating system where methodology classified into different quality levels, allowing ease of summary and comparison	
Development	Has been utilised in published systematic reviews	Still in development as main reporting paper based on conclusions of only one reviewer
Grading system used in COSMIN		Worst score counts with conclusions at measurement property level rather at individual item or components level, with subsequent loss of detail
General use of scoring system	Quality score combines information on several features into one value	"Numerical scales summarising different components can be misleading…better to describe" (Liberati et al., 2009)
COSMIN checklist		
	Pro	**Con**
Purpose	Allows separate standardised judgement of methodological quality of included studies and their results.	
Component based approach to methodological quality assessment	Uses component-based approach, where key dimensions of trial quality are individually examined without calculating a score, as recommended by (Liberati et al., 2009)	

Appendix 7: Data extraction fields

For each full text report:	Data extraction field:
Reference details - for records only	Reviewer Date of review
Descriptive features	Scale name First author, year Reference Country Setting Population Decision Purpose Description Language Citation Comments
Methodological quality **- COSMIN checklist and COSMIN 4-point scale**	Reliability - Internal consistency - Reliability - Measurement error Validity - Content validity - Face validity - Criterion validity - Construct validity - Structural validity - Hypotheses-testing - Cross-cultural validity - Responsiveness - Interpretability - Generalisability
Instrument quality **- Terwee et al, 2007: quality criteria for measurement properties of instruments**	Content validity Internal consistency Criterion validity Construct validity Reproducibility - Agreement - Reliability Responsiveness Floor and ceiling effect Interpretability
Map items onto deliberation/determination components	Option awareness Deliberation -Knowledge construct - Information search - Appraisal of knowledge sufficiency - Preference construction - Imagining counterfactuals - Affective forecasting - Ranking options Determination Time scale between deliberation and assessment

Appendix 8: Measurement scales

8.1. Decisional Conflict Scale (O'Connor, 1995)

1. I know which options are available to me.

2. I know the benefits of each option.

3. I know the risks and side effects of each option.

4. I am clear about which benefits matter most to me.

5. I am clear about which risks and side effects matter most.

6. I am clear about which is more important to me (the benefits or the risks and side effects).

7. I have enough support from others to make a choice.

8. I am choosing without pressure from others.

9. I have enough advice to make a choice.

10. I am clear about the best choice for me.

11. I feel sure about what to choose.

12. This decision is easy for me to make.

13. I feel I have made an informed choice.

14. My decision shows what is important to me.

15. I expect to stick with my decision.

16. I am satisfied with my decision.

8.2. The Satisfaction with Decision Scale (Holmes-Rovner et al., 1996)

1. I am satisfied that I am adequately informed about the issues important to my decision
2. The decision I made was the best decision possible for me personally
3. I am satisfied that my decision was consistent with my personal values
4. I expect to successfully carry out (or continue to carry out) the decision I made
5. I am satisfied that this was my decision to make
6. I am satisfied with my decision

8.3. Decision Attitude Scale, (Sainfort and Booske, 2006)

1. My decision is sound
2. I am comfortable with my decision
3. My decision is the right one for my situation
4. I am satisfied with my decision
5. It was difficult to make a choice
6. I had no problem using the information
7. The information was easy to understand
8. Consulting someone else would have been useful
9. More information would help

8.4. Preparation for Decision Making scale (Bennett et al., 2010)

Please show your opinion of [the educational material] by circling the number to show how much you agree with each statement.

- Help you recognize that a decision needs to be made?
- Prepare you to make a better decision?
- Help you think about the pros and cons of each option?
- Help you think about which pros and cons are most important?
- Help you know that the decision depends on what matters most to you?
- Help you organize your own thoughts about the decision?
- Help you think about how involved you want to be in this decision?
- Help you identify questions you want to ask your doctor?
- Prepare you to talk to your doctor about what matters most to you?
- Prepare you for a follow-up visit with your doctor?

8.5. SDM-Q (Simon et al., 2006)

1. In the selection of the treatment method, & my thoughts were taken into account just as much as the considerations of my doctor

2. There was enough time for questions

3. My doctor and I weighed up the different treatment options thoroughly

4. I was able to discuss the different treatment options with my doctor in detail

5. My doctor and I selected a treatment option together

6. I now know the advantages of the individual treatment options

7. I now know which treatment option is the best one for me

8. During the consultation, I felt included in the treatment decision

9. Through the consultation with the doctor, I felt jointly responsible for my further treatment

10. My doctor and I discussed the next steps of the treatment plan in detail

11. My doctor and I reached an agreement as to how we will proceed

8.6 The SURE test (Légaré et al., 2010a)

A response of yes scores 1 and a response of no scores 0; a score of < 4 is a positive result for decisional conflict.

1. Sure of myself – do you feel SURE about the best choice for you?

2. Understand information – do you know the benefits and risks of each option?

3. Risk-benefit ratio – are you clear about which benefits and risks matter most to you?

4. Encouragement – do you have enough support and advice to make a choice?

8.7. COMRADE (Edwards et al., 2003)

Satisfaction with communication

1. The doctor made me aware of the different treatments available

2. The doctor gave me the chance to express my opinions about the different treatments available

3. The doctor gave me the chance to ask for as much information as I needed about the different treatment choices

4. The doctor gave me enough information about the treatment choices available

5. The doctor gave enough explanation of the information about treatment choices

6. The information given to me was easy to understand

7. I know the advantages of treatment or not having treatment

8. I know the disadvantages of treatment or not having treatment

9. The doctor gave me a chance to decide which treatment I thought was best for me

10. The doctor gave me a chance to be involved in the decisions during the consultation

Confidence in decision

1. Overall, I am satisfied with the information I was given

2. My doctor and I agreed about which treatment (or no treatment) was best for me

3. I can easily discuss my condition again with my doctor

4. I am satisfied with the way in which the decision was made in the consultation

5. I am sure that the decision made was the right one for me personally

6. I am satisfied that I am adequately informed about the issues important to the decision

7. It is clear which choice is best for me

8. I am aware of the treatment choices I have

9. I feel an informed choice has been made

10. The decision shows what is most important to me

8.8. Decision Evaluation Scales (Stalmeier et al., 2005)

1. I expect to stick with my decision

2. I am satisfied with my decision

3. I am still doubtful about my choice

4. This is my own decision

5. I find it hard to make this choice

6. I am satisfied with the information I received

7. I know the pros and cons of the treatments

8. I want more information about this decision

9. I want a clearer advice

10. I made a well informed choice

11. This decision is made without me

12. I feel pressure from others in making this decision

13. I wish someone else would decide for me

14. My decision frightens me

15. I regret my decision

8.9. Decision Making Quality scale (Hollen, 1994)

"How true are these statements about your decision making choices?"

1. Searches for three or more choices
2. Takes into account values and all goals desired
3. Weighs the pros and cons of consequences
4. Finds more information about the pros and cons, when needed
5. Thinks about new information and what experts say, even if against first choice
6. Reviews carefully before making a final choice
7. Makes details plans with back up plans

8.10. Decision Self-Efficacy Scale (Bunn and O'Connor, 1996)

I feel confident that I can:

1 Get the facts about the medication choices available to me

2. Get the facts about the benefits of each choice

3. Get the facts about the risks and side effects of each choice

4. Understand the information enough to be able to make a choice

5. Ask questions without feeling dumb

6. Express my concerns about each choice

7. Ask for advice

8. Figure out the choice that best suits me

9. Handle unwanted pressure from others in making my choice

10. Let the clinic team know what's best for me

11. Delay my decision if I feel I need more time

10. References

AGORITSAS, T., HEEN, A. F., BRANDT, L., ALONSO-COELLO, P., KRISTIANSEN, A., AKL, E. A., NEUMANN, I., TIKKINEN, K. A., VAN DER WEIJDEN, T., ELWYN, G., MONTORI, V. M., GUYATT, G. H. & VANDVIK, P. O. 2015. Decision aids that really promote shared decision making: the pace quickens. *British Medical Journal* [Online], 350. Available: http://dx.doi.org/doi:10.1136/bmj.g7624.

BANDURA, A. 1977. *Social learning theory,* Englewood Cliffs, N.J., Prentice Hall.

BARR, P. J. & ELWYN, G. 2015. Measurement challenges in shared decision making: putting the "patient" in patient-reported measures. *Health Expectations* [Online]. Available: http://www.readcube.com/articles/10.1111%2Fhex.12380?tracking_referrer=onlinelibrary.wiley.com&parent_url=http%3A%2F%2Fonlinelibrary.wiley.com%2Fdoi%2F10.1111%2Fhex.12380%2Fepdf&preview=1 [Accessed Jun].

BARRY, M. J., CHERKIN, D. C., YUCHIAO, C., FOWLER, F. J. & SKATES, S. 1997. A Randomized Trial of a Multimedia Shared Decision-Making Program for Men Facing a Treatment Decision for Benign Prostatic Hyperplasia. *Disease Management and Clinical Outcomes,* 1, 5-14.

BENNETT, C., GRAHAM, I. D., KRISTJANSSON, E., KEARING, S. A., CLAY, K. F. & O&APOS;CONNOR, A. M. 2010. Validation of a Preparation for Decision Making scale. *Patient Education and Counseling,* 78, 130-133.

BOSSUYT, P. M. & LEEFLANG, M. M. 2008. *Developing criteria for including studies: Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* [Online]. The Cochrane Collaboration. Available: http://srdta.cochrane.org/sites/srdta.cochrane.org/files/uploads/Chapter06-Including-Studies %28September-2008%29.pdf [Accessed 21st August 2011].

BRISS, P., RIMER, B., REILLEY, B., COATES, R. C., LEE, N. C., MULLEN, P., CORSO, P., HUTCHINSON, A. B., HIATT, R., KERNER, J., GEORGE, P., WHITE, C., GANDHI, N., SARAIYA, M., BRESLOW, R., ISHAM, G., TEUTSCH, S. M., HINMAN, A. R. & LAWRENCE, R. 2004. Promoting Informed Decisions About Cancer Screening in Communities and Healthcare Systems. *American Journal of Preventive Medicine,* 26, 67-80.

BUNN, H. & O'CONNOR, A. 1996. Validation of client decision-making instruments in the context of psychiatry. *Canadian Journal of Nursing Research,* 28, 13-27.

CARDIFF UNIVERSITY IT AND LIBRARY SERVICES. 2012. *Selected A-Z of databases* [Online]. Available: http://www.cardiff.ac.uk/insrv/eresources/databases/ [Accessed Jan 2012].
CENTRE FOR REVIEWS AND DISSEMINATION. 2009. *Systematic Reviews: CRD's guidance for undertaking reviews in health care* [Online]. York: University of York. Available: http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf [Accessed 12th August 2011].
CHARLES, C., GAFNI, A. & WHELAN, T. 1997. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango) *Social Science and Medicine,* 44, 681-692.
COCHRANE CONSUMER AND COMMUNICATION REVIEW GROUP. *Overarching search strategy for CENTRAL* [Online]. Available: http://www.latrobe.edu.au/chcp/cochrane/resources.html [Accessed 1 March 2012].
COLLINGS, A. & COULTER, A. 2015. *Making shared decision-making a reality* [Online]. London: King's Fund. Available: http://www.kingsfund.org.uk/sites/files/kf/Making-shared-decision-making-a-reality-paper-Angela-Coulter-Alf-Collins-July-2011_0.pdf [Accessed January 2015].
CRITICAL APPRAISAL SKILLS PROGRAMME (CASP). 2011. *Systematic review critical appraisal* [Online]. Available: http://www.casp-uk.net [Accessed 20th July 2011].
DE VET, H.C.W., EISINGA, A., RIPHAGEN, I. I., AERTGEERTS, B. & PEWSNER, D. 2008. *Search for Studies: Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* [Online]. The Cochrane Collaboration. Available: http://srdta.cochrane.org/handbook-dta-reviews [Accessed 28th August 2011].
DORMAN, S. 2005. *Which measurement scales should we use to measure breathlessness in palliative care?* MSc, Cardiff University.
EDWARDS, A., ELWYN, G., HOOD, K., ROBLING, M., ATWELL, C., HOLMES-ROVNER, M., KINNERSLEY, P., HOUSTON, H. & RUSSELL, I. 2003. The development of COMRADE—a patient-based outcome measure to evaluate the effectiveness of risk communication and treatment decision making in consultations. *Patient Education and Counseling,* 50, 311-322.
EDWARDS, A. & ELWYN, G. 2006. Inside the black box of shared decision making: distinguishing between the process of involvement and who makes the decision. *Health Expectations,* 9, 307-320.
ELWYN, G., EDWARDS, A., ECCLES, M. & ROVNER, D. 2001a. Decision analysis in patient care. *The Lancet,* 358, 571-574.
ELWYN, G., EDWARDS, A., MOWLE, S., WENSING, M., WILKINSON, C., KINNERSLEY, P. & GROL, R. 2001b. Measuring the involvement of

patients in shared decision-making: a systematic review of instruments. *Patient Education and Counseling,* 43, 5-22.
ELWYN, G. & MIRON-SHATZ, T. 2010. Deliberation before determination: the definition and evaluation of good decision making. *Health Expectations,* 13, 139-147.
ELWYN, G., FROSCH, D., THOMSON, R., JOSEPH-WILLIAMS, N., LLOYD, A., KINNERSLEY, P., CORDING, E., TOMSON, D., DODD, C., ROLLNICK, S., EDWARDS, A. & BARRY, M. 2012. Shared Decision Making: A Model for Clinical Practice. *Journal of General Internal Medicine,* 27, 1361-1367.
ELWYN, G., BARR, P. J., GRANDE, S. W., THOMPSON, R., WALSH, T. & OZANNE, E. M. 2013. Developing CollaboRATE: A fast and frugal patient-reported measure of shared decision making in clinical encounters. *Patient Education and Counseling,* 93, 102-107.
ELWYN, G., LLOYD, A., MAY, C., VAN DER WEIJDEN, T., STIGGELBOUT, A., EDWARDS, A., FROSCH, D. L., RAPLEY, T., BARR, P., WALSH, T., GRANDE, S. W., MONTORI, V. & EPSTEIN, R. 2014. Collaborative deliberation: A model for patient care. *Patient Education and Counseling,* 97, 158-164.
ERCI, B. & ÖZDEMIR, S. 2008. Psychometric properties of the Treatment Decision Evaluation Scale in patients with cancer in Turkey. *European Journal of Oncology Nursing,* 12, 464-468.
FERRON PARAYRE, A., LABRECQUE, M., ROUSSEAU, M., TURCOTTE, S. & LEGARE, F. 2014. Validation of SURE, a four-item clinical checklist for detecting decisional conflict in patients. *Medical Decision Making,* 34, 54-62.
FINNELL, D. S. & LEE, J. 2011. Psychometric properties of the decisional balance for patient choice in substance abuse treatment. *Issues in Mental Health Nursing,* 32, 243-9.
GEIGER, F., LIETHMANN K FAU - HOFFMANN, F., HOFFMANN F FAU - PASCHEDAG, J., PASCHEDAG J FAU - KASPER, J. & KASPER, J. 2011. Investigating a training supporting Shared Decision Making (IT'S SDM 2011): study protocol for a randomized controlled trial. *Trials,* 12, 232.
GIGERENZER, G. & GAISSMAIER, W. 2011. Heuristic Decision Making. *Annual Review of Psychology,* 62, 451-482.
GOLDSTEIN, D. G. & GIGERENZER, G. 2009. Fast and frugal forecasting. *International Journal of Forecasting,* 25, 760-772.
GOPALAKRISHNAN, S. & GANESHKUMAR, P. 2013. Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare *Journal of Family Medicine and Primary Care,* 2, 9-14.
HEALTH KNOWLEDGE. 2011. *Health Care Evaluation Frameworks* [Online]. Available: http://www.healthknowledge.org.uk/public-health-

textbook/research-methods/1c-health-care-evaluation-health-care-assessment/hce-frameworks [Accessed December 2015].
HEMINGWAY, P. & BRERETON, N. 2009. *What is a systematic review?* [Online]. Hayward Medical Communications. Available: http://www.whatisseries.co.uk [Accessed October 2014].
HIGGINS, J. P. T. & DEEKS, J. J. 2011. Selecting studies and collecting data. *In:* HIGGINS, J. P. T. & GREEN, S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester: John Wiley and Sons Ltd.
HIGGINS, J. P. T. & GREEN, S. (eds.) 2011. *Cochrane Handbook for Systematic Reviews of Interventions,* Chichester: John Wiley and Sons Ltd.
HOLLEN, P. J. 1994. Psychometric properties of two instruments to measure quality decision making. *Research in Nursing and Health,* 17, 137-148.
HOLMES-ROVNER, M., KROLL, J., SCHMITT, N., ROVNER, D. R., BREER, M. L., ROTHERT, M. L., PADONU, G. & TALARCZYK, G. 1996. Patient satisfaction with health care decisions: the satisfaction with decision scale. *Medical Decision Making,* 16, 58-64.
HYDÉN, L. C. 1997. Illness and narrative. *Sociology of health & illness,* 19, 48-69.
JANIS, I. L. & MANN, L. 1977. *Decision making: a psychological analysis of conflict, choice and commitment,* New York, The Free Press.
JOHNSTON, M. V. & GRAVES, D. E. 2008. Towards Guidelines for Evaluation of Measures: An Introduction With Application to Spinal Cord Injury. *Journal of Spinal Cord Medicine,* 31, 13-26.
JOSEPH-WILLIAMS, N., EDWARDS, A. & ELWYN, G. 2011. The importance and complexity of regret in the measurement of 'good' decisions: a systematic review and a content analysis of existing assessment instruments. *Health Expectations,* 14, 59-83.
KALTOFT, M., CUNICH M FAU - SALKELD, G., SALKELD G FAU - DOWIE, J. & DOWIE, J. 2014. Assessing decision quality in patient-centred care requires a preference-sensitive measure. *Journal of Health Services and Policy,* 19, 110-7.
KATAPODI, M. C., MUNRO, M. L., PIERCE, P. F. & WILLIAMS, R. A. 2011. Psychometric Testing of the Decisional Conflict Scale. *Nursing Research,* 60, 368-377.
KHAN, K. S. 2005. Systematic reviews of diagnostic tests: a guide to methods and application. *Best Practice & Research: Clinical Obstetrics & Gynaecology,* 19, 37-46.
KNAPP, C., HUANG, I., MADDEN, V., VADAPARAMPIL, S., QUINN, G. & SHENKMAN, E. 2009. An evaluation of two decision-making scales for children with life-limiting illnesses. *Palliative Medicine,* 23, 518-525.
KOEDOOT, N., MOLENAAR, S., OOSTERVELD, P., BAKKER, P., DE GRAEFF, A., NOOY, M., VAREKAMP, I. & DE HAES, H. 2001. The

decisional conflict scale: further validation in two samples of Dutch oncology patients. *Patient Education and Counseling,* 45, 187-193.
KREMER, H. & IRONSON, G. 2008. Measuring the involvement of people with HIV in treatment decision making using the control preferences scale. *Medical Decision Making,* 28, 899-908.
KRISTON, L., SCHOLL, I., LZEL, L. H., SIMON, D., LOH, A. & RTER, M. H. 2010. The 9-item Shared Decision Making Questionnaire (SDM-Q-9). Development and psychometric properties in a primary care sample. *Patient Education and Counseling,* 80, 94-99.
LEFEBVRE, C., MANHEIMER, E. & GLANVILLE, J. 2009. Searching for studies. *In:* HIGGINS, J. P. T. & GREEN, S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester: John Wiley & Sons Ltd.
LÉGARÉ, F., KEARING, S., CLAY, K., GAGNON, S., DAMOURS, D., ROUSSEAU, M. & O'CONNOR, A. 2010a. Are you SURE? *Canadian Family Physician,* 56, e308-14.
LÉGARÉ, F., RATTE, S., STACEY, D., KRYWORUCHKO, J., GRAVEL, K., GRAHAM, I. & TURCOTTE, S. 2010b. Interventions for improving the adoption of shared decision making by healthcare professionals (Review). *The Cochrane Library*.
LIBERATI, A., ALTMAN, D. G., TETZLAFF, J., MULROW, C., GOTZSCHE, P. C., IOANNIDIS, J. P. A., CLARKE, M., DEVEREAUX, P. J., KLEIJNEN, J. & MOHER, D. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *British Medical Journal,* 339, b2700.
LINDER, S. K., SWANK, P. R., VERNON, S. W., MULLEN, P. D., MORGAN, R. O. & VOLK, R. J. 2011. Validity of a low literacy version of the Decisional Conflict Scale. *Patient Education and Counseling,* 85, 521-524.
LOCKWOOD, A. & MARSHALL, M. 1998. Assertive community treatment for people with severe mental disorders (Review). *The Cochrane Library*.
MAKOUL, G. & CLAYMAN, M. L. 2006. An integrative model of shared decision making in medical encounters. *Patient Education and Counseling,* 60, 301-312.
MANCINI, J., SANTIN, G., CHABAL, F. & JULIAN-REYNIER, C. 2006. Cross-cultural validation of the Decisional Conflict Scale in a sample of French patients. *Quality of Life Research,* 15, 1063-1068.
MANN, M. 2011. *Cardiff University Systematic Review Network - SysNET* [Online]. Available: http://www.cardiff.ac.uk/insrv/libraries/sure/sysnet/ [Accessed 21st August 2011].
MARSHALL, M., LOCKWOOD, A., BRADLEY, C., ADAMS, C., JOY, C. & FENTON, M. 2000. Unpublished rating scales: a major source of bias in

randomised controlled trials of treatments for schizophrenia. *British Journal of Psychiatry,* 176, 249-252.
MARTEAU, T. M., DORMANDY, E. & MICHIE, S. 2001. A measure of informed choice. *Health Expectations,* 4, 99-108.
MELBOURNE, E., SINCLAIR K FAU - DURAND, M.-A., DURAND MA FAU - LEGARE, F., LEGARE F FAU - ELWYN, G. & ELWYN, G. 2010. Developing a dyadic OPTION scale to measure perceptions of shared decision making. *Patient Education and Counseling,* 78, 177-83.
MILLER, V. A., REYNOLDS WW FAU - ITTENBACH, R. F., ITTENBACH RF FAU - LUCE, M. F., LUCE MF FAU - BEAUCHAMP, T. L., BEAUCHAMP TL FAU - NELSON, R. M. & NELSON, R. M. 2009. Challenges in measuring a new construct: perception of voluntariness for research and treatment decision making. *Journal of Empirical Research on Human Research Ethics,* 4, 21-31.
MOHER, D., LIBERATI, A., TETZLAFF, J., ALTMAN, D. & THE PRISMA GROUP 2009. Preferred Reporting Items for the Systematic Reviews and Meta-Analysis: The PRISMA statement. *British Medical Journal,* 339, 2535.
MOKKINK, L. B., TERWEE, C. B., STRATFORD, P. W., ALONSO, J., PATRICK, D. L., RIPHAGEN, I., KNOL, D. L., BOUTER, L. M. & DEVET, H. C. 2009. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research,* 18, 313-333.
MOKKINK, L. B., TERWEE, C. B., GIBBONS, E., STRATFORD, P. W., ALONSO, J., PATRICK, D. L., KNOL, D. L., BOUTER, L. M. & DE VET, H. C. 2010a. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) Checklist. *BMC Medical Research Methodology,* 10, 82.
MOKKINK, L. B., TERWEE, C. B., PATRICK, D. L., ALONSO, J., STRATFORD, P. W., KNOL, D. L., BOUTER, L. M. & DE VET, H. C. W. 2010b. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research,* 19, 539-549.
MOKKINK LB, TERWEE CB, PATRICK DL, ALONSO J, STRATFORD PW, KNOL DL, BOUTER LM, DE VET HCW. 2010c. International consensus on taxonomy, terminology, and definitions of measurement properties: results of the COSMIN study. *Journal of Clinical Epidemiology,* 63, 737-745.
MOKKINK, L. B., TERWEE, C. B., PATRICK, D. L., ALONSO, J., STRATFORD, P. W., KNOL, D. L., BOUTER, L. M. & DE VET, H. C. 2012. *COSMIN checklist manual* [Online]. Available: http://www.cosmin.nl [Accessed 2012].

O'CONNOR, A., PENNIE, R. A. & DALES, R. E. 1996. Framing effects on expectations, decisions, and side effects experienced: the case of influenza immunisation. *Journal of Clinical Epidemiology,* 49**,** 1271-1276.
O'CONNOR, A., ELWYN, G., STACEY, D., VOLK, R., THOMSON, R., BARRATT, A., BARRY, M., COULTER, A., HOLMES-ROVNER, M., LLEWELLYN-THOMAS, H., MOUMJID, N., WHELAN, T. & IPDAS. 2005. *IPDAS Collaboration Reaches Consensus on Indicators for Judging the Quality of Patient Decision Aids* [Online]. International Patient Decision Aid Standards (IPDAS) Collaboration. Available: http://ipdas.ohri.ca/resources.html [Accessed January 2015].
O'CONNOR, A., BENNETT, C. L., STACEY, D., BARRY, M., COL, N. F., EDEN, K. B., ENTWISTLE, V. A., FISET, V., HOLMES-ROVNER, M., KHANGURA, S., LLEWELLYN-THOMAS, H. & ROVNER, D. 2009. Decision aids for people facing health treatment or screening decisions (Review). *The Cochrane Library*.
O'CONNOR, A. M. 1995. Validation of a decisional conflict scale. *Medical Decision Making,* 15**,** 25-30.
O'CONNOR, A. M., LLEWELLYN-THOMAS, H. A., SAWKA, C., PINFOLD, S. P., TO, T. & HARRISON, D. E. 1997. Physicians' opinions about decision aids for patients considering systemic adjuvant therapy for axillary-node negative breast cancer. *Patient Education and Counseling,* 30**,** 143-153.
O'CONNOR, A. M., DRAKE, E. R., WELLS, G. A., TUGWELL, P., LAUPACIS, A. & ELMSLIE, T. 2003. A survey of the decision-making needs of Canadians faced with complex health decisions. *Health Expectations,* 6**,** 97-109.
O'CONNOR, A. M. 2006. *Ottawa Decision Support Framework to Address Decisional Conflict* [Online]. Available: https://decisionaid.ohri.ca/docs/develop/ODSF.pdf [Accessed January 2015].
O'CONNOR, A. M. 2010. *User manual - Decisional Conflict Scale* [Online]. Available: https://decisionaid.ohri.ca/docs/develop/User_Manuals/UM_Decisional_Conflict.pdf [Accessed January 2015].
OGDEN, J., DANIELLS, E. & BARNETT, J. 2008. The value of choice: development of a new measurement tool. *The British Journal of General Practice,* 58**,** 614-618.
OLSHANSKY, E., SACCO, D., BRAXTER, B., DODGE, P., HUGHES, E., ONDECK, M., STUBBS, M.L., UPVALL, M.J., 2005. Participatory action research to understand and reduce health disparities. *Nursing Outlook,* 53(3), pp.121–126.

PATRICK, D. L., GUYATT, G. H. & ACQUADRO, C. 2009. Patient-reported outcomes. *In:* HIGGINS, J. P. T. & GREEN, S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester: John Wiley and Sons Ltd.
PINK, J., PINK, K. & ELWYN, G. 2009. Measuring Patient Knowledge of Asthma: A Systematic Review of Outcome Measures. *Journal of Asthma,* 46**,** 980-987.
REEVES, B. C., DEEKS, J. J., HIGGINS, J. P. T. & WELLS, G. A. 2009. Including non-randomized studies. *In:* HIGGINS, J. P. T. & GREEN, S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester: John Wiley and Sons Ltd.
RYAN, R. 2013. *Cochrane Consumers and Communication Review Group: data synthesis and analysis* [Online]. Cochrane Consumers and Communication Review Group. Available: http://cccrg.cochrane.org [Accessed Jan 2015].
RYCROFT-MALONE, J., MCCORMACK, B., HUTCHINSON, A. M., DECORBY, K., BUCKNALL, T. K., KENT, B., SCHULTZ, A., SNELGROVE-CLARKE, E., STETLER, C. B., TITLER, M., WALLIN, L. & WILSON, V. 2012. Realist synthesis: illustrating the method for implementation research. *Implementation Science,* 7**,** 33.
SAINFORT, F. & BOOSKE, B. C. 2006. Measuring post-decision satisfaction. *Medical Decision Making,* 20**,** 51-61.
SCHELLINGERHOUT, J. M., VERHAGEN, A. P., HEYMANS, M. W., KOES, B. W., DE VET HC & TERWEE, C. B. 2012. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. *Quality of Life Research,* 21**,** 659-670.
SCHOLL, I., LOON, M. K.-V., SEPUCHA, K., ELWYN, G., LÉGARÉ, F., HÄRTER, M. & DIRMAIER, J. 2011. Measurement of shared decision making - a review of instruments. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen,* 105**,** 313-324.
SCIENTIFIC ADVISORY COMMITTEE OF THE MEDICAL OUTCOMES TRUST, 2002. Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research,* 11, pp.193–205.
SEPUCHA, K., OZANNE, E., SILVIA, K., PARTRIDGE, A. & MULLEY JR., A. G. 2007. An approach to measuring the quality of breast cancer decisions. *Patient Education and Counseling,* 65**,** 261-269.
SEPUCHA, K. & OZANNE, E. M. 2010. How to define and measure concordance between patients' preferences and medical treatments: A systematic review of approaches and recommendations for standardization. *Patient Education and Counseling,* 78**,** 12-23.
SEPUCHA, K. R., FOWLER JR, F. J. & MULLEY JR., A. G. 2004. Policy Support For Patient-Centered Care: The Need For Measurable Improvements In Decision Quality. *Health Affairs,* Suppl**,** 54-62.

SEPUCHA, K. R., STACEY D FAU - CLAY, C. F., CLAY CF FAU - CHANG, Y., CHANG Y FAU - COSENZA, C., COSENZA C FAU - DERVIN, G., DERVIN G FAU - DORRWACHTER, J., DORRWACHTER J FAU - FEIBELMANN, S., FEIBELMANN S FAU - KATZ, J. N., KATZ JN FAU - KEARING, S. A., KEARING SA FAU - MALCHAU, H., MALCHAU H FAU - TALJAARD, M., TALJAARD M FAU - TOMEK, I., TOMEK I FAU - TUGWELL, P., TUGWELL P FAU - LEVIN, C. A. & LEVIN, C. A. 2011. Decision quality instrument for treatment of hip and knee osteoarthritis: a psychometric evaluation. *BMC Musculoskeletal Disorders,* 12, 149.
SEPUCHA, K. R., FEIBELMANN S FAU - ABDU, W. A., ABDU WA FAU - CLAY, C. F., CLAY CF FAU - COSENZA, C., COSENZA C FAU - KEARING, S., KEARING S FAU - LEVIN, C. A., LEVIN CA FAU - ATLAS, S. J. & ATLAS, S. J. 2012. Psychometric evaluation of a decision quality instrument for treatment of lumbar herniated disc. *Spine,* 37, 1609-16.
SEPUCHA, K. R., BORKHOFF, C. M., LALLY, J., LEVIN, C. A., MATLOCK, D. D., NG, C. J., ROPKA, M. E., STACEY, D., JOSEPH-WILLIAMS, N., WILLS, C. E. & THOMSON, R. 2013. Establishing the effectiveness of patient decision aids: key constructs and measurement instruments. *BMC Medical Informatics and Decision Making,* 13, S12.
SHEA, B. J., GRIMSHAW, J. M., WELLS, G. A., BOERS, M., ANDERSSON, N., HAMEL, C., PORTER, A. C., TUGWELL, P., MOHER, D. & BOUTER, L. M. 2007. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology,* 7, 10.
SIMON, D., SCHORR, G., WIRTZ, M., VODERMAIER, A., CASPARI, C., NEUNER, B., SPIES, C., KRONES, T., KELLER, H., EDWARDS, A., LOH, A. & HÄRTER, M. 2006. Development and first validation of the shared decision-making questionnaire (SDM-Q). *Patient Education and Counseling,* 63, 319-327.
SIMON, D., LOH, A. & HÄRTER, M. 2007. Measuring (shared) decision-making - a review of psychometric instruments. *Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen - German Journal for Quality in Health Care,* 101, 259-267.
SLOVIC, P. 1995. The Construction of Preference. *American Psychologist,* 50, 364-371.
SONG, M.-K. & SEREIKA, S. M. 2006. An evaluation of the Decisional Conflict Scale for measuring the quality of end-of-life decision making. *Patient Education and Counseling,* 61, 397-404.
STACEY, D., LEGARE, F., COL, N. F., BENNETT, C. L., BARRY, M. J., EDEN, K. B., HOLMES-ROVNER, M., LLEWELLYN-THOMAS, H., LYDDIATT, A., THOMSON, R., TREVENA, L. & WU, J. 2014. Decision

aids for people facing health treatment or screening decisions (Review). *The Cochrane Library*.
STALMEIER, P. F. M., ROOSMALEN, M. S., VERHOEF, L. C. G., HOEKSTRA-WEEBERS, J. E. H. M., OOSTERWIJK, J. C., MOOG, U., HOOGERBRUGGE, N. & VAN DAAL, W. A. J. 2005. The decision evaluation scales. *Patient Education and Counseling,* 57, 286-293.
STREINER, D. L. & NORMAN, G. R. 2008. *Health Measurement Scales: a practical guide to their development and use,* Oxford, Oxford University Press.
SUNG, V. W., RAKER, C. A., MYERS, D. L. & CLARK, M. A. 2010. Treatment decision-making and information-seeking preferences in women with pelvic floor disorders. *International urogynecology journal,* 21, 1071-1078.
TERWEE, C. B., BOT, S. D. M., DE BOER, M. R., VAN DER WINDT, D. A. W. M., KNOL, D. L., DEKKER, J., BOUTER, L. M. & DE VET, H. C. W. 2007. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology,* 60, 34-42.
TERWEE, C. B., JANSMA, E. P., RIPHAGEN, I. I. & DE VET, H. C. W. 2009. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research,* 18, 1115-1123.
TERWEE, C. B., MOKKINK, L. B., KNOL, D. L., OSTELO, R. W. J. G., BOUTER, L. M. & DE VET, H. C. W. 2012. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research,* 21, 651-657.
VALDERAS, J. M., FERRER, M., MENDÍVIL, J., GARIN, O., RAJMIL, L., HERDMAN, M. & ALONSO, J. 2008. Development of EMPRO: A Tool for the Standardized Assessment of Patient-Reported Outcome Measures. *Value in Health,* 11, 700-708.
WILLIAMS, B. & HUMMELBRUNNER, R., 2010. *Systems Concepts in Action: A Practitioner's Toolkit*, Redwood City, Stanford University Press.
WILLS, C. E. & HOLMES-ROVNER, M. 2003. Preliminary validation of the Satisfaction With Decision scale with depressed primary care patients. *Health Expectations,* 6, 149-159.
WINSTEPS SOFTWARE. 2014. *WINSTEPS and Facets Rasch software* [Online]. Available: http://www.winsteps.com/index.htm [Accessed September 2014].